INTEGRACIÓN DE DATOS Y ANÁLISIS PREDICTIVO EN TRATAMIENTO DE DROGODEPENDENCIA

VICTOR HUGO MARISCAL CARHUAMACA

MÁSTER EN INGENIERÍA INFORMÁTICA, FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTENSE DE MADRID

Trabajo Fin de Máster en Ingeniería Informática
Curso 2015-2016
Convocatoria setiembre 2016
Calificación obtenida: 6,8

Directora:
María Victoria López López

Autorización de difusión y utilización

VICTOR HUGO MARISCAL CARHUAMACA

Madrid, Setiembre 2016

El abajo firmante, matriculado en el Máster en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "INTEGRACION DE DATOS Y ANALISIS PREDICTIVO EN TRATAMIENTO DE DROGODEPENDENCIA", realizado durante el curso académico 2014-2016 bajo la dirección de Victoria López en el Departamento de Arquitectura de Computadores, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Dedicatoria

A mi familia.

Agradecimientos

A Dios por darme la vida y haberme guiado y permitido que lograra uno más de mis objetivos. Le doy Gracias a María Victoria López López, mi Directora del Trabajo de Fin de Máster, por sus acertados consejos, elocuente paciencia y amplio conocimiento.

Un especial agradecimiento al Ministerio de Educación del Perú, por medio de PRONABEC (Programa Nacional de Becas y Crédito Educativo) por financiarme la estancia, estudios e investigación en la Universidad Complutense de Madrid.

Por ultimo quiero agradecer a todos los docentes del Master de Ingeniería Informática y también a la "Universidad Complutense de Madrid" por los años que me cobijo en sus aulas, para adquirir conocimiento y tener una formación profesional.

Resumen

El análisis de datos actual se enfrenta a problemas derivados de la combinación de datos procedentes de diversas fuentes de información. El valor de la información puede enriquecerse enormemente facilitando la integración de nuevas fuentes de datos y la industria es muy consciente de ello en la actualidad. Sin embargo, no solo el volumen sino también la gran diversidad de los datos constituye un problema previo al análisis. Una buena integración de los datos garantiza unos resultados fiables y por ello merece la pena detenerse en la mejora de procesos de especificación, recolección, limpieza e integración de los datos.

Este trabajo está dedicado a la fase de limpieza e integración de datos analizando los procedimientos existentes y proponiendo una solución que se aplica a datos médicos, centrándose así en los proyectos de predicción (con finalidad de prevención) en ciencias de la salud. Además de la implementación de los procesos de limpieza, se desarrollan algoritmos de detección de outliers que permiten mejorar la calidad del conjunto de datos tras su eliminación. El trabajo también incluye la implementación de un proceso de predicción que sirva de ayuda a la toma de decisiones.

Concretamente este trabajo realiza un análisis predictivo de los datos de pacientes drogodependientes de la Clínica Nuestra Señora de la Paz, con la finalidad de poder brindar un apoyo en la toma de decisiones del médico a cargo de admitir el internamiento de pacientes en dicha clínica. En la mayoría de los casos el estudio de los datos facilitados requiere un pre-procesado adecuado para que los resultados de los análisis estadísticos tradicionales sean fiables.

En tal sentido en este trabajo se implementan varias formas de detectar los outliers: un algoritmo propio (Detección de Outliers con Cadenas No Monótonas), que utiliza las ventajas del algoritmo Knuth-Morris-Pratt para reconocimiento de patrones, y las librerías *outliers* y *Rcmdr* de R. La aplicación de procedimientos de cleaning e integración de datos, así como de eliminación de datos atípicos proporciona una base de datos limpia y fiable sobre la que se implementarán procedimientos de predicción de los datos con el algoritmo de clasificación Naive Bayes en R.

Palabras Clave: Integración de Datos, Data Cleaning, Análisis de Datos, Predicción, Drogodependencia, Datos Atípicos, Regresión Lineal, Naive Bayes, Clasificación, Algoritmo KMP.

Abstract

The current data analysis faces problems arising from the combination of data from various sources. The value of information can be enhanced greatly facilitating the integration of new data sources and industry is well aware of it today. However, not only the volume but also the great diversity of data is a problem prior to analysis. A good integration of data ensures reliable results and therefore worth stopping in the specification process improvement, collecting, cleaning and data integration.

This work is dedicated to the cleaning phase and data integration analyzing existing procedures and proposing a solution that applies to medical data, thus focusing on projects prediction (with purpose of prevention) in health sciences. In addition to the implementation of cleaning processes, develop algorithms of detection of outliers that allow improving the quality of the data set after being eliminated. The work also includes the implementation of a process of prediction that serve as an aid to decision-making.

Specifically this work performs a predictive analysis of the data of patients drug addicts of the clinic Nuestra Señora de la Paz, in order to be able to offer support in decisions of the physician in charge admit the internment of patients in the clinic In the majority of cases the study of data provided requires a proper pre-procesado to traditional statistical analysis results to be reliable.

So in this paper are implemented various ways to detect the outliers: an own algorithm (Detection of Outliers not monotonous chains), that uses the advantages of the algorithm Knuth-Morris-Pratt for pattern recognition, and the bookshops outliers and Rcmdr of R. The application of cleaning procedures and data integration and elimination of outliers provides a clean and reliable base data on which prediction procedures be implemented data with Naive Bayes classification algorithm in R.

Keywords: Data integration, data cleaning, data analysis, prediction, drug dependence, outliers, linear regression, Naive Bayes classification algorithm KMP.

Índice de Contenidos

Autorización de Difusión ..... **III**
Dedicatoria ..... **V**
Agradecimientos ..... **VII**
Resumen y Palabras Clave ..... **IX**
Abstract and Keywords ..... **X**
Índice de Contenidos ..... **XI**
Índice de Figuras ..... **XIII**
Índice de Tablas ..... **XV**
Capítulo 1. Introducción **1**
1.1 Motivación **2**
1.2 Objetivos **3**
1.3 Trabajos Relacionados **3**
1.4 Estructura del Trabajo **5**
Capítulo 2. Visión General **7**
2.1 Drogodependencia **7**
2.2 Data Cleaning **9**
2.2.1 Data Wrangler **9**
2.2.2 Trifacta Wrangler Enterprise **10**
2.2.3 OpenRefine **11**
2.3 Análisis Predictivo **15**
2.3.1 Naïve Bayes Classifier **15**
2.3.2 Regresión Logística **16**
2.3.3 Redes Neuronales **19**
2.4 Fuentes de Información y Tratamiento de Datos de Carácter Personal **23**
Capítulo 3. Detección de Outliers y Análisis Predictivo **27**
3.1 Detección de Outliers en R **28**
3.1.1 Método 1: Detección de Outliers con Cadenas No Monótonas (DOCNM) **28**
3.1.1.1 Algoritmo Knuth-Morris-Pratt (KMP) **30**
3.1.2 Método 2: Detección de Outliers con la librería *Outliers* de R **32**
3.1.3 Método 3: Detección de Outliers con Regresión Lineal Simple **33**
3.2 Análisis Predictivo con R **36**
3.2.1 Algoritmo Naïve Bayes **36**
Capítulo 4. Resultados **37**
4.1 Base de Datos **37**
4.2 Data Cleaning **39**
4.3 Detección de Outliers **42**
4.3.1 Método 1: Detección de Outliers con Cadenas No Monótonas (DOCNM) **42**
4.3.2 Método 2: Detección de Outliers con la librería *Outliers* de R **47**
4.3.3 Método 3: Detección de Outliers con la librería *Rcmdr* de R **48**
4.4 Modelo de Predicción de Naïve Bayes **51**
4.4.1 Caso 1: Aplicación en la Base de Datos Original **63**
4.4.2 Caso 2: Aplicación en la Base de Datos libre de outliers **65**
Capítulo 5. Conclusiones y Trabajo Futuro **69**
Referencias Bibliográficas **71**
Anexos **75**

Índice de Figuras

Figura 1: Diseño del proceso general de la solución propuesta 1
Figura 2: Outlier generating asymmetry 5
Figura 3: Desarrollo de la dependencia 8
Figura 4: Limpieza de datos en Data Wrangler 10
Figura 5: Análisis y consumo Hadoop de Trifacta Wrangler Enterprise 11
Figura 6: Interfaz de Trifacta Wrangler Enterprise 11
Figura 7: Descripción de Google refine a OpenRefine 12
Figura 8: Interfaz web de OpenRefine 12
Figura 9: Transformación de datos en OpenRefine 13
Figura 10: Análisis predictivo en la herramienta RStudio de R 16
Figura 11: Implementación del caso con regresión logística 18
Figura 12: Grafico de regresión logística resuelto con la función glm() 18
Figura 13: Diagrama de los elementos de la red neuronal 19
Figura 14: Implementación en RStudio de la red neuronal del caso 21
Figura 15: Modelo de la red neuronal del caso 21
Figura 16: Detección de posibles outliers por CNM 29
Figura 17: Diseño del proceso general del algoritmo DOCNM 29
Figura 18: Comparación del patrón y el texto en una posición dada 30
Figura 19: Deslizamiento del patrón en posiciones 31
Figura 20: Pseudocódigo del algoritmo KMP 31
Figura 21: Diseño general de la detección de outliers con la librería "*Outliers*" de R ... 32
Figura 22: Diagrama de dispersión 33
Figura 23: Diagrama de la desviación de la recta regresión 34
Figura 24: Interfaz gráfico de usuario con R Commander 35
Figura 25: Diseño general de la detección de outliers con la librería "*Rcmdr*" de R 35
Figura 26: Diseño del modelo de clasificación 36
Figura 27: Clínica Nuestra Señora de la Paz - Madrid 37
Figura 28: Interfaz de Inicio de OpenRefine 39
Figura 29: Datos sin normalizar en OpenRefine 40
Figura 30: Datos normalizados en OpenRefine 40
Figura 31: Datos de las variables peso y altura en OpenRefine 40
Figura 32: Datos de la variable imc en OpenRefine 41
Figura 33: Datos de input para los dos métodos de detección de outliers 41
Figura 34: Implementación de patrones detectores de posibles outliers 44
Figura 35: Pseudocódigo del algoritmo DOCNM 45
Figura 36: Código del Algoritmo DOCNM implementado en R 46
Figura 37: Resultados de la ejecución del Algoritmo DOCNM 46
Figura 38: Gráfico de la detección de outliers con la librería *Outliers* de R 47
Figura 39: Diagrama de outliersTest 48
Figura 40: Gráfico de la distancia de Cook 49
Figura 41: Gráfico de la predicción con R en el caso 1 65
Figura 42: Gráfico de la predicción con R en el caso 2 67
Figura 43: Diagrama de dispersión 83
Figura 44: Diagrama de la desviación de la recta regresión 84
Figura 45: Diagrama de outliersTest 87
Figura 46: Gráfico de la distancia de Cook 87

Figura 47: Ajuste del modelo con RCommander 88
Figura 48: Modelo lineal con RCommander 88
Figura 49: Modelos con RCommander 89
Figura 50: Diagrama de dispersión con RCommander 90
Figura 51: Scatterplot de *edad* y *i_coca*, con RCommander 90
Figura 52: Diagrama de suceso seleccionar un envase defectuoso 93

Índice de tablas

Tabla 1: Ventajas y desventajas de las herramientas de limpieza de datos 13
Tabla 2: Características de las técnicas más utilizadas en el análisis predictivo con R . 15
Tabla 3: Ventajas y desventajas de las técnicas de análisis predictivo con R 22
Tabla 4: Base de datos de drogodependientes 38
Tabla 5: Frecuencia de consumo del paciente drogodependiente 38
Tabla 6: Clasificación de variables drogodependientes 42
Tabla 7: Codificación de datos de los pacientes drogodependientes..... 43
Tabla 8: Patrones detectores de posibles outliers 44
Tabla 9: Clasificación de datos para DO con la librería "*Outliers"* y *"Rcmdr"* de R 50
Tabla 10: Clasificación de datos históricos de pacientes con drogodependencias 58
Tabla 11: Datos históricos sin la presencia de Outliers..... 60
Tabla 12: Instancia nueva para predecir con el algoritmo Naïve Bayes..... 62

Capítulo 1. Introducción

Actualmente la drogodependencia constituye un problema de salud pública de primer orden en nuestra sociedad [1]. Una de las problemáticas que atraviesa el Centro de Atención Integral al Cocainómano de la Agencia de Salud de Madrid [2], se relaciona con los gastos administrativos que ocasionan los tratamientos no exitosos. En este trabajo se propone una solución a este tipo de problemas mediante el análisis predictivo de los datos históricos y se aplica a datos de pacientes drogodependientes con la intención de predecir si el tratamiento será eficiente o no en un paciente concreto. Esta solución puede ayudar a la toma de decisiones del médico a cargo del internamiento de nuevos pacientes drogodependientes, optimizando en cierta medida la inversión en este tipo de tratamientos.

Para poder realizar la predicción con éxito, es necesario realizar una fase de investigación sobre el tratamiento de los datos. Los datos proporcionados para este estudio provienen directamente de anotaciones tomadas por el médico o los auxiliares en distintos centros, mediante notas escritas, muchas veces incompletas y sin ningún tipo de formato establecido. El análisis directo de estos datos es inviable sin un pre-procesado adecuado. El sistema que hemos seguido en este trabajo para el pre-procesado (limpieza e integración) de los datos, es de interés para cualquier campo de aplicación hoy en día. La Figura 1 muestra el proceso general de la solución propuesta.

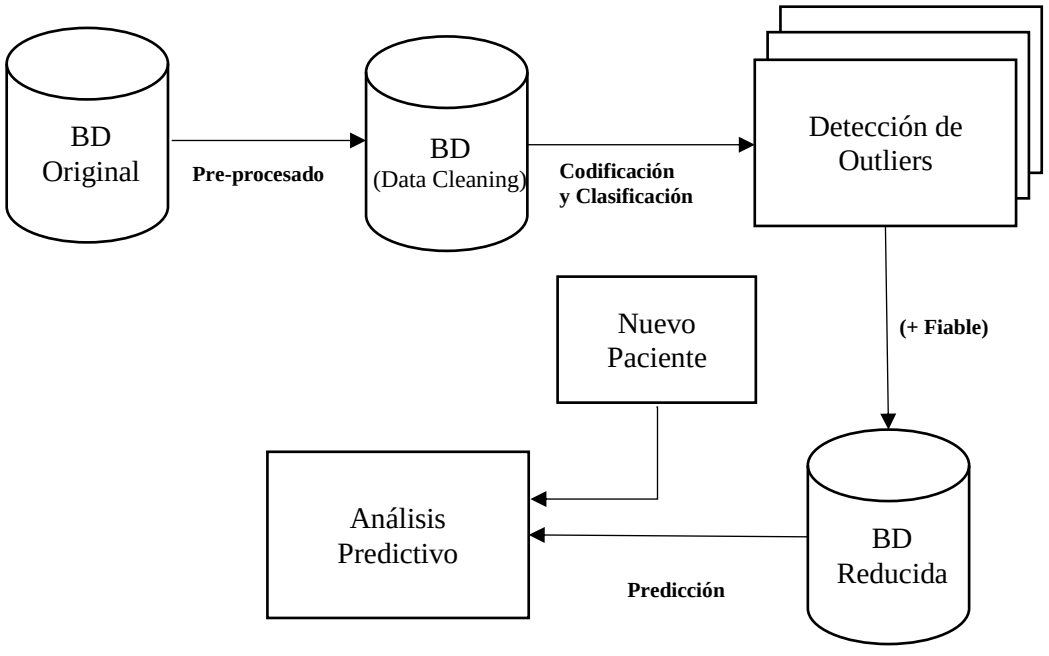


Figura 1.- Diseño del proceso general de la solución propuesta.

En primer lugar, el sistema recoge la base de datos original con datos de pacientes drogodependientes procedentes de distintas fuentes y como consecuencia con distintos formatos. En un segundo lugar, se realiza un pre-procesado para limpiar los datos, como normalizar los nombres y tipos de los campos. En un tercer lugar, se genera una base de datos preparada para realizar los procesos de codificación y clasificación, así como también datos de input para el procedimiento de detección de datos atípicos (outliers). Finalmente, se implementa el análisis predictivo sobre una base de datos reducida y más fiable con el propósito de obtener la predicción del resultado del tratamiento drogodependiente de un nuevo paciente.

1.1 Motivación

Teniendo en cuenta la gran cantidad y diversidad de datos que actualmente se utilizan para los análisis predictivos y la importancia de estos análisis en todo tipo de problemas (muy especialmente en el área de ciencias de la salud), nuestra motivación se basa en la calidad y fiabilidad que se requiere en los datos, antes de que estos puedan someterse a análisis, evitando resultados erróneos. Es un hecho bien conocido que cuando se combinan diversas fuentes de información el ruido que contienen se ve incrementado y aún más cuando la cantidad de datos es masiva, como ocurre hoy en día en proyectos big data. Es más, en algunos casos las bases de datos que se recogen para estudio pueden estar corrompidas por actuaciones ilícitas o malintencionadas. Por estas razones, una base de datos sin un pre-procesado adecuado puede producir patrones artificiales y falsas relaciones que pueden conducir a una toma de decisiones peligrosa. En este trabajo, estudiamos el pre-procesado de datos con fases de limpieza e integración y lo aplicamos sobre una base de datos de drogodependientes.

El problema de la drogodependencia es realmente de gran magnitud a escala mundial ya que existen cifras impactantes [3]. Con una esperanza de vida cada vez mayor, los tratamientos médicos están cambiando rápidamente, orientándose especialmente a los tratamientos individualizados y muchas de las decisiones detrás de esos cambios están siendo impulsados por los datos. El objetivo ahora es entender al máximo las características personales del paciente tanto como sea posible, desde el principio de su vida, con la esperanza de que las señales de advertencia de una enfermedad grave en una etapa lo suficientemente temprana sea posible y que el tratamiento sea mucho más sencillo (y menos costoso) [4, 5].

El interés por el uso de datos masivos en ciencias de la salud aumenta cada día debido a la mejora que aportan en procesos de predicción y detección de patrones, entre otros [6-11].

En este trabajo proponemos un método propio (Detección de Outlier mediante Cadenas No Monótonas). Esto facilitara enormemente en el pre-procesado de datos (Data Cleaning) y por ende en la calidad y fiabilidad que se requiere en los datos para un posterior análisis predictivo.

1.2 Objetivos

Este Trabajo Fin de Master tiene los siguientes objetivos:

- ✓ Aprender técnicas actuales de pre-procesado de datos, realizando una limpieza e integración de datos en la base de datos de pacientes drogodependientes.
- ✓ Aprender técnicas de análisis predictivo y su aplicación a las ciencias de la salud.
- ✓ Mejorar las técnicas estudiadas mediante la limpieza de datos atípicos implementando dos métodos tradicionales en R y un método propio basado en cadenas no monótonas y el algoritmo rápido KMP. Este objetivo es la principal aportación de este trabajo.
- ✓ Aplicando las técnicas aprendidas y las desarrolladas, obtener una base de datos fiable sobre la que implementar un análisis predictivo (en la base de datos reducida libre de ruido y outliers) para la toma de decisiones sobre el tratamiento de drogodependientes.

1.3 Trabajos Relacionados

Existen diversos estudios e investigaciones en el área de este trabajo. A continuación, se resumen los artículos estudiados relacionados por temas.

Integración y Limpieza de datos

En el artículo [12] M. Lenzerini. (2002), se explica brevemente el problema de combinar datos que residen en diferentes fuentes de información y muestra al usuario una visión unificada. Así mismo, explica el problema del diseño de sistemas de integración de datos y expone su importancia en aplicaciones en el mundo actual. También en este artículo se presentan instrucciones a modo de tutorial sobre la integración de datos centrándose en algunas de las cuestiones teóricas que son más relevantes, prestando una especial atención a los siguientes aspectos: modelado de una aplicación de integración de datos, procesamiento de consultas en integración de datos, relaciones con fuentes de datos inconsistentes y el razonamiento en consultas.

Además, en los artículos [13] Stonebraker, M (2013) y [14] S. Kandel (2015) los autores desarrollan ideas para hacer una integración automática de datos. En el primer artículo, que hace referencia a la aplicación Data Tamer [13], los autores del M.I.T, Brandeis e Instituto de investigación de computación de Qatar (QCRI), describen Data Tamer como un sistema que espera como entrada una secuencia de datos donde una nueva fuente se somete a algoritmos de machine learning para realizar identificación de atributos, agrupación de atributos en tablas, transformación y duplicación de datos.

También, Data Tamer incluye un componente de visualización de datos para que los usuarios puedan examinar y poder realizar manualmente cualquier cambio necesario. En el segundo artículo, que hace referencia a la aplicación Wrangler [14], los autores de la Universidad de Stanford y de la Universidad de California, muestran como Wrangler aprovecha los tipos de datos semánticos (por ejemplo, lugares geográficos, fechas, códigos de clasificación) para facilitar la validación y conversión de tipo. Los resultados del estudio muestran que Wrangler reduce significativamente el tiempo de especificación y promueve el uso robusto, auditable, en lugar de una edición manual. La debilidad de estos sistemas es que requieren la supervisión de un analista de datos por ser herramientas de uso general.

Actualmente también se está investigando en la línea de integración de datos semiautomática, reduciendo al máximo la supervisión por parte del científico de datos, en el sentido semántico que ya se ha presentado en Wrangler y haciendo uso de la naturaleza propia de los datos. Un trabajo en este aspecto es el desarrollado por Pavel Llamocca [15] en 2016.

En este trabajo, se ha utilizado OpenRefine como herramienta de integración y limpieza de datos. Data Wrangler y OpenRefine son herramientas muy similares, pero tienen objetivos diferentes y en cada uno de ellos se puede ahorrar mucho tiempo para el pre-procesado de los datos.

Detección de Outliers

Los valores atípicos (en inglés outliers) son valores que pueden encontrarse en las bases de datos con pequeña probabilidad pero que distorsionan las medidas notablemente. Su detección es fundamental para asegurar buenos resultados en los análisis. Por ejemplo, en el artículo [16] los autores describen un estudio que revela que revela dificultades para hacer frente a los valores atípicos. Los autores muestran que la detección de outliers mediante el tradicional intervalo centrado en la media, y de amplitud tres desviaciones estándar no es la mejor solución a pesar de ser de uso muy frecuente. La media y la desviación estándar son particularmente sensibles a los outliers y por ello los autores destacan las ventajas de otras medidas como la desviación media absoluta, una medida alternativa y más robusta de la dispersión que es igualmente fácil de implementar. También explican los procedimientos para el cálculo de este indicador en SPSS y el software R.

En la Figura 2, los autores consideran brevemente un caso ficticio donde se incluye un mayor número de observaciones. La Figura 2a muestra una distribución normal e informa de la media, la desviación estándar y la mediana. La Figura 2b muestra la misma distribución, pero con un valor (= 0,37) cambiado en un outlier (= 3). Podemos ver que la media y la desviación estándar han cambiado drásticamente mientras que la mediana sigue siendo la misma.

En este trabajo, se ha realizado la detección de outliers como mejora en el pre-procesado de los datos. Mediante la aplicación de un método propio (Detección de

Outliers mediante Cadenas No Monótonas) y dos métodos de la Librería de R (*Outliers* y *Rcmdr*).

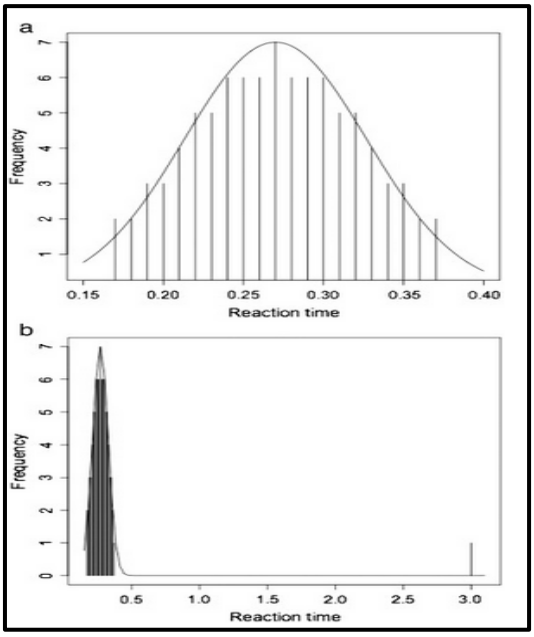


Figura 2. Outlier generating asymmetry.
a) Normal distribution, n = 91, mean = 0.27, median = 0.27, standard deviation = 0.06.
b) Asymmetry due to an outlier, n = 91, mean = 0.39, median = 0.27, standard deviation = 0.59.
Fuente: [16]

Análisis de datos médicos

En el artículo [17] los autores describen una investigación sobre el trastorno bipolar y el hecho de que a menudo conduce a períodos de baja por enfermedad causando problemas económicos y sociales en entornos familiares y de trabajo. Asimismo, los autores insisten que la mayor parte de estas crisis podrían ser evitadas si se lograra una predicción temprana. Así mismo, en este artículo se muestra la especificación de un sistema de predicción basado en datos de diversas fuentes obtenidos por monitorización con acelerómetros, testeos diarios mediante smartphones y datos clínicos recogidos en la consulta médica.

En este trabajo, también se viene utilizando Datos médicos (pacientes drogodependientes). Así mismo se requiere de un sistema de predicción temprana para poder predecir si el tratamiento de un nuevo paciente tendrá éxito o no.

1.4 Estructura del Trabajo

Esta memoria se organiza en cinco capítulos. En el capítulo 2 se describe la visión general de los temas realizados. También se exponen las normas y legislaciones vigentes en materia de protección de datos de carácter personal dada la importancia de esta temática en el caso de datos médicos.

En el capítulo 3 se exponen los procedimientos de detección de valores atípicos (outliers) y sobre el análisis predictivo de los datos, con la finalidad de hacer una predicción en el tratamiento drogodependiente.

En el capítulo 4 se exponen los datos utilizados en este trabajo y sus procesamientos previos para ejemplificar su tratamiento. Seguidamente se detallan las pruebas realizadas en la detección de posibles outliers. Así mismo se exponen las pruebas realizadas para el análisis predictivo con la librería “e1071” de R.

El capítulo 5 presenta las conclusiones y se describen las posibles líneas de trabajo futuro que conduzcan a enriquecer la investigación realizada.

Finalmente, se da una lista de referencias y anexos donde se pueden ver los detalles de la implementación de los procesos desarrollados.

Capítulo 2. Visión General

En este capítulo se muestran los conceptos más importantes sobre drogodependencia, data cleaning y análisis predictivo, por ser temas fundamentales para la comprensión y el desarrollo del trabajo realizado en esta memoria. También se muestran algunas de las técnicas más utilizadas en la limpieza de los datos (Data Wrangler, Trifacta Wrangler, Enterprise, Open Refine) y en el análisis predictivo con R (Naïve Bayes Classifier, Regresión Logística y Redes Neuronales). Así mismo se exponen las normas y legislaciones vigentes en materia de protección de datos de carácter personal.

2.1 Drogodependencia

La Organización Mundial de Salud (OMS) define la drogodependencia como "estado psíquico y en ocasiones también físico, debido a la interacción entre un organismo vivo y una droga y que se caracteriza por modificaciones del comportamiento y por otras reacciones, entre las que siempre se encuentra una pulsión a ingerir droga de forma continua o periódica con objeto de volver a experimentar sus efectos psíquicos y en ocasiones evitar su malestar en su abstinencia" [3].

La Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC) define la drogodependencia como una afección compleja y multifactorial en que intervienen factores de orden individual, cultural, biológico, social y ambiental. Uno de los principales obstáculos que impiden el tratamiento y la atención es el estigma y la discriminación implícitos en este trastorno de la salud que es tratable. El tratamiento de la drogodependencia requiere la aplicación de un enfoque integral y multidisciplinario con intervenciones tanto farmacológicas como psicosociales [18].

Tipos de dependencia

De acuerdo con UNODC básicamente existen dos tipos de dependencias [18]: dependencias físicas y dependencias psíquicas. La naturaleza de los datos es un conocimiento valioso para el analista antes de la realización de cualquier estudio por lo que es interesante explicar cada una:

Dependencia física. - Necesidad de mantener determinados niveles de una droga en el organismo. Tiene dos componentes: tolerancia y síndrome de abstinencia aguda, por ejemplo, sustancias depresógenas (alcohol, opiáceos, hipnóticos y sedantes).

- **Tolerancia:** Es la necesidad de cantidades crecientes de una sustancia en busca del efecto deseado o disminución del efecto ante una misma dosis.

- **Síndrome de abstinencia aguda:** Manifestaciones clínicas, psíquicas o físicas que se producen por el cese de la administración de una droga y desaparecen con la administración de la droga.

Dependencia psíquica. - Deseo irresistible o anhelo de repetir la administración de una droga para obtener la vivencia de sus efectos agradables, placenteros, evasivos o ambos para evitar el malestar psíquico que se siente con su ausencia, por ejemplo, sustancias psicoestimulantes (anfetaminas, cocaína, nicotina) y alucinógenos.

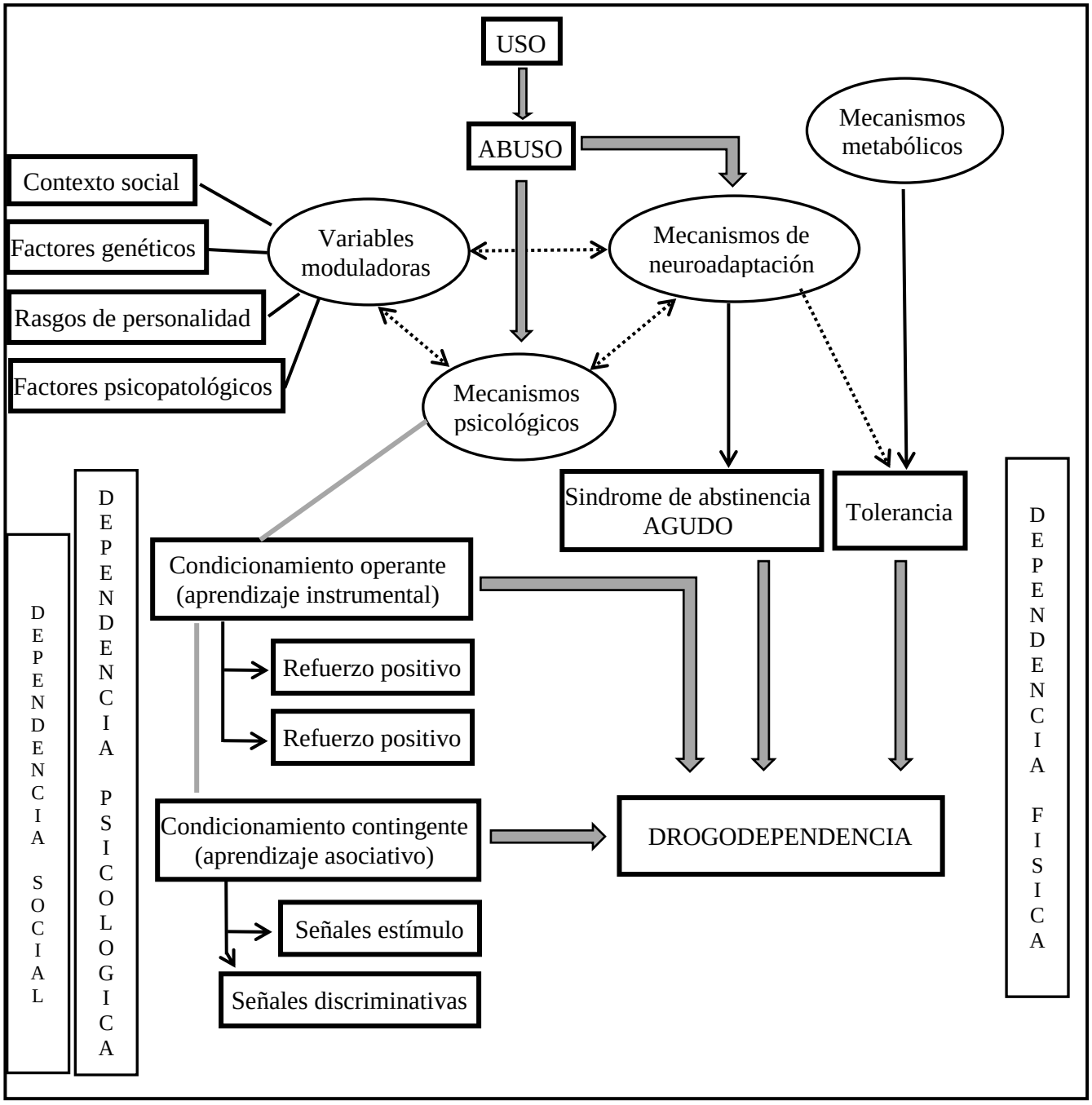


Figura 3. Desarrollo de la dependencia
Fuente: [18]

2.2 Data Cleaning

Data Cleaning o la limpieza de datos es uno de los pasos que consume más tiempo para conseguir un formato estructurado previo al análisis de los datos. Si los datos contienen ruido, éste puede causar confusión y resultados erróneos. Esta rutina de trabajo "limpiar" habitualmente incluye procedimientos de uso muy generalizado en la literatura, como el completado de valores ausentes, alisado de datos con ruido, la identificación y eliminación de valores atípicos y la resolución de inconsistencias. En la literatura existen algoritmos maduros para la resolución de estos procesos [19].

Otro paso importante en el proceso de limpieza de datos, es la detección de discrepancias. Las discrepancias pueden ser causadas por varios factores, incluyendo las formas de entrada de datos mal diseñados que tienen muchos campos opcionales, errores humanos en la entrada de datos, errores deliberados. Las discrepancias también pueden surgir de las representaciones de datos inconsistentes y uso inconsistente de códigos. Otras fuentes de discrepancias incluyen errores en los dispositivos de instrumentación que registran los datos y errores del sistema [20]. Para estos casos, las nuevas tecnologías en big data proporcionan la oportunidad de utilizar nuevas fuentes de información integrándolas mediante pre-procesado de datos: collecting, cleaning, integration son palabras clave que no podemos obviar si queremos que los resultados de los análisis estadísticos tradicionales sean fiables [21].

A continuación, se muestra algunas herramientas que utilizan distintas técnicas en el pre-procesado de los datos. Data Wrangler, Trifacta Wrangler Enterprise y Open Refine. Finalmente, seleccionamos una de ellas por ser una herramienta libre, de escritorio y su alto grado de robustez, basándonos en un análisis de ventajas y desventajas.

2.2.1 Data Wrangler

Es un servicio basado en la web del Grupo de Visualización de la Universidad de Stanford. Los autores del proyecto son: Sean Kandel, Andreas Paepcke, Joseph Hellerstein y Jeffrey Heer [22]. Wrangler [23] una herramienta interactiva para la limpieza y transformación de datos. Así mismo reduce el tiempo y esfuerzo en la estructuración de datos y la evaluación de los problemas de calidad de datos. Para ello, implica combinar la manipulación directa de los datos visualizados con la deducción automática de transformaciones relevantes. De esta manera permite al analista explorar de forma iterativa el espacio de operación aplicable y una vista previa de sus efectos.

Wrangler aprovecha los datos de tipos semánticos (por ejemplo, ubicaciones, fechas, códigos de clasificación geográficas) ayudando en la validación y el tipo de conversión. Actualmente el proyecto Wrangler realizado en Stanford, se encuentra ya en fase comercial en la empresa Trifacta Wrangler Enterprise, aunque todavía existe una versión de prueba gratuita [24].

En la Figura 4 se muestra la interfaz del proceso de limpieza de los datos de Wrangler. El panel izquierdo contiene (a partir de arriba hacia abajo) historial de conversiones, un menú de selección convertir y sugerencias automáticas en base a la selección actual. El texto en negrita dentro de las descripciones convertir, indican los parámetros en los que se puede hacer clic y revisar. El panel de la derecha contiene una tabla de datos genérico y por encima de cada columna se encuentra un medidor de calidad de los datos.

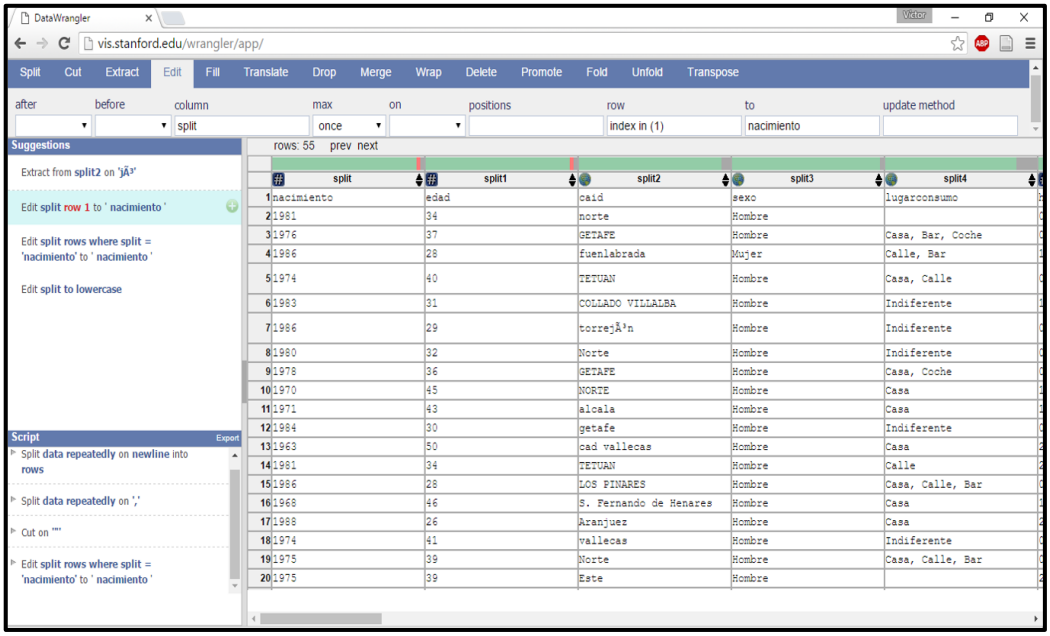


Figura 4. Limpieza de datos en Data Wrangler.

2.2.2 Trifacta Wrangler Enterprise

Este software ha existido desde 2012 y como ya se ha comentado, nació de un proyecto conjunto de investigación, entre el Grupo de Visualización la Universidad de Stanford y la Universidad de Berkeley, llamado Data Wrangler. Existen tres fundadores: Sean Kandel, un estudiante del doctorado de la Universidad de Stanford y dos profesores universitarios, Joe Hellerstein y Jeffrey Heer [25].

Trifacta Wrangler Enterprise [24] proporciona un flujo de trabajo optimizado para la transformación de los datos a escala. La solución permite a los usuarios la visualización del contenido de los datos almacenados en Hadoop e interactuar con ese contenido. Así mismo define reglas de transformación cuando se ejecuta un trabajo Hadoop (usando Spark o MapReduce) para procesar e imprimir los datos en la forma deseada para su análisis. Por otro lado, está diseñado para ayudar a los analistas de datos a realizar un trabajo asociado en la preparación de los datos sin tener que escribir manualmente el código.

El objetivo de Trifacta Wrangler Enterprise es tomar grandes conjuntos de datos y hacer muy sencillo el análisis y consumo de Hadoop para personas sin conocimientos técnicos, (ver Figura 5)

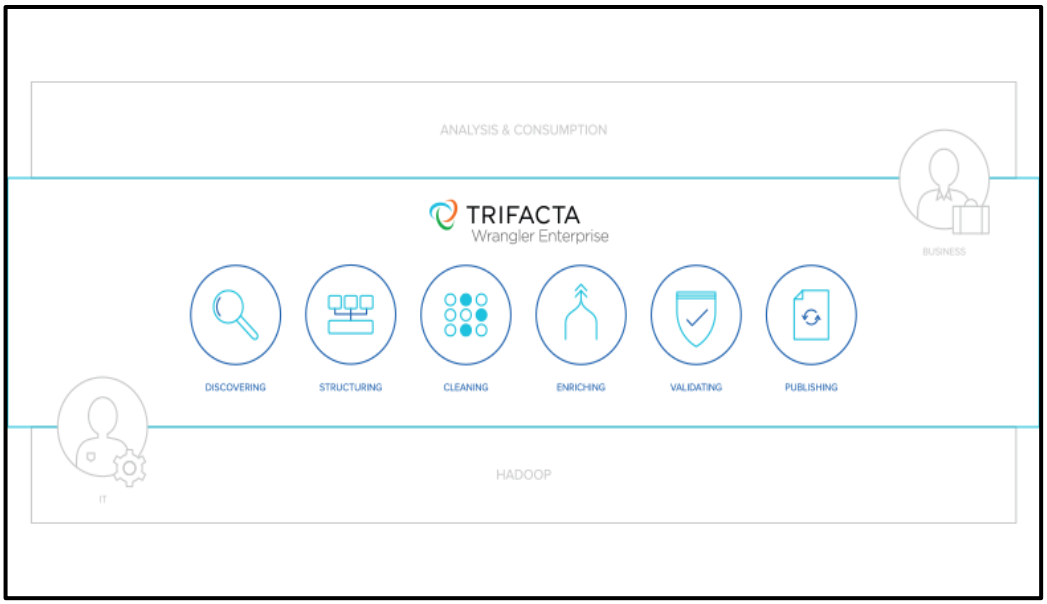


Figura 5. Análisis y consumo Hadoop de Trifacta Wrangler Enterprise. Fuente: [24]

En la Figura 6 se muestra la interfaz de ejecución de Trifacta Wrangler Enterprise.

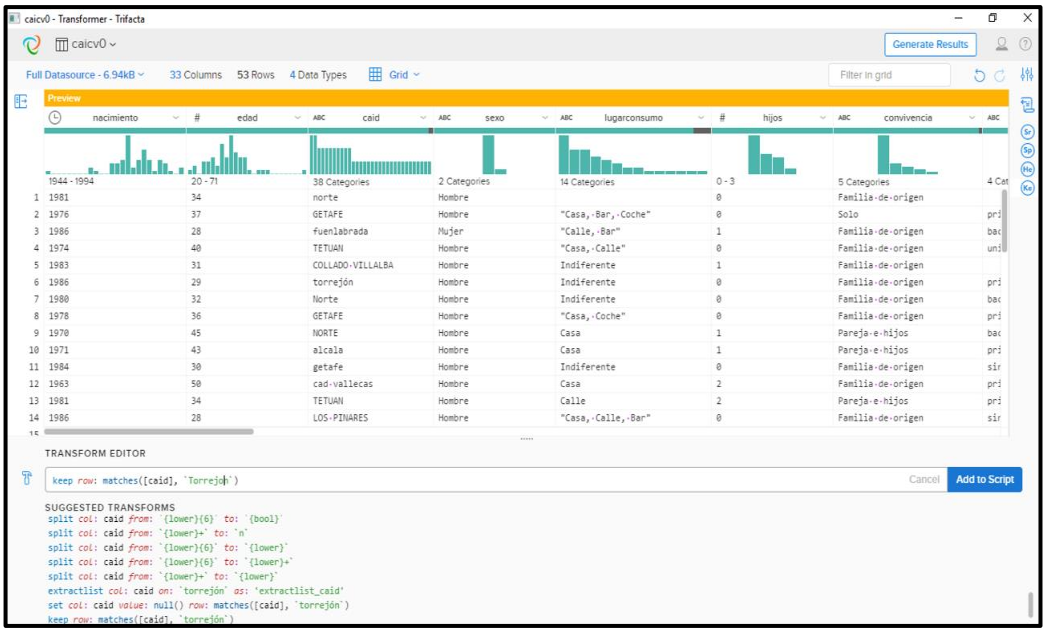


Figura 6. Interfaz de Trifacta Wrangler Enterprise.

2.2.3 OpenRefine

Este software [26] fue creado por Metaweb Technologies y originalmente escritos y concebidos por David Huynh. Fue adquirida por Google en julio de 2010 y el producto pasó a llamarse Google Refine (Figura 7). En octubre de 2012, se renombró a OpenRefine como transición a un producto de apoyo comunitario. OpenRefine es un potente software de limpieza de datos y su objetivo principal es ayudar a entender la estructura y calidad de los datos, permitiendo corregir determinados tipos de errores comunes en ellos. Su enfoque en la “limpieza de datos” es sencillo e intuitivo, por lo

que OpenRefine es una herramienta muy potente para trabajar con grandes conjuntos de datos. Así mismo, utiliza el navegador web como interfaz, pero no necesita conexión a Internet para su ejecución.

Cabe mencionar que desde el 2 de octubre de 2012 Google no está apoyando activamente a OpenRefine. El desarrollo del proyecto, documentación y promoción ahora son totalmente apoyados por voluntarios (OpenRefine es una herramienta libre).

Figura 7. Descripción de Google refine a OpenRefine.
Fuente: [26]

En la Figura 8 se muestra el interfaz web de la poderosa herramienta de escritorio OpenRefine.

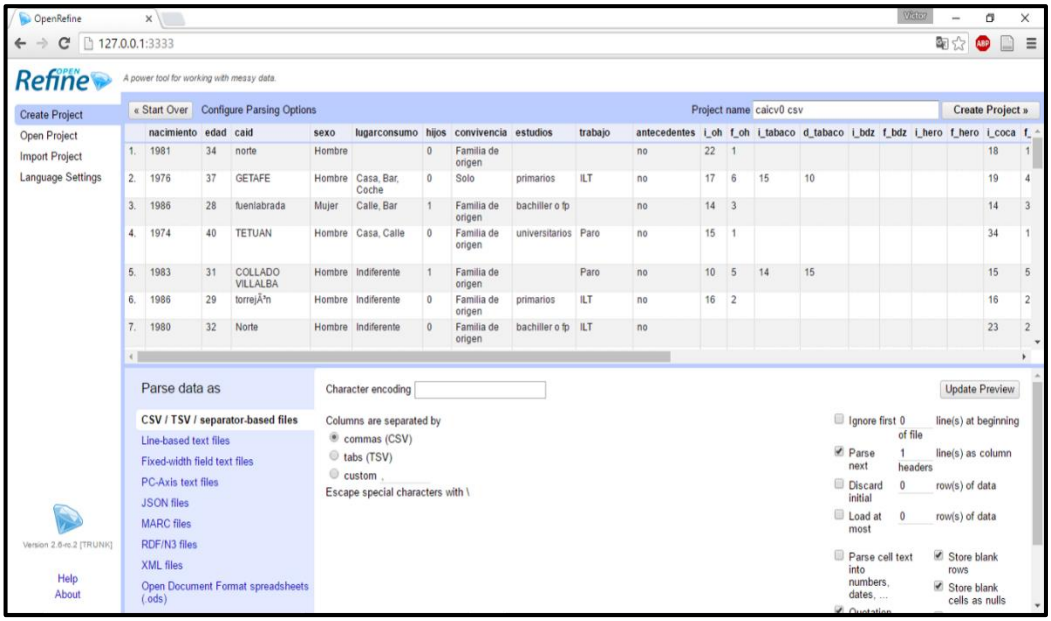


Figura 8. Interfaz web de OpenRefine.

En la Figura 9 se muestra el interfaz de OpenRefine. En este caso se muestra la transformación de datos. Como ejemplo el nombre de la ciudad del paciente drogodependiente (la letra inicial en mayúscula seguida de minúsculas)

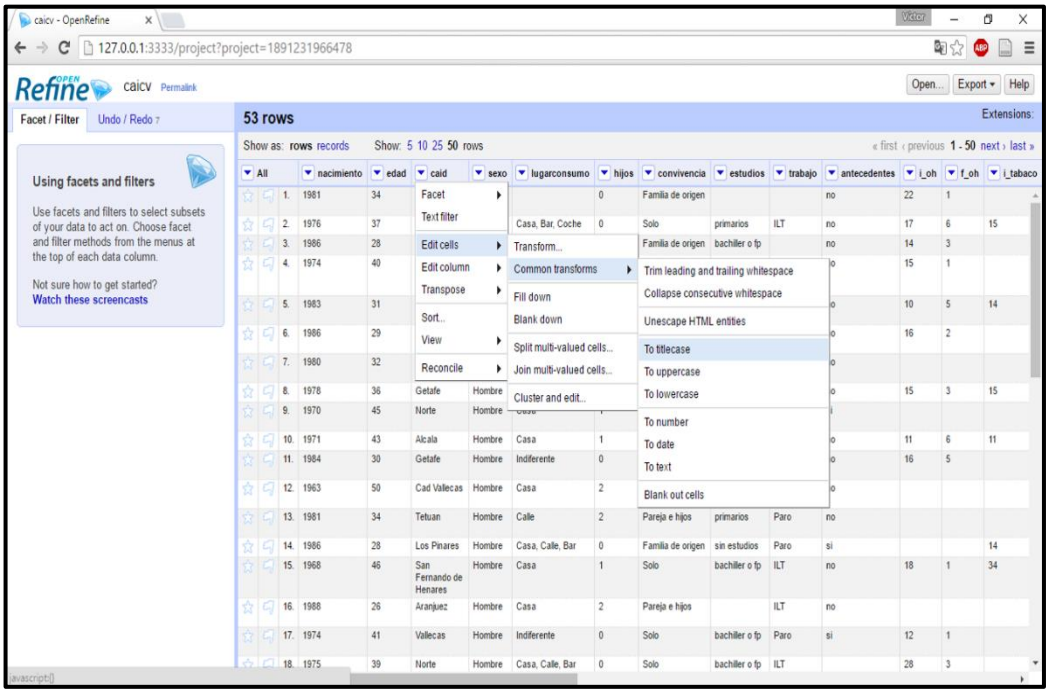


Figura 9. Transformación de datos en OpenRefine.

Estas herramientas son similares, pero tienen diferentes objetivos, y con cada uno de ellos se puede ahorrar mucho tiempo. En la Tabla 1 se muestra las ventajas y desventajas de cada software.

Tabla 1. Ventajas y desventajas de las herramientas de limpieza de datos.

Software	Ventajas	Desventajas
Data Wrangler	• Es un software gratuito. • Es una aplicación web, que no requiere de mucho espacio en disco. • ***Reorganiza los datos en vistas fáciles de entender.*** **-** Ofreciendo 14 opciones diferentes para dar formato a su conjunto de datos. Estas funciones permiten dividir, fusionar, eliminar, plegar, desplegar filas o columnas, así como también cortar, extraer y editar los datos. • Admite los scripts de salida en dos idiomas: Python (para cálculo de datos en la parte final) y JavaScript (en caso de que desee transformar en el navegador, o el uso de Node.js).	• Los servicios web dependen del equipo de desarrollo así lo requieran o por fallos en los servidores. • Los datos están en la aplicación web y se puede rastrear cualquier actividad que el usuario haga. Esto puede traer problemas de privacidad de los datos. • Actualmente el proyecto solo está disponible a través de su versión comercial gratuita a través de Trifacta Wrangler Enterprise.

Trifacta Wrangler Enterprise	• Es una aplicación de escritorio y se beneficia de actualizaciones y metadatos a través de una conexión de internet. • Cuenta con un foro de ayuda de la comunidad Wrangler, desde su página web oficial (https://www.trifacta.com/support/community/). • ***Acelera el proceso de análisis*****.** - Aprovechar un enfoque más eficiente, intuitivo y visual para la preparación de datos para la visualización. • ***Permite a los analistas explorar y transformar los datos*****.** – para todas las estructuras y tamaños en Hadoop sin necesidad de escribir código. • Permite integrar varias fuentes de datos como también trabajar con grandes conjuntos de datos.	• Es un software comercial. • Para utilizar el software necesariamente tenemos que estar conectados a Internet. • Está solo disponible para Windows y Mac. No se encuentra disponible para Linux.
OpenRefine	• Es un software gratuito y se encuentra como proyecto abierto (https://github.com/OpenRefine). • Es un aplicativo de escritorio, con una interfaz web. Funciona como ejecutable sobre cualquier navegador web y está disponible para Windows, Mac y Linux. • Cuenta con la ayuda de la comunidad OpenRefine (http://openrefine.org/community). • Ayuda a entender la estructura y la calidad de los datos, y permite corregir determinados errores comunes. • Posee dos opciones más comunes, tales como: ***-Limpieza de datos.*** **–** nos permite realizar cambios en los contenidos de celdas y unificación de los campos, de forma manual o sugerido por el propio software (brinda sugerencias de manera automática como optimizar nuestros datos). **- *Transformación de Datos.* –** nos permite dividir columnas, crear nuevas columnas según el valor de otra. Así mismo toma parte del contenido una columna para crear otra nueva, etc.	• Actualmente el proyecto ya no es financiado por Google.

Para este trabajo se ha utilizado OpenRefine como herramienta en el proceso de limpieza de datos. Para ello se han tomado en cuenta dos características fundamentales: ser una aplicación de escritorio y de uso libre.

2.3 Análisis Predictivo

El análisis predictivo se apoya en la analítica y estadística para hacer predicciones sobre sucesos futuros. Básicamente se construyen modelos para inferir información sobre una muestra. Para ello se utilizan varias técnicas de minería de datos y Machine Learning.

En el Trabajo de Fin de Grado [27] los autores lo definen de la siguiente manera. "El análisis predictivo utiliza estadística junto con algoritmos de minería de datos. Se basa en el análisis de los datos actuales e históricos para hacer predicciones sobre futuros eventos. Estas predicciones raramente suelen ser afirmaciones absolutas, pareciéndose más a eventos y su probabilidad de que suceda en el futuro". Así mismo mencionan tres tipos de análisis predictivo: los modelos predictivos, modelos descriptivos y modelos de decisión.

En este apartado se muestran las características de las técnicas más utilizadas en el análisis predictivo con R (Tabla 2). Posteriormente se explica cada una de ellas y se exponen las ventajas y desventajas de las mismas (Tabla 3), seleccionando la más adecuada para este trabajo.

Tabla 2. Características de las técnicas más utilizadas en el análisis predictivo con R.

Nº	Técnica	Paquete de R	Método	Tipo de método	Argum. Value
1	Naïve Bayes	"e1071"	Naïve Bayes Classifier	Clasificación	naiveBayes
2	Regresión Logística	"caret"	Generalized Linear Model	Dual (Clasificación-regresión)	glm
3	Redes Neuronales	"neuralnet"	Training of neural networks	Dual (Clasificación-regresión)	neuralnet

2.3.1 Naïve Bayes Classifer

Es una función del paquete "e1071" de R. Calcula la probabilidad condicional a posteriori de una variable de clase categórica dadas las variables predictoras independientes utilizando la regla de Bayes. [28].

La técnica del algoritmo de clasificación Naïve Bayes, se basa en el teorema de Bayes y es adecuado cuando la dimensionalidad de los datos de entrada es grande. Así mismo Naïve Bayes es uno de los algoritmos de clasificación más simples, pero es muy utilizado por ser muy preciso. En primera instancia, Naïve Bayes trata de clasificar casos basados en las probabilidades de atributos o casos antes vistos, asumiendo completamente la independencia de atributos.

En la Figura 10 se muestra la codificación del análisis predictivo del clasificador Naïve Bayes, en la herramienta RStudio de R.

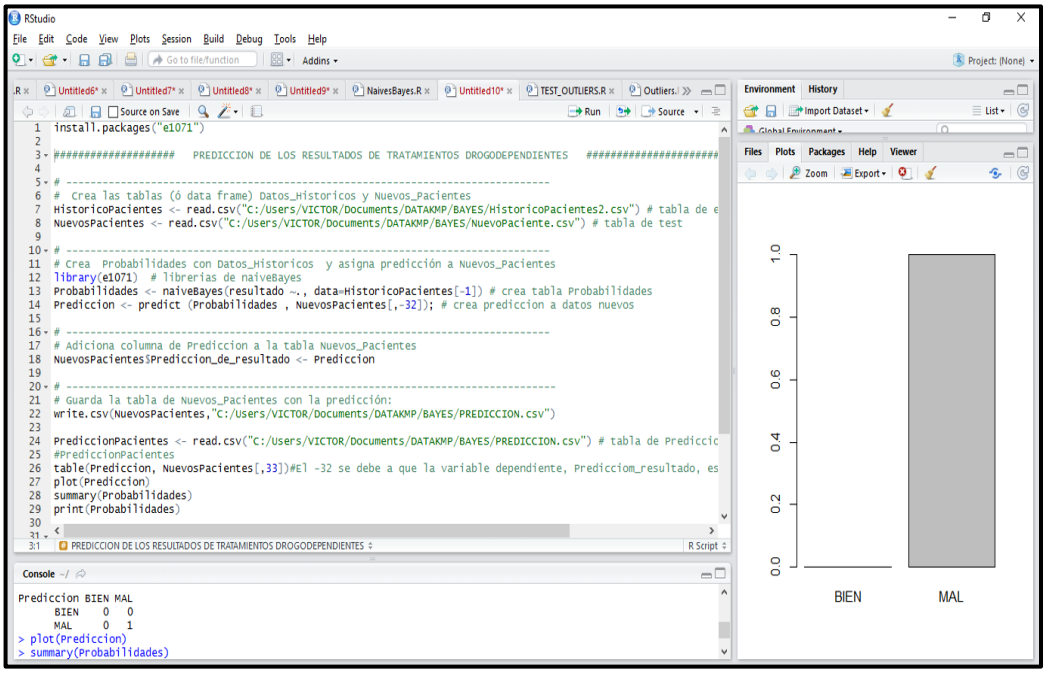


Figura 10. Análisis predictivo en la herramienta RStudio de R.

2.3.2 Regresión Logística

Es una técnica multivariante, en la que la variable dependiente es categórica y las variables independientes son de cualquier naturaleza, es decir, cuantitativas o cualitativas. Esta técnica es muy utilizada en investigación clínica y epidemiologia, pero también se aplica en otras áreas del conocimiento [29].

Para el caso de regresión logística con R, en este trabajo se ha utilizado el paquete *caret*, que contiene numerosas herramientas para desarrollar y construir modelos predictivos usando otros paquetes de R [30]. Por ejemplo, contiene funciones para pre-procesamiento de datos, construcción y evaluación de modelos de diferentes parámetros, predicción de nuevas observaciones, evaluación de la importancia de las variables, visualización de modelos y selección de variables. En este trabajo hemos utilizado el modelo lineal generalizado (Generalized Linear Model).

Para ajustar un modelo lineal generalizado, la función genérica que se usa en R es *glm*. Cuyos argumentos pueden consultarse mediante la instancia siguiente:

```
args(glm)
function (formula, family = gaussian, data, weights, subset, na.action, start = N
ULL, etastart, mustart, offset, control = list(...), model = TRUE, method = "glm.
fit", x = FALSE, y = TRUE, contrasts = NULL, ...)
NULL
```

El argumento “formula” es ampliamente usado en la modelización con R y permite una sintaxis comprensible para expresar relaciones entre variables. La sintaxis de formula describe la relación entre la variable respuesta y las variables predictoras.

En lenguaje R [31], cuando la variable dependiente es binaria debe venir expresada como un booleano (0 FALSE, 1 TRUE), o como un factor, en cuyo caso la primera categoría representa los fracasos y la otra los éxitos. La primera categoría es la que tenga el menor número de código, o la primera en orden alfabético si se ha creado el factor a partir de una variable de tipo carácter. La categoría de referencia se puede cambiar, bien recodificando la variable dependiente o, utilizando funciones de R como *relevel*.

También se puede considerar el caso general de una variable dependiente binomial, dónde la variable dependiente es el número de éxitos en uno o más ensayos. En este caso la sintaxis de *glm* varía levemente, como en el ajuste del modelo con datos agrupados.

Un ejemplo típico de la sintaxis de la función glm es.

```
glm(y ~ x, family = binomial, data=mis.datos)
```

Donde "y" es una variable discreta con valores 0, 1 y "x" una variable continua o categórica. Estas variables están en el data frame "mis.datos". Se ha especificado la familia binomial, la cual toma por defecto la función *logit* como función de enlace.

La función *glm* es la más utilizada para ajustar modelos lineales generalizados, si bien también existen alternativas en algunos paquetes desarrollados por la comunidad de R, tales como la función *lrm* en el paquete *rms* o *vglm* del paquete VGAM, o se pueden crear funciones propias implementando algún algoritmo iterativo de ajuste, como el de Newton-Raphson. En el apéndice A se pueden ver algunas funciones sencillas para ajustar el modelo mediante otros procedimientos.

A continuación, se implementa en la herramienta RStudio de R, el siguiente caso: de la base de datos de pacientes drogodependientes se requiere predecir las probabilidades de éxito del tratamiento para pacientes cuyo tiempo de consumo de cocaína están entre 5 y 10 años.

Para ello se han utilizado las variables de i_coca y edad. Con el propósito de implementar tres variables (TiempoConsumoCoca, TiempoSinConsumoCoca y Exito_Tratamiento), las dos primeras calculadas y la tercera estimada.

```
Drogodependientes <- read.csv ("C:/Users/PruebaCAIC.csv", header=T)
head(Drogodependientes)
TiempoSinConsumoCoca<- Drogodependientes$edad - Drogodependientes$i_coca
TiempoConsumoCoca<- Drogodependientes$i_coca
(Exito_Tratamiento <- TiempoConsumoCoca/(TiempoConsumoCoca +
TiempoSinConsumoCoca))
```

Seguidamente optamos por especificar la variable dependiente como una matriz de dos columnas.

```
formula<-cbind(TiempoConsumoCoca,TiempoSinConsumoCoca)~TiempoConsumoCoca
RegresionLogistica <- glm(formula, family=binomial)
summary(RegresionLogistica)
```

La función para predecir con el modelo se instancia de la siguiente forma:

```
FuncionPredecir <- function(x) predict(RegresionLogistica,
newdata=data.frame(TiempoConsumoCoca=x), type="response")
```

De acuerdo con el modelo, las probabilidades de éxito del tratamiento para pacientes cuyos tiempos de consumo de cocaína están entre 5 y 10 años se obtienen en la variable:

```
FuncionPredecir(c(5,10))
```

Los datos se pueden visualizar mediante plot:

```
plot(TiempoConsumoCoca, Exito_Tratamiento, pch=19, col="red")
```

Sobre la imagen podemos visualizar la curva correspondiente al modelo:

```
curve(FuncionPredecir, col="blue", add=T)
```

Los resultados del caso se muestran en la Figura 11 y 12.

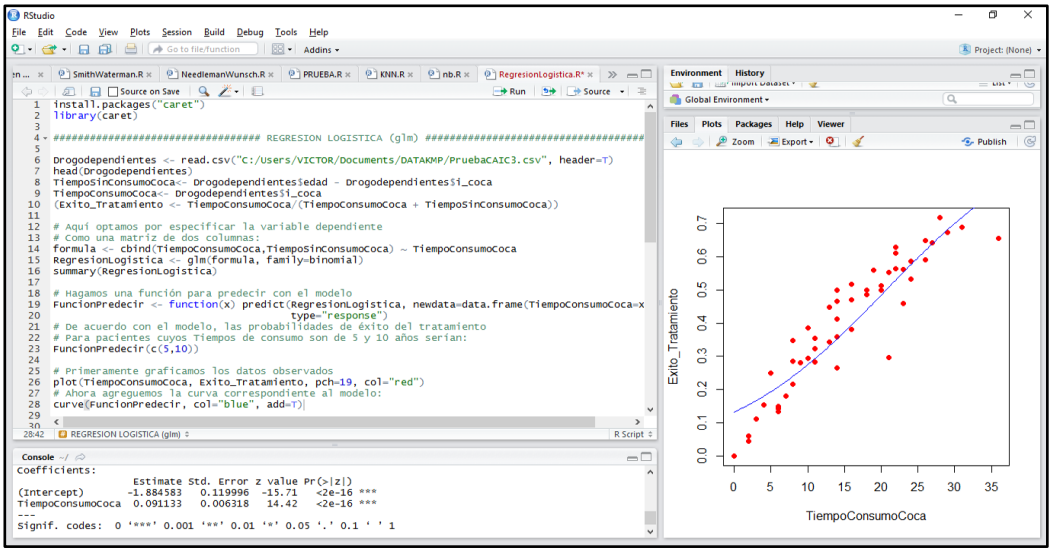


Figura 11. Implementación del caso con regresión logística.

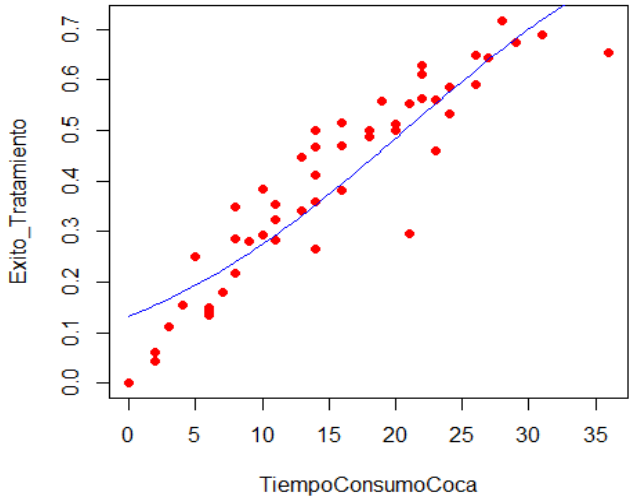


Figura 12. Gráfico de regresión logística resuelto con la función glm().

2.3.3 Redes Neuronales

El primer modelo de red neuronal [32] fue propuesto en 1943 por McCulloch y Pitts en términos de un modelo computacional de actividad nerviosa. Éste era un modelo binario donde cada neurona tenía un escalón o umbral prefijado y sirvió de base para los modelos posteriores.

Existen distintos modelos de redes neuronales siendo uno de los más utilizados el del "Perceptron". Esta red se basa en una "caja negra" donde lo importante es la predicción, y no cómo se hace. El proceso incluye una fase de entrenamiento (training) para la optimización de las predicciones.

Los elementos de la red son:

- ❖ Las neuronas o nodos
- ❖ Las capas
  - De entrada
  - De salida
  - Oculta (puede tener a su vez varias capas)
- ❖ Los pesos
- ❖ La función de combinación
- ❖ La función de activación
- ❖ El objetivo (target)

En la Figura 13 se muestra como los nodos (neuronas) de la capa de entrada, se combinan con los nodos de la capa oculta mediante la función de combinación, que suele ser una combinación lineal de los nodos de entrada mediante los pesos. A las neuronas de las capas ocultas, se les aplica una función de activación, que suele ser la tangente hiperbólica de la anterior combinación más un parámetro por nodo oculto, con lo que estimamos las neuronas de la capa de salida y sus errores.

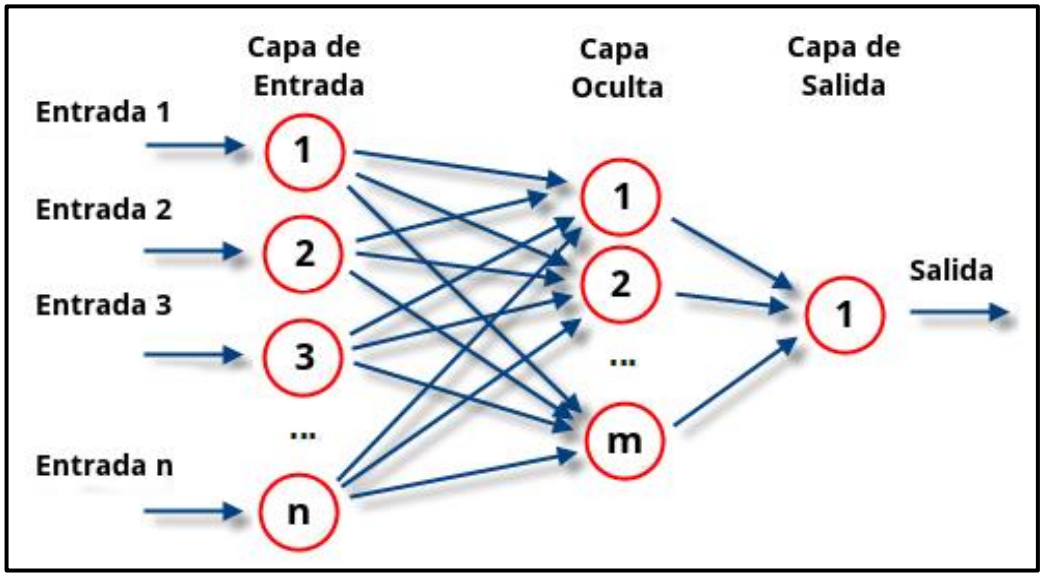


Figura 13. Diagrama de los elementos de la red neuronal.
Fuente [32]

Para el caso de redes neuronales con R, en este trabajo hemos usado el método del paquete "neuralnet", específicamente la función "Entrenamiento de redes neuronales". El paquete neuralnet [33, 34] permite la estimación de modelos MLP (Modelo Lineal de Probabilidad) con el algoritmo RPROP (Riedmiller, 1994). La función permite una configuración flexible a través de la opción de la función error y activación. Además, se implementa el cálculo de pesos generalizados (Intrator O. y Intrator N., 1993).

Para ejemplificar esta técnica, se ha implementado con R el siguiente caso: de la base de datos de pacientes drogodependientes se requiere generar una red neuronal para predecir si el tratamiento de los pacientes drogodependientes tendrá éxito (BIEN) o fracaso (MAL).

Para ello se han utilizado las variables de adicciones y la variable resultado (i_coca, i_tabaco, i_bdz, i_hero, i_coca, i_canna, i_anfet y resultado). El código se muestra a continuación:

```
# DATOS Y LIBRERIA A USAR
#---------------------------------------------------------------------
BDPacientes <- read.csv("C:/Users/VICTOR/DBAdicciones2.csv", header=T)
library(neuralnet)
n     <- sample(1:54,10)
Train <- BDPacientes[-n,]
Test  <- BDPacientes[n,]
clase <- c("MAL","BIEN")

# CLASES
#---------------------------------------------------------------------
Train$MAL     <- ifelse(Train$resultado == "Mal"     , TRUE, FALSE)
Train$BIEN <- ifelse(Train$resultado == "Bien" , TRUE, FALSE)

# FORMULA
#---------------------------------------------------------------------
frml <- as.formula(paste("MAL+BIEN ~ ", #variables a Predecir
                         paste(names(within(
                           Train,rm(resultado,
                                    MAL,BIEN))),
collapse="+") ))

# MODELO
#---------------------------------------------------------------------
modelo.net <- neuralnet(frml,
                        data      = Train,
                        algorithm = "rprop+", # ver en rprop+
                        threshold = 0.1,   # ver en threshold
                        hidden    = 3         # ver en hidden
)

# PREDICCION
#---------------------------------------------------------------------
predict_prb        <- as.data.frame(compute(modelo.net,
within(Test,rm(resultado)))$net.result)
names(predict_prb) <- clase
predict_class      <-
colnames(predict_prb)[apply(predict_prb,1,which.max)]

# MATRIZ CONFUCION Y GRAFICO DE RED
#---------------------------------------------------------------------
(MC <- table(Test$resultado,predict_class))
plot(modelo.net)
```

El parámetro **threshol** = 0.1 indica que las iteraciones se detendrán cuando el cambio del error sea menor a 10% entre una iteración de optimización y otra. Este cambio es calculado como la derivada parcial de la función de error respecto a los pesos. El parámetro **algoithm** = "rprop+" refiere al algoritmo "Resilient Backpropagation", que actualiza los pesos considerando únicamente el signo del cambio, es decir, si el cambio del error es en aumento (+) o disminución (-) entre una iteración y otra. El parámetro **hidden** = 3 especifica una capa oculta con 3 neuronas. Si se quisieran dos capas ocultas con tres neuronas cada una, seria hidden = c(3,3).

Los resultados del caso se muestran a continuación. En la Figura 14 se muestra el desarrollo de la codificación en RStudio de R y los resultados de la misma. Así mismo en la Figura 15 se visualiza el modelo de la red neuronal del caso.

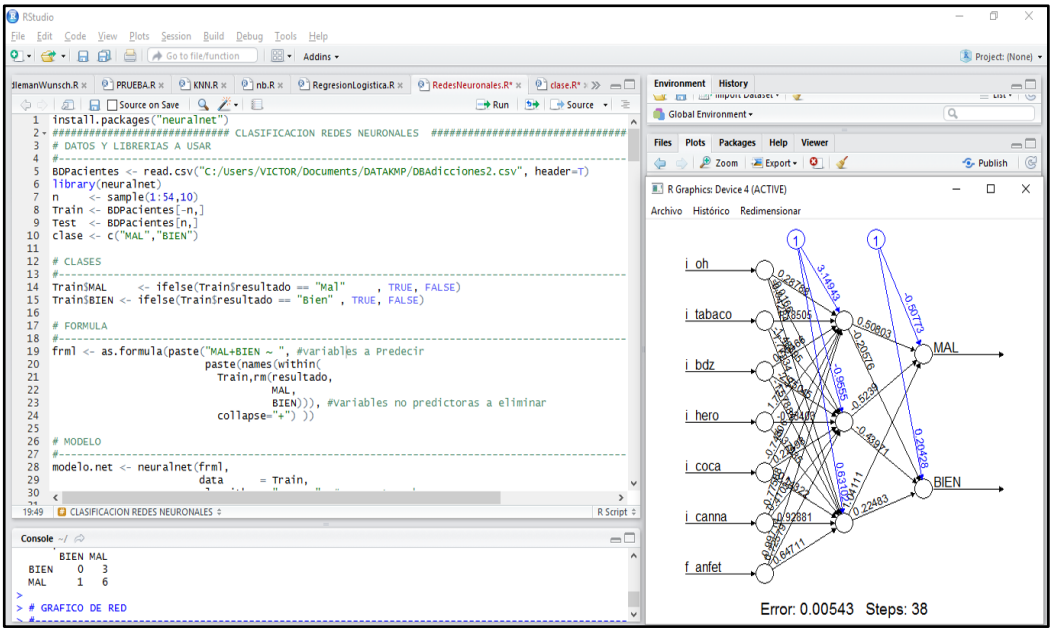


Figura 14. Implementación en RStudio de la red neuronal del caso.

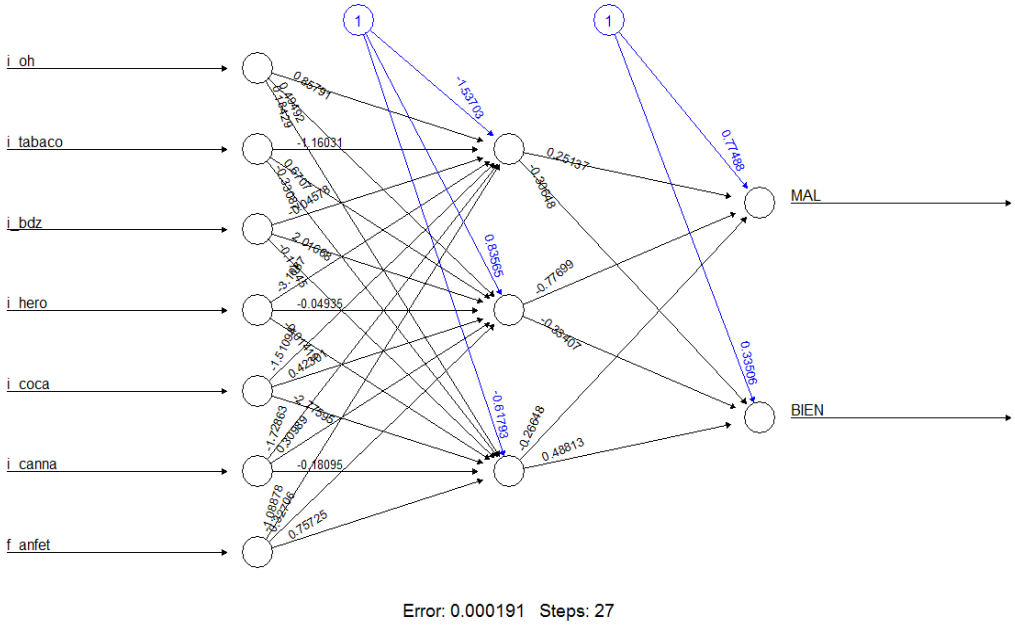


Figura 15. Modelo de la red neuronal del caso.

El análisis de las ventajas y desventajas de las técnicas más utilizadas en R se muestra en la Tabla 3.

Tabla 3. Ventajas y desventajas de las técnicas de análisis predictivo con R.

Técnica	Ventajas	Desventajas
Naïve Bayes Classifier	• Solo se requiere una pequeña cantidad de datos de entrenamiento para estimar los parámetros (las medias y las varianzas de las variables) necesarias para la clasificación. • Rápido para entrenar (solo escaneo). Rápido para clasificar. • No es sensible a las características irrelevantes. • Maneja bien los flujos de datos. • Maneja datos reales y discretos.	• Si el valor de la clase y de la función dada no ocurren juntas en los datos de entrenamiento, entonces la estimación basada en la probabilidad de frecuencia será cero. • Asume la independencia de las características.
Regresión Logística	• Funciona bien para predecir los resultados categóricos. • Puede manejar los resultados no lineales • Las variables independientes no tienen que tener una distribución normal, o tiene la misma diferencia en cada grupo.	• No puede predecir los resultados continuos. • requiere muchos más datos para alcanzar resultados estables y significativos. Datos de al menos 50 puntos por predictor es necesaria para alcanzar resultados estables. • Cuando se incluyen las variables independientes equivocadas, el modelo tendrá poco o ningún valor predictivo.
Redes Neuronales	• Pueden utilizarse para llevar a cabo la modelización estadística no lineal y ofrecer una nueva alternativa a la regresión logística. • Tiene la capacidad de detectar de forma implícita complejas relaciones no lineales entre las variables dependientes e independientes. • Así mismo tiene la capacidad de detectar todas las posibles interacciones entre las variables predictoras y la disponibilidad de múltiples algoritmos de entrenamiento.	• Tiene una mayor carga computacional y es propenso al sobreajuste. • Una desventaja principal es la incapacidad para visualizar los modelos. De hecho, las redes neuronales son criticados comúnmente por su naturaleza de “caja negra” que ofrecen poca información sobre las relaciones causales entre las variables.

En este trabajo se ha utilizado **Naïve Bayes Classifier**, como técnica de análisis predictivo en R ya que es adecuado para nuestro caso. Así mismo se ha tomado en cuenta el análisis de las ventajas y desventajas de la Tabla 3.

2.4 Fuentes de Información y Tratamiento de Datos de Carácter Personal

Los datos de carácter personal (DCP) cobran un interés especial y requieren tratamientos de anonimización y custodia adecuados. Es por ello que en este trabajo se han estudiado y aplicado meticulosamente. Concretamente los datos médicos tienen un nivel de protección máximo y su tratamiento debe someterse a la legislación vigente en cada país. Cabe destacar que las protecciones de los datos en el mundo cuentan con normas específicas y con autoridades encargadas de garantizar su aplicación. Europa es un continente donde la protección de datos ha alcanzado un nivel muy elevado ya que la práctica totalidad de países europeos posee estos elementos. América del Norte es también una región en que la protección de datos y la privacidad, ha alcanzado un alto nivel de desarrollo. En los últimos años también se han producido avances significativos en materia de legislación e institucionalización en países de Iberoamérica y el Pacifico, así como en algunas regiones de África.

De acuerdo con el Artículo 18, sección 1, capítulo 2 de los Derechos y Deberes Fundamentales, de la Constitución Española [35]:

1. Se garantiza el derecho al honor, a la intimidad personal y familiar y a la propia imagen.
2. El domicilio es inviolable. Ninguna entrada o registro podrá hacerse en él sin consentimiento del titular o resolución judicial, salvo en caso de flagrante delito.
3. Se garantiza el secreto de las comunicaciones y, en especial, de las postales, telegráficas y telefónicas, salvo resolución judicial.
4. La ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos.

Al entrar en vigor la Ley Orgánica de Protección de Datos de Carácter Personal 15/1999, de 13 de diciembre, se prescriben diversas obligaciones para las aquellas empresas que cuenten con ficheros con datos de carácter personal. “La presente Ley Orgánica tiene por objeto garantizar y proteger, en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor e intimidad personal y familiar” [36]

El Registro General de Protección de Datos es el Órgano de la Agencia Española de Protección de Datos responsable de cautelar la presencia de los ficheros y tratamientos de datos de carácter personal, regulados de acuerdo al artículo 14 de la

Ley Orgánica 15/1999, de 13 de diciembre de protección de datos de carácter personal. Son objetos de inscripción en el Registro General de Protección de Datos:

- Los ficheros de las Administraciones Públicas.
- Los ficheros de titularidad privada.
- Las autorizaciones de transferencias internacionales de datos de carácter personal con destino a países que no presten un nivel de protección equiparable al que presta la LOPD a que se refiere el ART.33.1 de la citada Ley.
- Los códigos tipo, a que se refiere el artículo 32 de la LOPD.
- Los datos relativos a los ficheros que sean necesarios para el ejercicio de los derechos de información, acceso, rectificación, cancelación y oposición [37].

La Ley Orgánica de Protección de Datos de Carácter Personal 15/1999, de 13 de diciembre, no contiene una definición exacta de los datos médicos, pero sí la tiene sobre los datos personales, la cual concierne a cualquier persona física identificada o identificable. La Ley aplicable a los datos médicos es la Ley Orgánica de Protección de Datos de Carácter Personal 15/1999, de 13 de diciembre, en la cual se regulan las reglas generales de los tratamientos de datos. Asimismo, la Ley 41/2002 regula las cuestiones estrictamente sanitarias. Ambas constituyen el marco normativo interno de los datos sobre salud y su tratamiento.

En el artículo 8 de la Ley Orgánica de Protección de Datos de Carácter Personal 15/1999, de 13 de diciembre, se establece específicamente para los datos relativos a la salud, lo siguiente: “Sin perjuicio de lo que se dispone en el artículo 11 respecto de la cesión, las instituciones y los centros sanitarios públicos y privados y los profesionales correspondientes podrán proceder al tratamiento de los datos de carácter personal relativos a la salud de las personas que a ellos acudan o hayan de ser tratados en los mismos, de acuerdo con lo dispuesto en la legislación estatal o autonómica sobre sanidad” [36].

La ley 41/2002 de 14 de noviembre, es una ley básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica. Esta ley tiene como objeto la regulación de los derechos y obligaciones de los pacientes, usuarios y profesionales, así como de los centros y servicios sanitarios, públicos y privados, en materia de autonomía del paciente y de la información y documentación clínica. La persona que elabore o tenga acceso a la información y la documentación clínica está obligada a guardar la reserva debida. En el artículo 16 de esta ley, se establece específicamente para el uso de la historia clínica, lo siguiente: “El acceso a la historia clínica con fines judiciales, epidemiológicos, de salud pública, de investigación o de docencia, se rige por lo dispuesto en la Ley Orgánica 15/1999, de Protección de Datos de Carácter Personal, y en la Ley 14/1986, General de Sanidad, y demás normas de aplicación en cada caso. El acceso a los datos y documentos de la historia clínica queda limitado estrictamente a los fines específicos de cada caso” [38].

En este trabajo se han tenido en cuenta las legislaciones en materia de protección de datos como complemento a nuestro análisis dado que utilizamos datos de carácter personal. Al no existir procedimientos de comunicación externas, en este trabajo basta utilizar los datos tras una convenientemente anonimización. Siguiendo las pautas de los procesos de Auditoría Informática y las recomendaciones sobre los roles en procesos de revisión, se ha decidido que el autor de este trabajo actúe como encargado del tratamiento y como responsable de seguridad del fichero don Diego Urgelés, doctor responsable del hospital Nuestra Señora de la Paz, siendo estos los únicos usuarios del sistema por el momento.

A partir de los datos originales custodiados por el responsable de seguridad, se ha generado una base de datos manipulable por el encargado del tratamiento. Para ello se han eliminado datos de carácter personal como nombre, datos de localización, número de teléfono, etc. Algunos de estos datos, como en el caso de los datos de localización (lugar de consumo, dirección del trabajo, ruta domicilio-trabajo), son importantes para el estudio, por lo que se han sustituido por codificaciones de forma que el valor del dato siga presente pero se impida la decodificación y por lo tanto se asegure la privacidad de los datos. Estas transformaciones son codificaciones no invertibles, es decir, la manera de garantizar la privacidad de los datos se fundamenta en la no posibilidad de deducir a partir de los datos transformados, la identidad del paciente del que se obtuvieron. Además se han realizado algunas fusiones para reducir la dimensionalidad del problema, permitiendo optimizar los tiempos de ejecución sin eliminar información.

Como consecuencia se ha definido una estructura de la base de datos que constituye el formulario tal y como debería aparecer en una futura notificación oficial para la Agencia Española de Protección de Datos, de acuerdo con su Guía del Responsable de Ficheros [37]. La notificación telemática inicial a la Agencia Española de Protección de Datos, se realizará cuando los responsables del hospital decidan la finalización de la fase de pruebas y el comienzo de la fase de explotación del proyecto.

Capítulo 3. Detección de Outliers y Análisis Predictivo con R

En este capítulo se exponen los procedimientos de detección de valores atípicos (outliers) y el análisis predictivo de los datos implementados en este trabajo. Para ello, en primer lugar, se definen los outliers como datos con valores muy diferentes a otros datos de la muestra y su detección. En un segundo lugar, se exponen tres métodos para la detección de outliers en R. Uno de ellos propio y otros dos elegidos de la literatura. Finalmente, se expone el algoritmo de aprendizaje supervisado Naïve Bayes en R, como una técnica de clasificación y predicción que construye modelos que predicen la probabilidad de posibles resultados.

Definición de Outliers

Los valores atípicos son datos con valores muy diferentes a otros datos de la muestra. Estos datos atípicos distorsionan los resultados del análisis, debido a esta razón hay que identificarlos y tratarlos de manera adecuada. Si los valores atípicos del conjunto de datos se ignoran, puede haber cambios importantes en las conclusiones obtenidas del estudio.

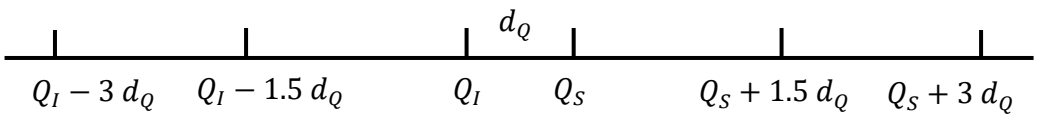
Identificación de Outliers

Para una muestra de datos $x_1, \dots, x_n$ se denota $x_{(1)} x_{(2)}, \dots, x_{(n)}$ la muestra reordenada de menor a mayor. Es decir, para la muestra de 5 elementos $x_1 = 3,\ x_2 = 0,\ x_3 = 7,\ x_4 = 10,\ x_5 = 4$ tendríamos $x_{(1)} = 0,\ x_{(2)} = 3,\ x_{(3)} = 4,\ x_{(4)} = 7,\ x_{(5)} = 10$.

Aunque hay muchas formas de determinar outliers, es muy conocida la forma basada en cuartiles y el rango intercuartílico, que veremos a continuación (análogamente podría hacerse con deciles o percentiles o cualquier tipo de cuantil).

Usamos la distancia intercuantílica (d_Q) como medida de dispersión, y definimos los puntos de corte o cotas para detectar outliers de la siguiente manera:

$$\text{Cota Interna Inferior} = Q_I - 1.5\ d_Q$$
$$\text{Cota Interna Superior} = Q_S + 1.5\ d_Q$$
$$\text{Cota Externa Inferior} = Q_I - 3\ d_Q$$
$$\text{Cota Externa Superior} = Q_S + 3\ d_Q$$



A partir de estas cotas se define el valor adyacente inferior (VAI) como el valor más cercano, mayor o igual, a la cota interna inferior; y el valor adyacente superior (VAS) como el valor más cercano, mayor o igual, a la cota interna inferior.

Se observa que si: $VAI = X_{(1)} = Mínimo$ **y** $VAS = X_{(n)} = Máximo$, no hay valores atípicos.

En este trabajo se ha utilizado tres métodos para la detección de outliers. Dos de estos métodos utilizan la librería R (*Outliers* y *Rcmdr*) y ofrecen diferentes resultados. Por ello, se ha decidido incorporar un método propio para aportar variedad a los resultados. Con el propósito de eliminar solo los outliers detectados por los tres métodos.

3.1 Detección de Outliers en R

3.1.1 Método 1: Detección de Outliers con Cadenas No Monótonas (DOCNM)

Se dice que una cadena es monótona si preserva el orden, concretamente una cadena $x_1, x_2, \dots, x_n$ es monótona creciente si se cumple que:

$$\forall_i \, \forall_j, i < j \;\; => x_i \leq x_j$$

Análogamente, una cadena $x_1, x_2, \dots, x_n$ es monótona decreciente si se cumple que:

$$\forall_i \forall_j, i < j \;\; => x_i \geq x_j$$

Por lo tanto, una cadena es monótona si solo si es monótona creciente o monótona decreciente.

En este trabajo se ha implementado un algoritmo propio en R. La cual se ha denominado algoritmo "DOCNM" (Detección de Outliers con Cadenas No Monótonas). El código de la implementación en R se encuentra en el Anexo 1.

La idea se basa en los procesos de regresión lineal de los datos. Es conocido que cuando un dato es un outlier sus características lo posicionan gráficamente fuera del rango o rangos respecto a una o varias características. Al analizar la muestra correspondiente se presenta típicamente un crecimiento-decrecimiento (o viceversa) entorno al dato. Como se muestra en la Figura 16.

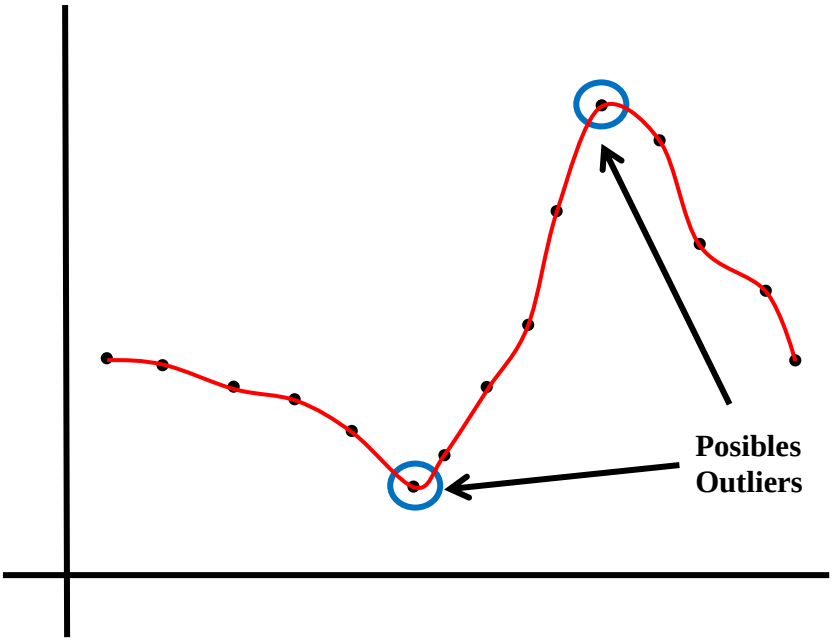


Figura 16. Detección de posibles outliers por CNM.

Los procesos previos de los datos utilizados como input se muestran en el Capítulo 4. Así mismo para la aplicación del algoritmo DOCNM se ha etiquetado los datos ya que se basa en los procedimientos del algoritmo KMP. El proceso general de la aplicación del algoritmo DCONM se muestra en la Figura 17.

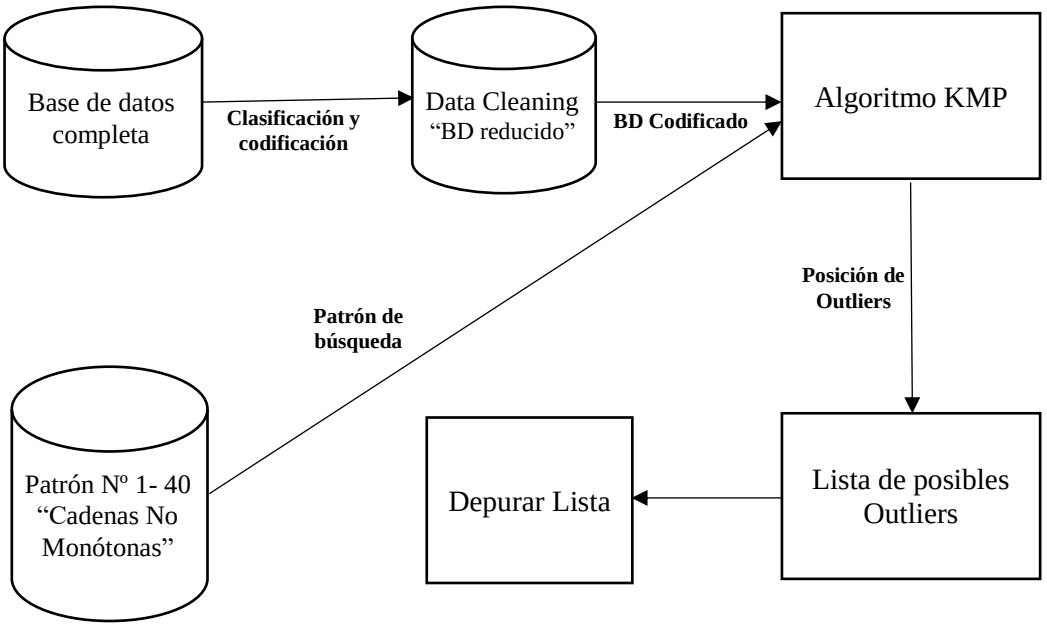


Figura 17. Diseño del proceso general del algoritmo DOCNM.

3.1.1.1 Algoritmo Knuth-Morris-Pratt (KMP)

La búsqueda de patrones en un texto es un problema muy importante en la práctica. El algoritmo Knuth-Morris-Pratt (KMP) [39, 40] es un algoritmo en tiempo real para la búsqueda de patrones en una cadena especificada, en otras palabras, dada una cadena S, y un patrón P, el problema consiste en devolver todas las posiciones de S donde se detecta el comienzo del patrón P.

Existen otros algoritmos que realizan búsquedas similares sin embargo KMP es especialmente interesante y consigue gran rapidez gracias a las siguientes características:

- Realiza las comparaciones de izquierda a derecha.
- Procesamiento previo de complejidad O(m) en espacio y tiempo, correspondiente a un análisis sobre el patrón.
- Búsqueda de complejidad temporal O(n+m) independiente del tamaño del alfabeto.

El algoritmo KMP ha sido aplicado ampliamente en la práctica por ejemplo en [41]. Este algoritmo cuando está comparando el patrón en el texto, si encuentra una disconformidad, retrocede hasta una determinada posición reiniciándose la búsqueda en una posición más avanzada.

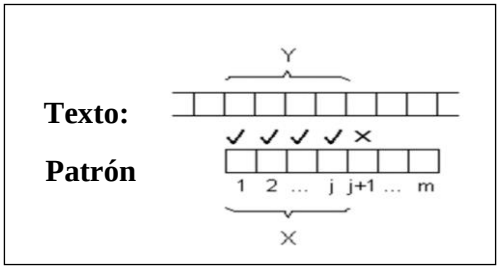


Figura 18. Comparación del patrón y el texto en una posición dada.

En la Figura 18, X es la parte del patrón que se alinea con el texto, e Y es la correspondiente parte del texto. La longitud de X es j. Un algoritmo de fuerza bruta movería el patrón una posición hacia la derecha, sin embargo, esto puede no ser lo correcto en el sentido de que los primeros j-1 caracteres de X pueden o no alinear los últimos j-1 caracteres de Y.

La observación clave que realiza el algoritmo KMP es que X es igual a Y, por lo que la pregunta planteada en el párrafo anterior puede ser respondida mirando solamente el patrón de búsqueda, lo cual permite pre-calcular la respuesta y almacenarla en una tabla. Por lo tanto, si al deslizar el patrón en una posición no funciona, se puede intentar deslizarlo en 2, 3, ..., hasta j posiciones (Figura 19).

Para modelizar matemáticamente el problema, KMP define la función de fracaso (failure function) como:

$$f(j) = max\left(i < j \mid b_i, .. b_i = \ b_{j-i+1} .. b_j\right)$$

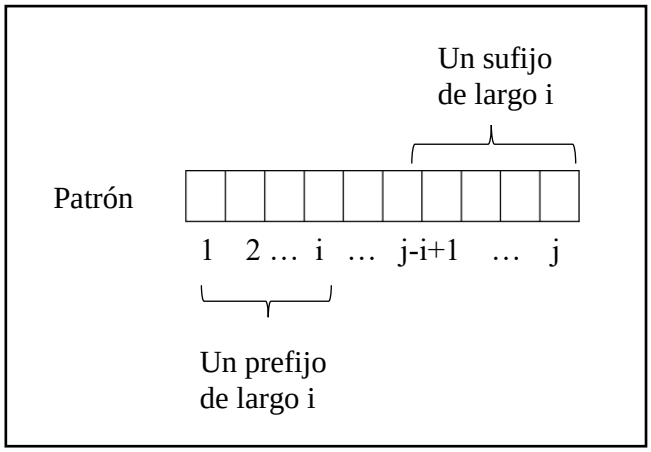


Figura 19. Deslizamiento del patrón en posiciones.

En el Anexo 2 se expone con más detalle los componentes del algoritmo KMP y se muestra un ejemplo de ejecución.

En la Figura 20 se muestra el pseudocódigo del algoritmo KMP.

- **Compute-Prefix-Function (Π)**

```
1 m ← length[p]   //Patrón 'p'.
2 Π [1] ← 0
3 k ← 0
4         for q ← 2 to m
5         do while k > 0 and p[k+1]! = p[q]
6               do k ← Π[k]
7         If p[k+1] = p[q]
8         then k ← k +1
9         Π[q] ← k
10        return Π
```

- **KMP Matcher**

```
1 n ← length[S]
2 m ← length[p]
3 Π ← Compute-Prefix-Function(p)
4 q ← 0
5 for i ← 1 to n
6    do while q > 0 and p[q+1]! = S[i]
7         do q ← Π[q]
8         if p[q+1] = S[i]
9         then q ← q + 1
10        if q = m
11        then print "Patrón ocurre en la posición" i – m
12        q ← Π [ q]
```

Figura 20. Pseudocódigo del algoritmo KMP.
Fuente [40].

3.1.2 Método 2: Detección de Outliers con la Librería "*Outliers*" de R.

La librería Outliers [42] de R, permite la detección de valores atípicos desde un conjunto de datos. En este trabajo hemos utilizado la versión 0.14 de febrero del 2015 titulada "test for outliers" escrita y mantenida por Lukaz Komsta. Se trata de una colección de test usado tradicionalmente para la detección de outliers.

Para este trabajo he probado los test chisq.out.test y cochran.test, eligiendo el primero finalmente para las comparativas. El test elegido se basa en la distribución de Chi-cuadrado para encontrar diferencias de los datos y la media de la muestra. Los datos de las variables de drogodependencia se muestran en el capítulo 4.

En este trabajo se han realizado algunos procesos para búsqueda de outliers en la base de datos de pacientes drogodependientes, siguiendo el esquema de la Figura 21. Los resultados se muestran en un capítulo posterior.

Como en el caso anterior ha sido necesario realizar un pre-procesado de la base de datos. También en este caso el resultado es una lista de pacientes con perfil outlier.

Los resultados obtenidos durante la aplicación de la librería outliers no coincide totalmente con los resultados obtenidos con el algoritmo DOCNM siendo muy interesante la comparación de estos resultados. Así mismo estos resultados pueden compararse y combinarse con otros métodos, concretamente con la detección de outliers con regresión lineal simple, que veremos en el apartado siguiente.

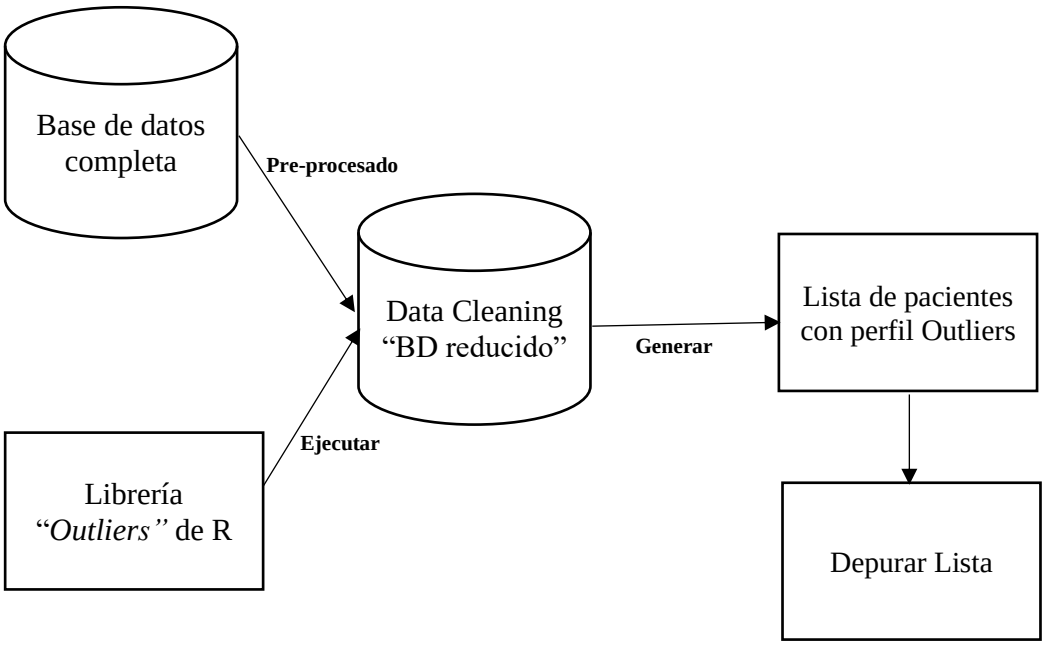


Figura 21.- Diseño general de la detección de outliers con la librería "*Outliers*" de R.

3.1.3 Método 3: Detección de Outliers con Regresión Lineal Simple.

En este apartado se muestran los procesos previos para la detección de valores atípicos con regresión lineal simple. La relación de las variables no siempre será cierta, para ello necesitaremos algún instrumento que nos permita decidir la existencia de una relación entre variables, como por ejemplo la variable edad con la de años de consumo de coca, la edad con años de consumo de tabaco, etc. La forma más sencilla de comenzar consiste en realizar representaciones gráficas.

El principio de parsimonia nos dice que un modelo de regresión lineal es una manera eficaz de explicar la relación entre las variables. Cabe resaltar que el test de valores atípicos de Bonferroni nos permite detectar la presencia u observaciones atípicas. Para ello vamos a establecer el modelo que relaciona las variables. Por ejemplo, analizaremos la regresión entre i_coca con edad, pero antes de esto estudiaremos la normalidad de los datos y calcularemos la correlación entre edad y i_coca, realizando además el correspondiente gráfico de dispersión como paso previo al modelo de regresión lineal propiamente dicho.

El proceso completo requiere que se apliquen los siguientes pasos:

1. Analizar la normalidad de los datos, lo que garantiza la calidad de los resultados.
2. Cálculo de la correlación de las variables mediante el test de Pearson.
3. Dibujar el gráfico de dispersión para visualizar las relaciones (Figura 22).
4. Ajustar el modelo de regresión lineal mediante la función lm de R (Figura 23).
5. Creación del modelo mediante ModelDrogo.
6. Cálculo de estimadores, errores y p-valores.
7. Obtención de los intervalos de confianza asociados.
8. Obtención de los valores atípicos.

En el Anexo 3 se muestran los detalles de los desarrollos aplicados para este trabajo sobre la base de datos de pacientes drogodependientes.

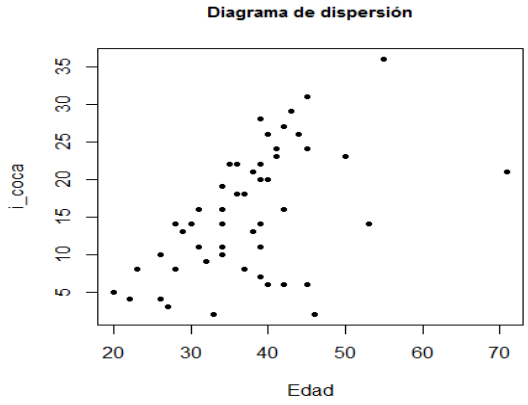


Figura 22. Diagrama de dispersión

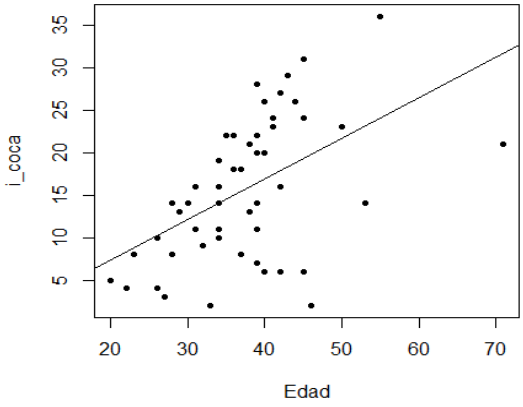


Figura 23. Diagrama de la desviación de la recta regresión

Para llevar a cabo estas ideas, se ha utilizado la librería *Rcmdr* de R.

La librería Rcmdr de R, proporciona una interfaz gráfica de usuario (GUI), como se muestra en la Figura 24, que nos permite de una manera sencilla interpretar los coeficientes de un modelo lineal.

En este trabajo hemos utilizado la versión 2.2-4 de abril del 2016 titulada "R Commander" escrita por varios autores [43] y mantenida por John Fox. Se trata de una plataforma independiente de estadística básica GUI para R, basada en el paquete de tcltk. Para ello se ha utilizado el test de valores atípicos de Bonferroni. Este test se basa en detectar la presencia de valores atípicos y observar el gráfico "Residuals vs Leverage", que detecta valores no influyentes en la estimación del modelo. El test de Bonferroni de Rcommander se basa en el método de corrección de Bonferroni para comparaciones múltiples [44, 45]. Se encuentra dentro de la opción del menú Modelos / Diagnósticos numéricos, de la plataforma de RCommander.

Así mismo para este trabajo se ha realizado algunos procesos de búsqueda de outliers con *Rcmdr*, siguiendo el esquema de la Figura 25. Los resultados se muestran en un capítulo posterior. Por otro lado, en el Anexo 4 se muestra un ejemplo de cómo se utiliza la herramienta Rcommander de R para la detección de outliers. Como en casos anteriores ha sido necesario realizar un pre-procesado a la base de datos.

Cabe resaltar que los resultados obtenidos durante la aplicación de la librería *Rcmdr* no coinciden totalmente con los resultados obtenidos con el algoritmo DOCNM, siendo muy interesante la comparación de estos resultados por parte del experto médico.

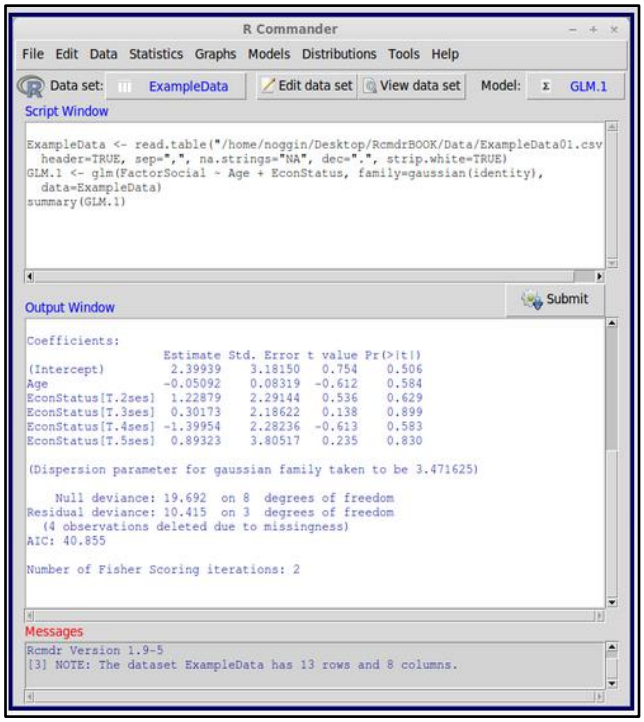


Figura 24. Interfaz gráfico de usuario con R Commander.

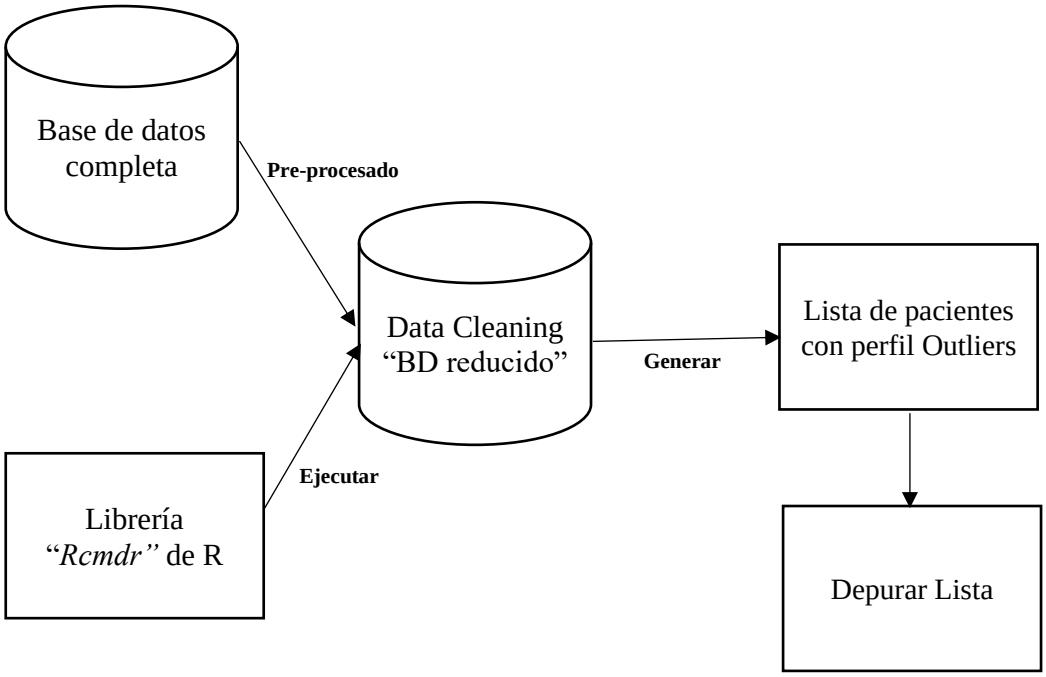


Figura 25. Diseño general de la detección de outliers con la librería "*Rcmdr*" de R.

3.2. Análisis Predictivo con R

En este apartado se muestra el algoritmo de clasificación Naïve Bayes [46, 47], lo que buscamos es predecir si el resultado del tratamiento en nuevos pacientes drogodependientes tendrá éxito o no.

El análisis se ha desarrollado con R, para ser más específico con la librería "e1071", que calcula la probabilidad condicional a posteriori de una variable de clase categórica dadas las variables predictoras independientes utilizando la regla de Bayes.

En el Capítulo 2 ya se hace un estudio de los métodos de predicción, eligiendo Bayes para ello. Así mismo la aplicación de los procesos que se describen en este capítulo se muestran en el capítulo 4.

3.2.1 Algoritmo Naïve Bayes

El problema consiste en que a partir de una tabla de aprendizaje (con ciertas características de drogodependencia) y usando el algoritmo Naïve Bayes predecir si un nuevo paciente drogodependiente tendrá éxito o no en su tratamiento.

Su nombre original es "Naïve Bayesian Clasiffier". Es un algoritmo de aprendizaje supervisado con una técnica de clasificación y predicción que construye modelos que predicen la probabilidad de posibles resultados, como se muestra en la Figura 26. Este algoritmo utiliza datos históricos para entrenar y encontrar asociaciones y relaciones entre las variables a partir de las cuales realizar las predicciones. El algoritmo se basa en la probabilidad condicional y en el Teorema de Bayes. Así mismo este algoritmo de clasificación está diseñado para atributos categóricos por lo que en el caso de atributos continuos es necesario categorizar o discretizar a partir de la función de densidad de cada variable.

En el Anexo 5 se explican las generalidades y los conceptos de probabilidad [48] que se necesitan para la comprensión del clasificador bayesiano.

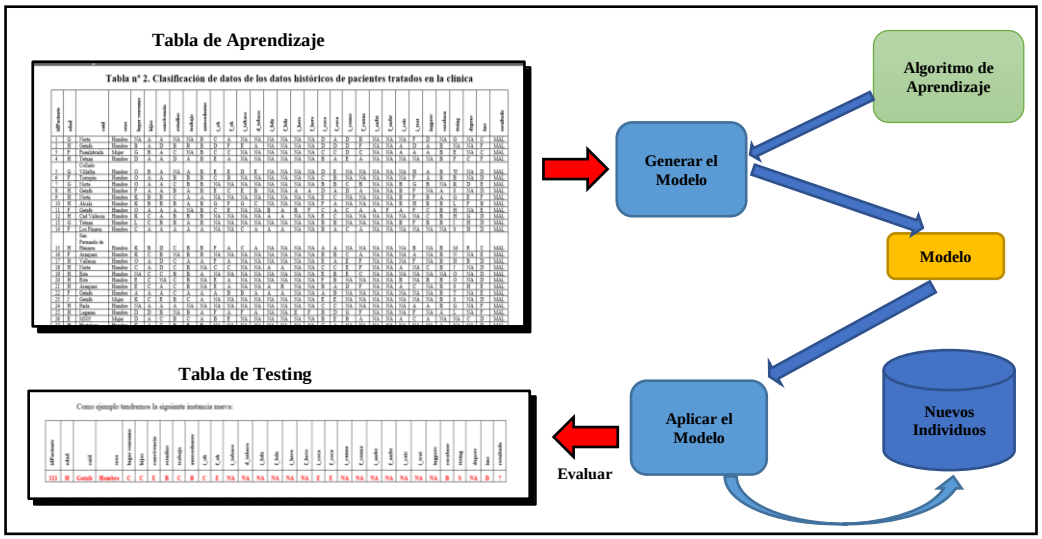


Figura 26.- Diseño del modelo de clasificación.

Capítulo 4: Resultados

En este capítulo se describe la base de datos utilizada en este trabajo. Seguidamente se describen los pasos realizados para: (1) realizar la limpieza de datos tal como se describe en el Capítulo 2, usando la herramienta de OpenRefine, (2) detección de outliers mediante los tres métodos descritos en el Capítulo 3, y (3) construcción del modelo de predicción según Naïve Bayes, descrito también en el Capítulo 3.

4.1 Base de Datos

En este trabajo utilizamos los datos de los pacientes drogodependientes del Centro de Atención Integral al Cocainómano (CAIC), el cual es un dispositivo de evaluación y tratamiento intensivo para los problemas de abuso y/o dependencia a la cocaína, asociados o no, al consumo de otras sustancias.

El CAIC se encuentra ubicado en la Clínica Nuestra Señora de la Paz (ver Figura 27). Esta clínica es gestionada por la Orden Hospitalaria San Juan de Dios, entidad sin ánimo de lucro. Este centro está financiado por la Consejería de Sanidad de la Comunidad de Madrid. La Unidad de Hospitalización del CAIC se puso en marcha en 2001, mientras que el Centro de Día de Cocaína comenzó a funcionar en 2007.

Figura 27.- Clínica Nuestra Señora de la Paz – Madrid.

En el año 2015, hubo un total de 137 ingresos en la Unidad de Hospitalización y 91 en el Centro de Día de Cocaína. El grupo multiprofesional se encuentra formado por expertos en Psiquiatría, Psicología Clínica, Medicina Interna, Enfermería, Terapeutas Ocupacionales, Educadores Sociales, Auxiliares de Enfermería, Agentes de Pastoral y Trabajadores Sociales que trabajan con el fin de lograr la desintoxicación, deshabituación y prevención de recaídas en el paciente [2].

A continuación, se describe los datos de pacientes drogodependientes procedentes del Hospital Nuestra Señora de La Paz de Madrid (Tabla 4). En el Anexo 6 se exponen con mayor detalle los datos (archivo histórico).

Tabla 4. Base de datos de drogodependientes

Campo	Tipo	Descriptor
nacimiento	Fecha	Fecha de nacimiento
edad	Numérico	Edad del paciente
caid	Carácter	Ciudad de procedencia
sexo	Carácter	Identificador de la sexualidad
lugarconsumo	Carácter	lugar de consumo del paciente
hijos	Numérico	Cantidad de hijos del paciente
convivencia	Carácter	Situación de la convivencia
estudios	Carácter	Grado de instrucción
trabajo	Carácter	Situación laboral
antecedentes	Carácter	Antecedentes penales
i_oh	Numérico	Edad de inicio del consumo de alcohol
f_oh	Numérico	Frecuencia de consumo de alcohol
i_tabaco	Numérico	Edad de inicio del consumo de tabaco
d_tabaco	Numérico	Frecuencia de consumo de tabaco
i_bdz	Numérico	Edad de inicio del consumo de benzodiacepina
f_bdz	Numérico	Frecuencia de consumo de benzodiacepina
i_hero	Numérico	Edad de inicio del consumo de heroína
f_hero	Numérico	Frecuencia de consumo de heroína
i_coca	Numérico	Edad de inicio del consumo de cocaína
f_coca	Numérico	Frecuencia de consumo de cocaína
i_canna	Numérico	Edad de inicio del consumo de cannabis
f_canna	Numérico	Frecuencia de consumo de cannabis
i_anfet	Numérico	Edad de inicio del consumo de anfetamina
f_anfet	Numérico	Frecuencia de consumo de anfetamina
i_caic	Numérico	Ingreso previo a un centro de salud
i_trat	Numérico	Edad que inicio el tratamiento
ingprev	Numérico	Ingresos previos a la clínica
cocabase	Carácter	Consumo de cocaína base
ttoing	Carácter	Tratamiento de ingreso
dxprev	Carácter	Diagnóstico previo
peso	Numérico	Peso del paciente drogodependiente
altura	Numérico	Altura del paciente drogodependiente
resultado	Carácter	Resultado del tratamiento

Así mismo se muestra la frecuencia de consumo del paciente drogodependiente, esta frecuencia es una variable cualitativa no ordinal y se encuentra en los rangos de 1-9, como se puede apreciar en la Tabla 5.

Tabla 5. Frecuencia de consumo del paciente drogodependiente

Id Frecuencia	Descriptor
1	Todos los días
2	(4 a 5) días a la semana
3	(2 a 3) días a la semana
4	(1) día a la semana
5	Menos de un f a la semana
6	No consumió
9	Desconocido

4.2 Data Cleaning

Cuando trabajamos con datos buena parte de nuestros esfuerzos y tiempo se va en la preparación, limpieza y puesta en orden de los mismos. Los problemas de valores ausentes, datos con ruido, valores atípicos, entre otros, se convierten en uno de los primeros obstáculos a superar en el camino hacia la generación de datos fiables y de calidad.

En este trabajo utilizamos la herramienta OpenRefine para el pre-procesado de datos de los pacientes drogodependientes. Esta herramienta es de código abierto y ofrece múltiples funcionalidades que van desde limpiar bases de datos, exportarlas en diferentes formatos, y transformarlas para un mejor uso.

A continuación, se muestra como se ha realizado la limpieza de datos en este trabajo. En primer lugar, abrimos la herramienta OpenRefine y creamos un proyecto nuevo (ver Figura 28).

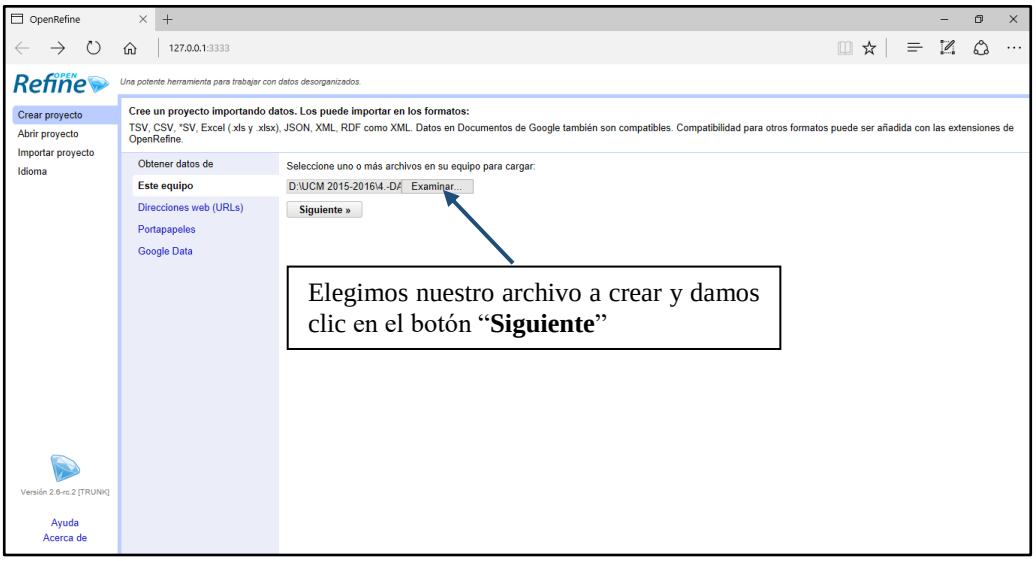


Figura 28.- Interfaz de Inicio de OpenRefine.

Seguidamente realizamos la normalización, integración y limpieza de datos. Para ello en la Figura 29 se muestran los datos antes de su transformación. En la Figura 30 se muestra como se ha normalizado el campo "caid", sin tildes y sin puntuación, pero sí número. Por ejemplo:

1. GETAFE ⟶ Getafe
2. alcorcón ⟶ Alcorcon
3. m-105 ⟶ M105

Así mismo se ha realizado la fusión de dos variables peso y altura (ver Figura 31). El resultado de esta fusión es la variable índice de masa corporal (imc) y en la Figura 32 se muestra la fusión de las mismas.

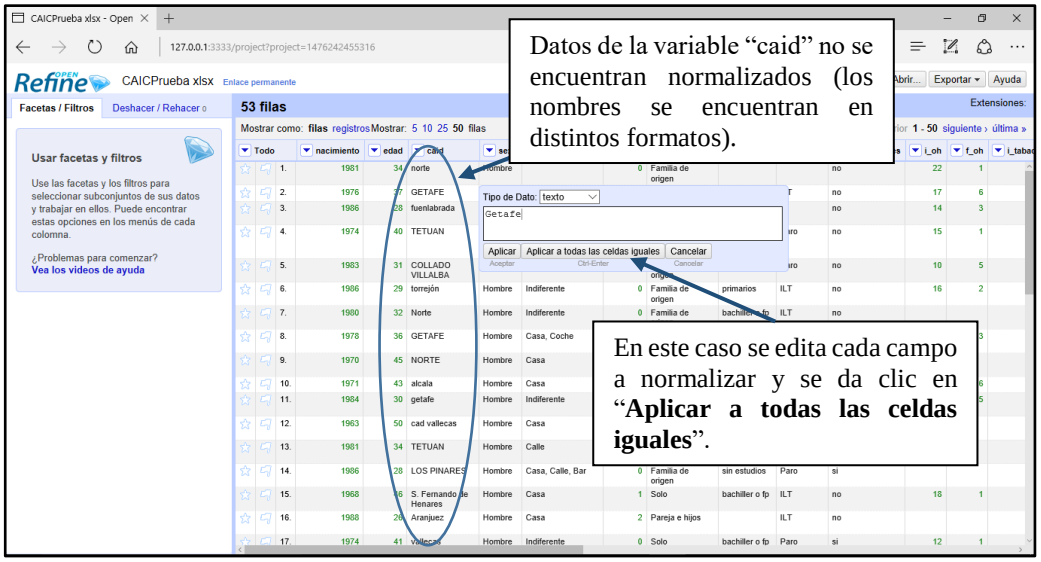


Figura 29.- Datos sin normalizar en OpenRefine.

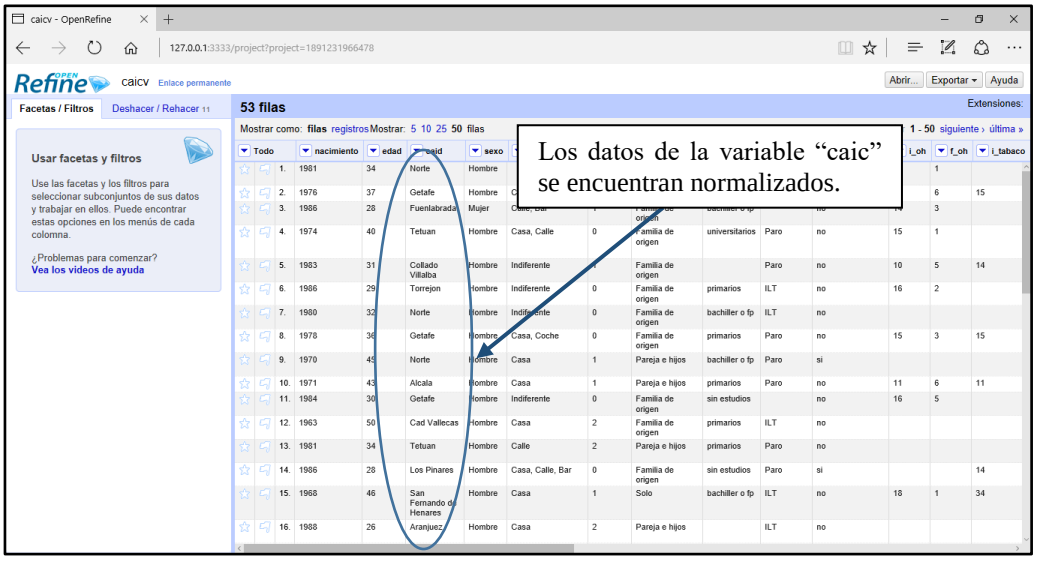


Figura 30.- Datos normalizados en OpenRefine.

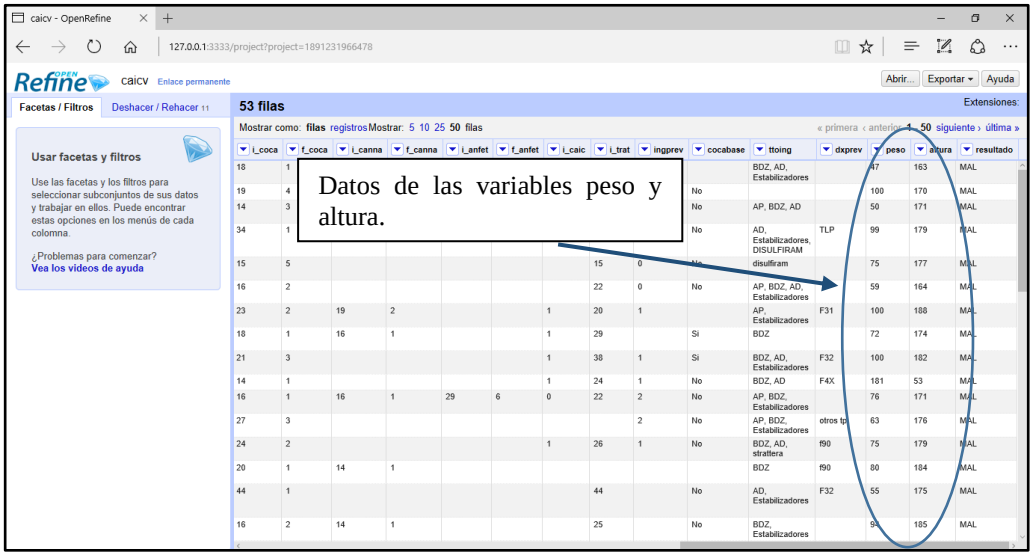


Figura 31.- Datos de las variables peso y altura en OpenRefine.

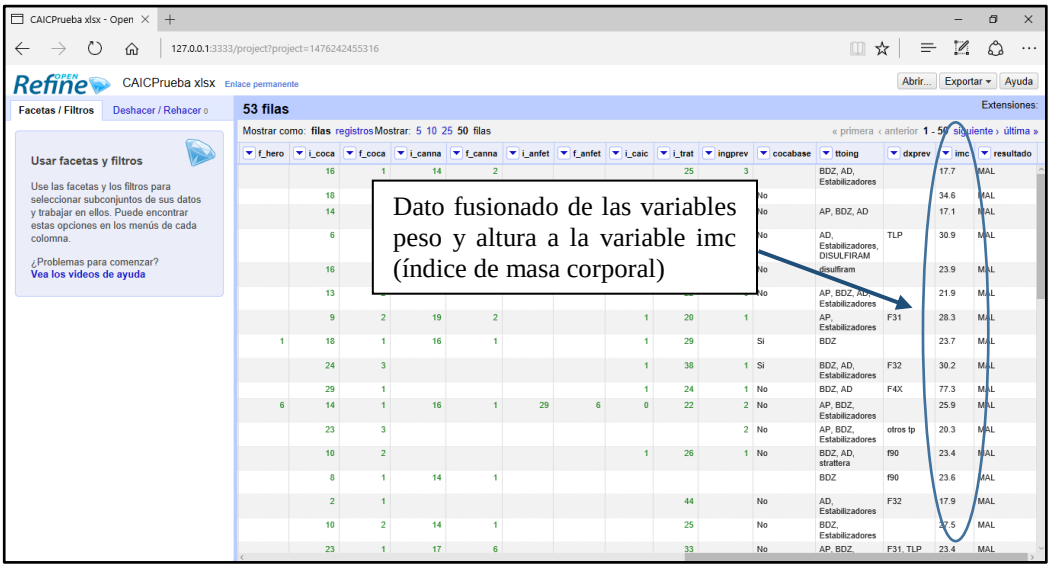


Figura 32.- Datos de la variable imc en OpenRefine.

Cabe resaltar que adicionalmente a estos procesos de limpieza de datos se han realizado pre-procesado de datos de acuerdo a la necesidad del caso. Por ejemplo, para los datos de input de los dos métodos de detección de outliers (librería *Outliers* y *Rcmdr* de R), se han utilizado solamente las variables drogodependientes y la variable edad. Estos datos se muestran en la Figura 33.

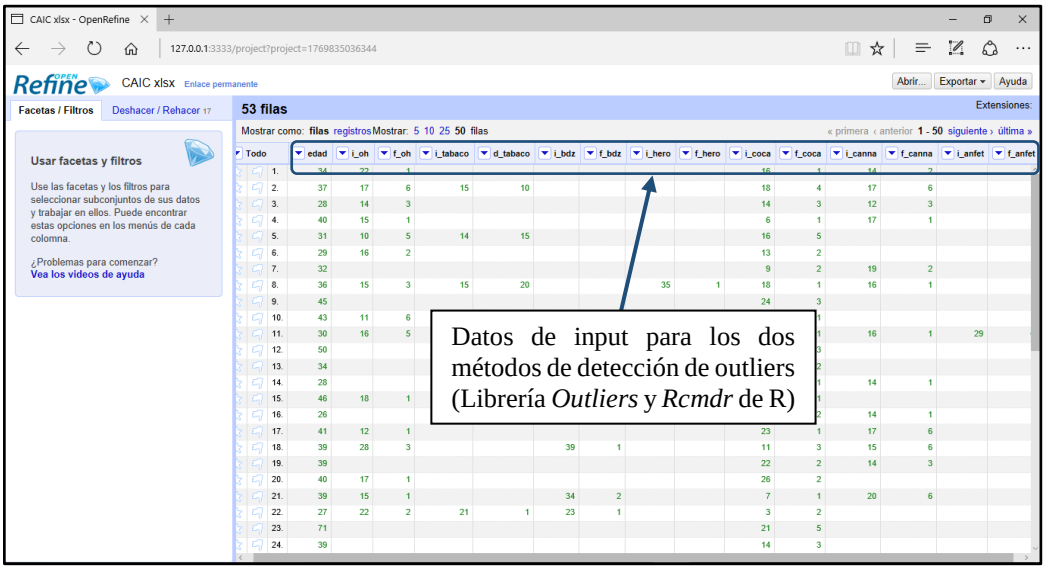


Figura 33.- Datos de input para los dos métodos de detección de outliers.

4.3 Detección de Outliers

4.3.1 Método 1: Detección de Outliers con Cadenas No Monótonas (DOCNM)

Para el proceso de detección de outliers con cadenas no monótonas es necesario realizar previamente una limpieza de los datos originarios, luego implementar los procesos de clasificación y codificación de los datos, como se muestra en la Tabla 6 y 7. Estos procesos sirven para preparar los datos como input de procedimientos de detección de datos atípicos.

Los caracteres de clasificación son los siguientes:

N=nada, **P**=poco, **R**=regular, **F**= Frecuente, **M**=muy frecuente y **O**=sin datos.

Para la codificación de los datos se ha utilizado la siguiente fórmula:

$$Valor\ Numérico = \frac{Máximo - Mínimo}{Nº\ de\ Términos}$$

Tabla 6. Clasificación de variables drogodependientes

Clase → Adicción↓	**O**	**N**	**P**	**R**	**F**	**M**
i_oh	vacío	4 - 11.4	11.4 – 18.8	18.8 - 26.2	26.2 – 33.6	33.6 - 41
f_oh	vacío	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6
i_tabaco	vacío	6- 13	13 - 20	20 - 27	27 - 34	34 - 41
d_tabaco	vacío	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6
i_bdz	vacío	0- 1.8	1.8 – 3.6	3.6 - 5.4	5.4 – 7.2	7.2 - 9
f_bdz	vacío	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6
i_hero	vacío	0 – 7.2	7.2 – 14.4	14.4 - 21.6	21.6 - 28.8	28.8 - 36
f_hero	vacío	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6
i_coca	vacío	2 - 8.8	8.8 - 15.6	15.6 - 22.4	22.4 - 29.2	29.2 - 36
f_coca	vacío	1 – 1.8	1.8 – 2.6	2.6 – 3.4	3.4 – 4.2	4.2 - 5
i_canna	vacío	6 – 12.6	12.6 – 19.2	19.2 – 25.8	25.8 – 32.4	32.4 - 39
f_canna	vacío	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6
i_anfet	vacío	1– 5.4	5.4 – 9.8	9.8 – 14.2	14.2 – 18.6	18.6 - 23
f_anfet	vacío	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6

Tabla 7.- Codificación de datos de los pacientes drogodependientes

Edad → Adicción↓	**20**	**22**	**23**	**26**	**26**	**27**	**28**	**28**	**29**	**30**	**31**	**31**	**31**	**32**	**33**	**34**	**34**	**34**	**34**	**34**	**34**	**35**	**36**	**36**	**37**	**37**	**38**	**38**	**38**	**39**	**39**	**39**	**39**	**39**	**39**	**40**	**40**	**40**	**41**	**41**	**42**	**42**	**42**	**43**	**44**	**45**	**45**	**45**	**46**	**50**	**53**	**55**	**71**
i_oh	N	N	N	O	O	N	P	O	P	P	R	P	O	O	R	P	O	O	R	P	O	F	R	R	R	P	R	R	O	N	O	R	O	O	R	R	R	R	F	F	O	O	F	F	R	O	F	O	F	O	O	M	O
f_oh	N	N	F	O	O	N	P	O	N	F	F	N	F	O	F	N	O	O	P	N	O	N	P	N	M	F	F	N	O	P	O	N	O	O	F	N	N	N	N	P	O	O	N	M	N	O	N	O	N	O	O	N	O
i_tabaco	N	N	O	O	O	N	O	P	O	O	P	P	O	O	O	O	O	P	O	O	O	O	R	O	R	P	O	O	O	O	O	O	O	O	O	O	O	R	O	O	O	R	F	F	F	O	F	O	N	O	O	M	O
d_tabaco	N	F	O	O	O	N	O	N	O	O	R	F	O	O	O	O	O	F	O	O	O	O	F	O	P	M	O	O	O	O	O	O	O	O	O	O	O	N	O	O	O	F	N	M	N	O	N	O	N	O	O	N	O
i_bdz	F	O	O	O	O	R	O	P	O	M	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	N	O	R	O	O	O	O	O	O	O	O	O	F	O	O	O	O	O	O	O	N	O	O	O	O
f_bdz	N	O	O	O	O	N	O	N	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	N	O	N	O	N	O	O	O	O	O	O	O	O	M	O	O	O	O	O	O	O	N	O	O	O	O
i_hero	O	O	O	O	N	O	O	O	O	P	O	O	O	O	O	O	O	O	O	O	O	O	N	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	F	O	F	O	O	O	F	M	O	O
f_hero	O	O	O	O	M	O	O	O	O	M	O	O	O	O	O	O	O	O	O	O	O	O	N	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	M	O	M	O	O	O	M	M	O	O
i_coca	N	N	N	P	N	N	P	N	P	P	R	P	R	P	N	R	P	R	P	P	P	R	R	R	R	N	R	R	P	P	R	N	P	F	R	N	F	R	F	F	R	N	F	F	F	F	N	M	N	F	P	M	R
f_coca	N	N	M	P	N	P	R	N	P	N	M	R	F	P	N	N	P	N	M	R	R	N	N	R	F	M	N	M	N	R	P	N	R	R	M	N	P	N	N	N	M	P	N	N	N	R	F	N	N	R	N	N	M
i_canna	N	N	N	N	N	O	P	P	O	P	O	O	O	P	P	R	O	R	R	P	O	F	R	R	R	N	R	O	O	R	R	P	O	O	R	R	O	O	R	O	O	O	F	O	F	O	F	O	O	O	O	M	O
f_canna	N	O	N	N	P	O	P	N	O	N	O	O	O	N	N	N	O	N	P	N	O	N	N	N	M	N	N	O	O	M	P	M	O	O	F	N	O	O	M	O	O	O	P	O	M	O	M	O	O	O	O	N	O
i_anfet	O	O	O	O	O	O	O	O	O	N	O	O	F	O	O	O	O	O	R	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	M	O	O	O	O	O	O	O	O	O	O
f_anfet	O	O	O	O	O	O	O	O	O	M	O	O	M	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	M	O	O	O	O	O	O	O	O	O	O
Posición	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**

Tras los procesos previos, hemos implementado un algoritmo (DOCNM) que integra la generación de cadenas no monótonas como patrones con instancias al algoritmo KMP. Para la creación de patrones hemos escogido (sin pérdida de generalidad) cadenas de longitud 3, al ser 5 códigos y palabras de longitud 3, obteniendo un total de:

$$\boldsymbol{VR}_{5,3} = 5^3 = 125\ \boldsymbol{patrones}$$

De estos patrones, sólo nos interesan las cadenas que no presentan monotonía (creciente o decreciente), por ser indicador de un posible valor atípico (correspondiente al centro de dicho patrón). Las posibles cadenas detectoras se obtienen como se muestra en la Figura 34.

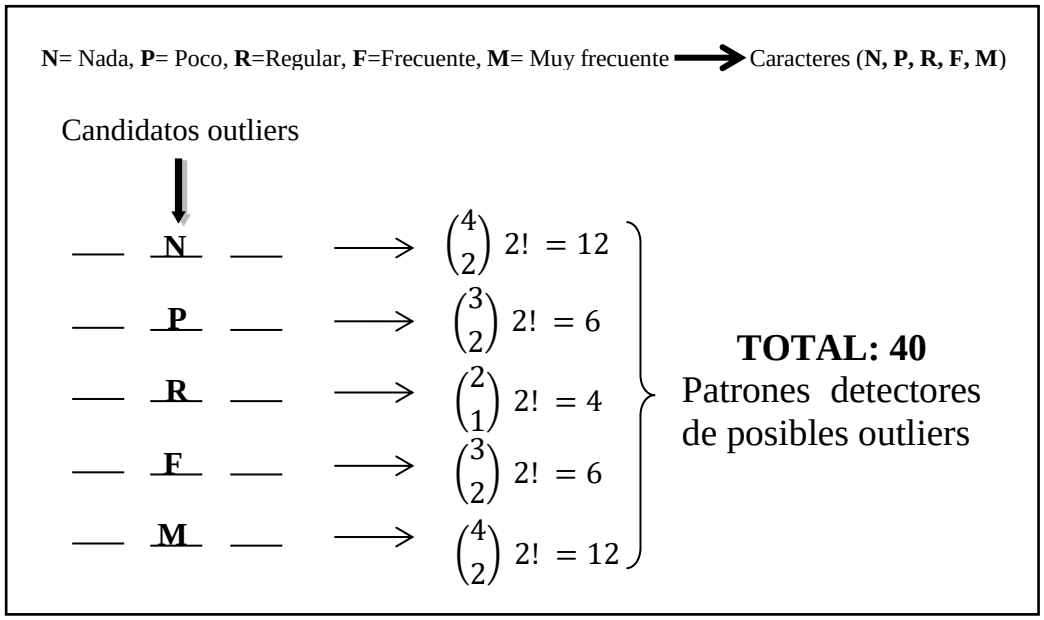


Figura 34. Implementación de patrones detectores de posibles outliers.

En la Tabla 8, se muestra la clasificación de los 40 patrones detectores de posibles outliers, producto de las combinaciones de cadenas no monótonas.

Tabla 8.- Patrones detectores de posibles outliers

Nº	CARACTERES														
	Carácter N			Carácter P			Carácter R			Carácter F			Carácter M		
1	P	N	M	M	P	F	M	R	F	N	F	P	N	M	P
2	P	N	R	M	P	R	F	R	M	N	F	R	N	M	R
3	P	N	F	F	P	R	N	R	P	P	F	N	N	M	F
4	R	N	M	F	P	M	P	R	N	P	F	R	P	M	N
5	R	N	P	R	P	F				R	F	N	P	M	R
6	R	N	F	R	P	M				R	F	P	P	M	F
7	F	N	M										F	M	N
8	F	N	R										F	N	P
9	F	N	P										F	M	R
10	M	N	P										R	M	P
11	M	N	R										R	M	N
12	M	N	F										R	M	F

Obteniéndose el conjunto de 40 patrones:

$$Patrones_{40} = \{PNM, PNR, PNF, RNM, RNP, RNF, FNM, FNR, FNP,$$
$$MNP, MNR, MNF, MPF, MPR, FPR, FPM, RPF, RPM, MRF, FRM, NRP, PRN, NFP, NFR, PFN, PFR,$$
$$RFN, RFP, NMP, NMR, NMF, PMN, PMR, PMF, FMN, FMP, FMR, RMP, RMN, RMF\}$$

La Figura 35 muestra el pseudocódigo del algoritmo DCONM (el código en R puede verse en el Anexo 1) que hace referencia al algoritmo KMP.

$\{Pre: TRUE\}$

$Lista = [\,];$

$texto = scan(Base\ de\ Datos\ Codificado\ );$

$\forall_{p\in}\ Patrones_{40}$

$P' = preKMP\ (P, texto)$

$KMP\ (P', texto, lista)$

$\{Post: lista\ contiene\ las\ posiciones\ de\ texto,\ donde\ se\ ha\ localizado\ un\ coincidente\ de\ P\}$

Figura 35. Pseudocódigo DOCNM.

Previamente a la implementación del algoritmo se ha realizado la clasificación y codificación de los datos de los pacientes drogodependientes como se muestran en la Tabla 6, 7 y 8.

En la Figura 36, se muestra un ejemplo realizado con el código del algoritmo DOCNM en R. Así mismo en la Figura 37, se muestran los resultados de la aplicación.

A continuación, se realizará un breve resumen del ejemplo. Para ello, en primer lugar, se realiza una lectura de datos de la cadena de texto de la variable i_coca (i_coca.txt) como se muestra en la Tabla 7.

En un segundo lugar, se realiza la lectura de datos de la cadena de texto del patrón (candidato5.txt), cabe mencionar que se han encontrado 40 patrones, los cuales se pueden visualizar en la Tabla 8.

En un tercer lugar, se aplica la función de PreKMP, esta función realiza previamente un procesado de los caracteres del patrón para encontrar unas coincidencias de prefijo con el patrón en sí, para ello se define el prefijo más grande del patrón [0..j-1] que es también un sufijo del patrón [1..j]. La función PreKMP puede determinar qué cambio (s) no es válido y directamente será descartado. También indica la cantidad de la última comparación la cual puede ser reutilizada si falla, de esta manera se evita el retroceso en la cadena de texto.

Finalmente, se ejecuta la función KMP. Esta función recibe la longitud del texto y la del patrón y procede a buscar de izquierda a derecha la coincidencia del patrón en la cadena de texto de la variable i_coca. Una vez ubicado uno o más coincidencias nos devuelve la posición en donde ha sido ubicado la coincidencia del patrón en la cadena de texto de la variable i_coca.

```
> texto<- scan("DATAKMP/DATA_ADICCION/i_coca.txt", what = 'character')
Read 1 item
> patron<- scan("DATAKMP/PATRON_ADICCION/Candidato33.txt", what = 'character')
Read 1 item
>
> PreKMP<-function(patron){
+   m<-nchar(patron)
+   prefix <- numeric(0)
+   prefix[1]<-0
+   a<-0
+   for(b in 2:m){
+     while (a>0 && substr(patron,a+1,a+1) != substr(patron,b,b))
+     {
+       a<-prefix[a]
+     }
+     if(substr(patron,a+1,a+1) == substr(patron,b,b))
+     {
+       a<-a+1
+     }
+     prefix[b]<-a
+   }
+   return (prefix)
+ }
>
> KMP<-function(texto,patron){
+   n<-nchar(texto)
+   m<-nchar(patron)
+   PRE<-c(PreKMP(patron))
+   q<-0
+   for(i in 1:n){
+     while (q>0 && substr(patron,q+1,q+1) != substr(texto,i,i))
+     {
+       q<-PRE[q]
+     }
+     if(substr(patron,q+1,q+1) == substr(texto,i,i))
+     {
+       q<-q+1
+     }
+     if(q==m){
+       #OJO SE HA AUMENTADO EN 2 PARA QUE EL USUARIO SE ENCUENTRE EN EL CENTRO
+       cat("\n EL USUARIO CANDIDATO OUTLIER ESTA EN LA POSICION : ",(i-m +2))
+       q<-PRE[q]
+     }
+   }
+ }
>
> PreKMP(patron)
[1] 0 0 0
> KMP(texto,patron)
```

Figura 36.- Código del Algoritmo DOCNM implementado en R.

```
EL USUARIO CANDIDATO OUTLIER ESTA EN LA POSICION:  52
```

Figura 37.- Resultados de la ejecución del Algoritmo DCONM.

Cabe mencionar que los resultados de la ejecución del algoritmo DOCNM a todas las variables drogodependientes se encuentra en el Anexo 7.

4.3.2 Método 2: Detección de Outliers con la Librería "*Outliers*" de R

En este apartado se detalla el resultado de un ejemplo para la detección de outliers con la librería "*outliers*" de R. Así mismo los datos utilizados para este proceso se muestran en la Tabla 9.

A continuación, se expone como se ha implementado la detección de outliers con la librería "*outliers*" de R. Para ello, en primer lugar, se instala la librería "*outliers*".

```
install.packages("outliers")
```

En un segundo lugar, se llama a la librería outliers

```
library(outliers)
```

En un tercer lugar, se realiza una lectura de los datos de la variable i_coca (Tabla 9)

```
i_coca<-scan("/DATA_ADICCION/PatronOutliers/i_coca.txt",sep = ",")
```

En un cuarto lugar, se realiza una prueba de chisq.out.test para la detección de outliers en un vector.

```
chisq.out.test(i_coca, variance=var(i_coca), opposite = FALSE)
```

```
        chi-squared test for outlier

data:  i_coca
X-squared = 6.2119, p-value = 0.01269
alternative hypothesis: highest value 36 is an outlier
```

Finalmente, Representamos gráficamente los outliers (Figura 38)

```
boxplot(i_coca, main='Boxplot i_coca outliers')
```

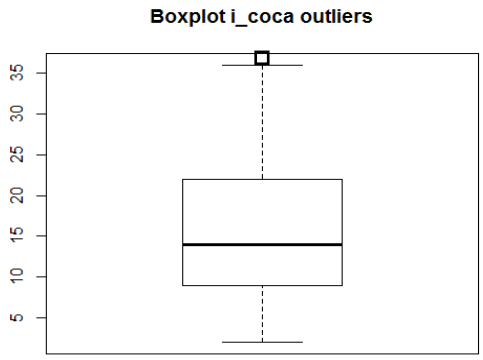


Figura 38. Gráfico de la detección de outliers con la librería *outliers* de R.

El resultado del ejemplo nos muestra un valor 36. Si observamos en la Tabla 9, específicamente en los datos de la variable i_coca, se puede observar que el dato con valor 36 se encuentra en la posición **52** de los datos de pacientes drogodependientes.

Así mismo los resultados de la ejecución del código con la librería "*outlier"* de R, a todas las variables drogodependientes se encuentra en el Anexo 8.

4.3.3 Método 3: Detección de Outliers con la Librería "*Rcmdr"* de R

En este apartado se detalla el resultado de un ejemplo para la detección de outliers con la librería "*Rcmdr"* de R, a través de regresión lineal simple. Para ver al detalle el ejemplo completo se encuentra en el Anexo 3. Así mismo los datos utilizados para este proceso se pueden observar en la Tabla 9.

La librería *Rcmdr* de R, proporciona una interfaz gráfica de usuario, que nos permite de una manera sencilla interpretar los coeficientes de un modelo lineal. En este trabajo he probado el test de valores atípicos de Bonferroni.

El test elegido se basa en detectar la presencia de valores atípicos y observar el gráfico "Residuals vs Leverage", que detecta valores no influyentes en la estimación del modelo.

El test de Bonferroni de Rcommander se basa en el método de corrección de Bonferroni para comparaciones múltiples [43, 44]. En la Figura 39 se muestra el test y el gráfico que nos indica que la observación número **53** es un valor atípico. Las observaciones **47**, **49**, **52** y **53** que vemos en el gráfico son medidas influyentes si llegan a ser atípicos. En la Figura 40 se muestra un gráfico de las distancias de Cook (J.Faraway, 2009).

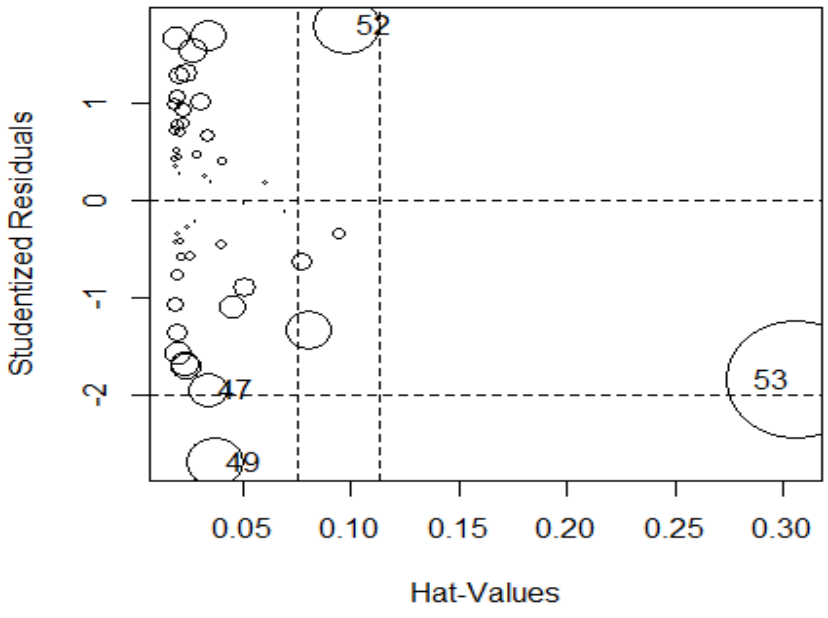


Figura 39. Diagrama de outliersTest

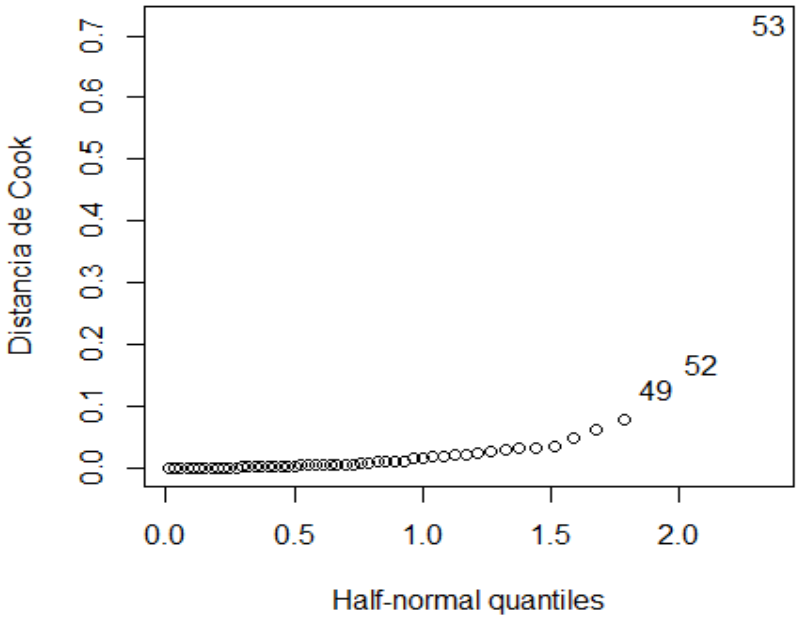


Figura 40. Gráfico de la distancia de Cook

En la Tabla 9 se muestran las variables (edad, i_oh, f_oh, i_tabaco, d_tabaco, i_bdz, f_bdz, i_hero, f_hero, i_coca, f_coca, i_canna, f_canna, i_anfet, f_anfet). Estos datos son utilizados por los dos métodos de detección de outliers (librería *Outliers* y *Rcmdr* de R). Para ello, se ha realizado una limpieza de datos previamente.

Tabla 9.- Clasificación de datos para detección de outliers con la librería "*outliers"* y *"Rcmdr"* de R

Edad → Adicción↓	**20**	**22**	**23**	**26**	**26**	**27**	**28**	**28**	**29**	**30**	**31**	**31**	**31**	**32**	**33**	**34**	**34**	**34**	**34**	**34**	**34**	**35**	**36**	**36**	**37**	**37**	**38**	**38**	**38**	**39**	**39**	**39**	**39**	**39**	**40**	**40**	**40**	**41**	**41**	**42**	**42**	**42**	**43**	**44**	**45**	**45**	**45**	**46**	**50**	**53**	**55**	**71**
i_oh	9	4	8			5	14		13	14	21	17			19	12			22	16		28	21	24	20	16	25	20		11		24			22	25	23	25	29	28			29	32	28		29		28			41	
f_oh	1	1	5			2	3		2	5	5	2	5		5	1			3	1		1	3	1	6	5	5	1		3		1			5	1	1	2	1	3			1	6	1		1		1			1	
i_tabaco	9	7				6		14			17	15						19					21		22	16												26				26	29	32	30		29		12			41	
d_tabaco	1	20				1		1			15	20						20					20		10	30												1				20	1	30	6		1		1			5	
i_bdz	7					4		3		9																				0		5										6							1			
f_bdz	1					1		1		1																				1		1										6							1			
i_hero					0					9													1																					26		22				26	36	
f_hero					6					6													1																					6		6				6	6	
i_coca	5	4	8	10	4	3	14	8	13	14	16	11	16	9	2	16	10	19	14	14	11	22	18	22	18	8	21	21	13	11	22	7	14	28	20	6	26	20	23	24	16	6	27	29	26	24	6	31	2	23	14	36	21
f_coca	1	1	5	2	1	2	3	1	2	1	5	3	4	2	1	1	2	1	5	3	3	1	1	3	4	5	1	5	1	3	2	1	3	3	5	1	2	1	1	1	5	2	1	1	1	3	4	1	1	3	1	1	5
i_canna	9	7	8	12	10		16	14		14				13	17	20		21	21	16		27	20	24	20	6	21			24	25	19			23	23			24				29		28		32					39	
f_canna	1		1	1	3		3	1		1				2	1	2		1	3	1		1	1	1	6	1	1			6	3	6			5	1			6				3		6		6					2	
i_anfet										1			15						12																									23										
f_anfet										6			6																															6										
Posición	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**

Cabe resaltar que la lectura de datos se realiza por cada variable en el software R. Por ejemplo, el formato de lectura de datos es de la siguiente manera:

5,4,8,10,4,3,14,8,13,14,16,11,16,9,2,16,10,19,14,14,11,22,18,22,18,8,21,21,13,11,22,7,14,28,20,6,26,20,23,24,16,6,27,29,26,24,6,31,2,23,14,36,21

Como se puede ver, el método 1, 2 y 3 arroja como outliers al paciente que se encuentra en la posición **52**. Este paciente es un candidato potencial a outliers y tiene que ser eliminado de los datos, para obtener unos datos fiables para posterior análisis. Así mismo para estos ejemplos se ha utilizado la variable "i_coca".

Los resultados de la detección de los outliers con la librería *Rcmdr* de R a todas las variables drogodependientes se encuentra en el Anexo 9.

4.4 Modelo de Predicción de Naïve Bayes

El algoritmo Naïve Bayes está basado en el teorema de Bayes y la probabilidad condicional. Dado un conjunto de instancias conocidas y una instancia nueva, predice la clase de esta instancia nueva.

En este trabajo se ha utilizado el algoritmo Naïve Bayes. Este algoritmo permite generar un modelo a partir de los datos históricos de pacientes drogodependientes (tabla de entrenamiento) y aplicar este modelo a los datos de un nuevo paciente drogodependiente (tabla de test). Con la finalidad de que arroje el resultado de su tratamiento (éxito o fracaso). Para ello se ha realizado la codificación y clasificación de los datos como se muestra en la Tabla 10. Este conjunto de datos sirve como una tabla de entrenamiento para el algoritmo de clasificación Naive Bayes.

A continuación, se muestra cómo se ha realizado la clasificación de datos por variables de interés:

EDAD (Edad del paciente cuando se registra en el sistema):

Edad (Años)	Código
15-16	A
17-18	B
19-20	C
21-22	D
23-24	E
25-30	F
31-35	G
36-50	H
51-70	I
71-90	J
NA	Z

SEXO:

Sexo	Código
M	A
F	B
NA	Z

LUGAR DE CONSUMO:

Lugar de Consumo						Código
Casa	**Calle**	**Bar**	**Trabajo**	**Coche**	**Indiferente**	
X	X	X	X			A
X		X		X		B
X	X	X				C
X	X					D
X		X				E
X				X		F
	X	X				G
X			X			H
			X	X		I
		X		X		J
X						K
	X					L
		X				M
			X			N
					X	O
NA						Z

HIJOS:

Hijos	Código
0	A
1	B
2	C
3	D
NA	Z

CONVIVENCIA:

Convivencia	Código
Familia de Origen	A
Pareja e Hijos	B
Pareja	C
Solo	D
Otros	E
NA	Z

ESTUDIOS:

Estudios	Código
Sin estudios	A
Primarios	B
Bachiller o fp	C
Universitarios	D
NA	Z

TRABAJO:

Trabajo	Código
Paro	A
ILT	B
Pensionista	C
NA	Z

ANTECEDENTES PENALES:

Antecedentes Penales	Código
Si	A
No	B
NA	Z

En lo que respecta a las adicciones (i_oh, i_tabaco, i_bdz, i_hero, i_coca, i_canna, i_ anfet), se realizó la clasificación de datos de acuerdo al rango de la edad de inicio de consumo como también en la frecuencia de consumo. Como primer paso se ha determinado trabajar con el año de consumo de la adicción, para eso es necesario tener un rango sobre el tiempo de consumo de la adicción correspondiente.

Usamos la siguiente fórmula para describir los años de consumo de la adicción:

$$\textbf{Año de consumo} = \textbf{Edad} - \textbf{Edad de Inicio del Consumo}$$

En este primer paso, el inicio de consumo se clasificó de la siguiente manera:

Año de consumo	Código
0-5	A
6-10	B
11-15	C
16-20	D
21-25	E
26-30	F
31-35	G
36-40	H
41-50	I
51-60	J
61-80	K
NA	Z

Como un segundo paso tenemos el dato de la frecuencia de consumo la cual está estipulada de la siguiente manera:

1. Todos los días
2. (4 a 5) días a la semana
3. (2 a 3) días a la semana
4. (1) día a la semana
5. Menos de un día a la semana
6. No consumió
9. Desconocido

Por lo cual, la frecuencia de consumo será la misma para todas las adicciones, quedando del siguiente modo:

Frecuencia de Consumo	Código
Todos los días	A
(4 a 5) días a la semana	B
(2 a 3) días a la semana	C
(1) día a la semana	D
Menos de un día a la semana	E
No consumió	F
Desconocido	G
NA	Z

I_CAIC:

Ingresos	Código
0	A
1	B
2	C
3	D
NA	Z

I_TRAT:

Para definir el año de inicio del tratamiento, se propuso la siguiente fórmula:

$$\textbf{Tiempo del Tratamiento} = \textbf{Edad} - \textbf{Edad de Inicio de Tratamiento}$$

De esta manera el inicio del tratamiento se clasificó de la siguiente forma:

Tiempo del Tratamiento	Código
0-1	A
2	B
3	C
4	D
5	E
6-10	F
11-15	G
16-20	H
21-30	I
31-50	J
51-80	K
NA	Z

INGPREV:

Ingresos Previos	Código
0	A
1	B
2	C
3	D
4	E
5	F
6	G

7-10	H
NA	Z

COCABASE:

Coca Base	Código
Si	A
No	B
NA	Z

TTOING:

Tratamiento de Ingreso						Código
AP	**BDZ**	**AD**	**Estabilizadores**	**DISULFIRAM**	**Strattera**	
X	X	X	X	X		A
X	X	X	X			B
X	X		X	X		C
	X	X	X			D
X	X	X				E
		X	X	X		F
	X	X	X			G
X	X		X			H
	X	X			X	I
X		X	X			J
X			X			K
	X	X				L
		X	X			M
	X		X			N
X	X					O
			X		X	P
X		X				Q
	X			X		Y
X						R
	X					S
		X				T
			X			U
				X		W
					X	X
NA						Z

DXPREV:

Diagnósticos Previos								Código
F2X	**Epilepsia Mioclonica J.**	**F31**	**TLP**	**F32**	**F4X**	**Otros TP**	**F90**	
X	X							A
		X	X					B
			X					C
		X						D
				X				E
					X			F
						X		G
							X	H
NA								Z

PESO:
El peso se encuentra en un rango establecido, de la siguiente manera:

Peso (Kg)	**Código**
20-30	A
31-40	B
41-45	C
46-50	D
51-55	E
56-60	F
61-65	G
66-70	H
71-75	I
76-80	J
81-85	K
86-90	L
91-95	M
96-100	N
101-110	O
111-130	P
131-150	Q
151-180	R
181-200	S
NA	Z

ALTURA:
La altura se encuentra en un rango establecido, de la siguiente manera:

Altura (cm)	**Código**
140-150	A
151-155	B
156-160	C
161-165	D
166-170	E
171-175	F
176-180	G
181-185	H
186-190	I
191-200	J
NA	Z

RESULTADO:

Resultado	Código
Bien	A
Mal	B
NA	Z

$$IMC = \frac{Peso(Kg)}{Altura^{2}\ (mts)}$$

IMC	´Descripción	Código
IMC < 16	Desnutrición Grado 3	A
16 IMC < 17	Desnutrición Grado 2	B
17 IMC < 18.5	Desnutrición Grado 1	C
18.5 IMC < 24.9	Normal	D
25 IMC < 29.9	Sobrepeso Grado 1	E
30 IMC < 40	Sobrepeso Grado 2	F
IMC < 40	Sobrepeso Grado 3	G
NA	NA	Z

En la Tabla 10 se muestra la clasificación de los datos históricos de pacientes drogodependientes.

En la Tabla 11 se muestra los candidatos posibles a outliers (con línea roja). Estos candidatos a outliers deben ser eliminados ya que han sido detectados por los tres métodos expuestos en el capítulo 3 (Algoritmo DOCNM, librerías *Outliers* y *Rcmdr* de R). Por tal motivo la Tabla 11 será utilizada para el análisis de predicción en el Caso 2 de la sección 4.2.3.

Tabla 10. Clasificación de datos históricos de pacientes con drogodependencias

idPaciente	edad	caid	sexo	lugar consumo	hijos	convivencia	estudios	trabajo	antecedentes	i_oh	f_oh	i_tabaco	d_tabaco	i_bdz	f_bdz	i_hero	f_hero	i_coca	f_coca	i_canna	f_canna	i_anfet	f_anfet	i_caic	i_trat	ingprev	cocabase	ttoing	dxprev	imc	resultado
1	G	Norte	Hombre	NA	A	A	NA	NA	B	C	A	NA	NA	NA	NA	NA	NA	D	A	D	B	NA	NA	NA	F	D	NA	G	NA	C	MAL
2	H	Getafe	Hombre	B	A	D	B	B	B	D	F	E	A	NA	NA	NA	NA	D	D	D	F	NA	NA	A	D	A	B	NA	NA	F	MAL
3	F	Fuenlabrada	Mujer	G	B	A	C	NA	B	C	C	NA	NA	NA	NA	NA	NA	C	C	D	C	NA	NA	A	A	A	B	E	NA	C	MAL
4	H	Tetuan	Hombre	D	A	A	D	A	B	E	A	NA	NA	NA	NA	NA	NA	B	A	E	A	NA	NA	NA	NA	NA	B	F	C	F	MAL
5	G	Collado Villalba	Hombre	O	B	A	NA	A	B	E	E	D	E	NA	NA	NA	NA	D	E	NA	NA	NA	NA	NA	H	A	B	W	NA	D	MAL
6	F	Torrejón	Hombre	O	A	A	B	B	B	C	B	NA	NA	NA	NA	NA	NA	C	B	NA	NA	NA	NA	NA	F	A	B	B	NA	D	MAL
7	G	Norte	Hombre	O	A	A	C	B	B	NA	NA	NA	NA	NA	NA	NA	NA	B	B	C	B	NA	NA	B	G	B	NA	K	D	E	MAL
8	H	Getafe	Hombre	F	A	A	B	A	B	E	C	E	B	NA	NA	A	A	D	A	D	A	NA	NA	B	F	NA	A	S	NA	D	MAL
9	H	Norte	Hombre	K	B	B	C	A	A	NA	NA	NA	NA	NA	NA	NA	NA	E	C	NA	NA	NA	NA	B	F	B	A	G	E	F	MAL
10	H	Alcala	Hombre	K	B	B	B	A	B	G	F	G	C	NA	NA	NA	NA	F	A	NA	NA	NA	NA	B	H	B	B	L	F	B	MAL
11	F	Getafe	Hombre	O	A	A	A	NA	B	C	E	NA	NA	B	A	B	F	C	A	C	A	A	F	A	F	C	B	H	NA	E	MAL
12	H	Cad Vallecas	Hombre	K	C	A	B	B	B	NA	NA	NA	NA	A	A	NA	NA	E	C	NA	NA	NA	NA	NA	NA	C	B	H	G	D	MAL
13	G	Tetuan	Hombre	L	C	B	B	A	B	NA	NA	NA	NA	NA	NA	NA	NA	B	B	NA	NA	NA	NA	B	F	B	B	I	H	D	MAL
14	F	Los Pinares	Hombre	C	A	A	A	A	A	NA	NA	C	A	A	A	NA	NA	B	A	C	A	NA	NA	NA	NA	NA	NA	S	H	D	MAL
15	H	San Fernando de Henares	Hombre	K	B	D	C	B	B	F	A	C	A	NA	NA	NA	NA	A	A	NA	NA	NA	NA	NA	B	NA	B	M	E	C	MAL
16	F	Aranjuez	Hombre	K	C	B	NA	B	B	NA	NA	NA	NA	NA	NA	NA	NA	B	B	C	A	NA	NA	NA	A	NA	B	N	NA	E	MAL
17	H	Vallecas	Hombre	O	A	D	C	A	A	F	A	NA	NA	NA	NA	NA	NA	E	A	E	F	NA	NA	NA	F	NA	B	H	B	D	MAL
18	H	Norte	Hombre	C	A	D	C	B	NA	C	C	NA	NA	A	A	NA	NA	C	C	E	F	NA	NA	A	NA	C	B	J	NA	D	MAL
19	H	Este	Hombre	NA	C	C	B	B	A	NA	NA	NA	NA	NA	NA	NA	NA	E	B	E	C	NA	NA	NA	NA	NA	NA	O	NA	D	MAL
20	H	Este	Hombre	E	C	NA	C	B	NA	E	A	NA	NA	NA	NA	NA	NA	F	B	NA	NA	NA	NA	B	NA	B	B	O	NA	D	MAL
21	H	Aranjuez	Hombre	E	C	A	C	B	NA	E	A	NA	NA	A	B	NA	NA	B	A	D	F	NA	NA	A	C	NA	B	S	H	E	MAL
22	F	Getafe	Hombre	A	A	A	C	A	A	A	B	B	A	A	A	NA	NA	A	B	NA	NA	NA	NA	NA	NA	NA	B	T	NA	E	MAL
23	J	Getafe	Mujer	K	C	E	B	C	A	NA	NA	NA	NA	NA	NA	NA	NA	E	E	NA	NA	NA	NA	NA	NA	NA	B	S	NA	D	MAL
24	H	Parla	Hombre	NA	A	A	A	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	C	C	NA	NA	NA	NA	NA	A	A	B	G	NA	F	MAL

25	H	Leganes	Hombre	D	D	B	NA	B	A	F	A	F	A	NA	NA	E	F	B	D	G	F	NA	NA	NA	F	NA	A	L	NA	F	MAL
26	E	M105	Mujer	D	A	C	B	C	A	B	E	NA	NA	NA	NA	NA	NA	B	E	B	A	NA	NA	A	C	A	NA	NA	C	D	MAL
27	H	Hortaleza	Hombre	K	A	C	B	B	B	NA	NA	NA	NA	NA	NA	NA	NA	G	A	NA	NA	NA	NA	NA	NA	NA	A	NA	NA	D	MAL
28	C	Parla	Mujer	NA	A	A	B	A	B	B	1	B	A	B	A	NA	NA	A	A	D	A	NA	NA	NA	NA	NA	B	NA	NA	E	MAL
29	H	Getafe	Hombre	D	A	C	A	A	A	E	1	NA	NA	NA	NA	NA	NA	E	C	E	A	NA	NA	NA	G	A	B	NA	NA	D	MAL
30	G	Parla	Hombre	D	C	A	B	A	B	NA	NA	D	B	NA	NA	NA	NA	D	A	E	A	NA	NA	C	G	C	B	A	NA	D	MAL
31	H	Este	Hombre	M	A	E	C	A	B	NA	NA	NA	NA	NA	NA	NA	NA	D	E	NA	NA	NA	NA	B	G	C	B	G	NA	NA	MAL
32	F	San Blas	Hombre	L	A	A	C	A	B	NA	NA	NA	NA	NA	NA	A	F	A	A	B	C	NA	NA	A	NA	A	B	NA	NA	C	MAL
33	H	Vallecas	Hombre	NA	D	B	B	B	B	F	C	NA	NA	NA	NA	NA	NA	E	A	NA	NA	NA	NA	NA	NA	NA	B	T	NA	F	MAL
34	G	Majadahonda	Hombre	O	A	A	D	B	B	D	E	NA	NA	NA	NA	NA	NA	A	A	D	A	NA	NA	NA	G	C	B	K	NA	D	MAL
35	G	Norte	Hombre	K	B	A	A	B	B	F	A	NA	NA	NA	NA	NA	NA	E	A	F	A	NA	NA	C	G	C	B	NA	NA	D	MAL
36	G	NA	Hombre	C	B	A	C	A	B	E	C	NA	NA	NA	NA	NA	NA	C	E	E	C	C	NA	B	F	B	B	E	NA	E	MAL
37	I	Vallecas	Hombre	NA	B	C	NA	C	B	I	A	J	E	NA	NA	H	F	H	A	H	B	NA	NA	A	I	NA	A	S	NA	D	MAL
38	H	Majadahonda	Hombre	K	A	A	B	A	B	E	E	NA	NA	NA	NA	NA	NA	E	A	E	A	NA	NA	NA	NA	NA	B	K	A	D	BIEN
39	G	Sur	Hombre	I	A	A	B	B	B	D	B	C	B	NA	NA	NA	NA	C	C	NA	NA	NA	NA	A	F	A	B	NA	H	F	BIEN
40	G	Este	Hombre	L	B	A	C	A	B	NA	E	NA	NA	NA	NA	NA	NA	D	D	NA	NA	C	F	A	F	A	B	C	F	D	BIEN
41	H	Cad Villaverde	Hombre	D	C	B	C	A	B	NA	NA	NA	NA	NA	NA	NA	NA	F	C	NA	NA	NA	NA	NA	A	A	B	P	H	NA	BIEN
42	H	Alcorcon	Hombre	K	C	A	C	A	B	E	E	NA	NA	NA	NA	NA	NA	D	E	E	E	NA	NA	NA	NA	NA	B	Q	E	D	BIEN
43	H	Cad Vallecas	Hombre	M	D	A	B	A	B	D	A	NA	NA	NA	NA	NA	NA	E	E	NA	NA	NA	NA	NA	B	A	B	U	NA	E	BIEN
44	G	Leganes	Hombre	L	B	A	B	A	B	D	A	NA	NA	NA	NA	NA	NA	C	C	D	A	NA	NA	B	G	B	B	NA	NA	D	BIEN
45	G	Mostoles	Hombre	F	A	A	B	A	B	NA	NA	NA	NA	NA	NA	NA	NA	C	C	NA	NA	NA	NA	NA	E	A	NA	NA	NA	E	BIEN
46	D	Aranjuez	Hombre	J	A	A	B	B	A	B	A	A	B	NA	NA	NA	NA	A	A	NA	NA	NA	NA	NA	NA	NA	B	NA	H	E	BIEN
47	H	Alcala	Hombre	C	B	C	B	B	A	D	E	D	C	NA	NA	NA	NA	B	E	E	A	NA	NA	B	A	C	B	S	E	F	BIEN
48	H	Mostoles	Mujer	H	C	B	C	B	B	NA	NA	F	B	B	F	NA	NA	B	B	NA	NA	NA	NA	A	A	A	B	E	NA	D	BIEN
49	H	Cad Tetuan	Hombre	D	A	A	B	A	B	E	B	F	A	NA	NA	NA	NA	D	A	NA	NA	NA	NA	NA	A	NA	A	L	NA	D	BIEN
50	H	NA	Hombre	O	D	A	B	B	B	F	A	F	A	NA	NA	NA	NA	F	A	F	C	E	F	NA	F	A	B	Y	NA	E	BIEN
51	H	Alcorcon	Mujer	M	B	A	C	A	B	NA	NA	NA	NA	NA	NA	NA	NA	C	A	NA	NA	NA	NA	B	NA	C	B	T	C	D	BIEN
52	H	M105	Hombre	L	A	A	C	C	B	F	A	F	F	NA	NA	F	F	F	A	F	F	NA	NA	B	I	G	B	B	D	D	BIEN
53	I	Ctd Centro	Hombre	K	B	B	D	A	NA	NA	NA	NA	NA	NA	NA	F	F	C	1	NA	NA	NA	NA	C	G	C	A	N	NA	D	BIEN

Tabla 11. Datos históricos sin la presencia de Outliers (Detectados en los tres métodos del estudio)

idPaciente	edad	caid	sexo	lugar consumo	hijos	convivencia	estudios	trabajo	antecedentes	i_oh	f_oh	i_tabaco	d_tabaco	i_bdz	f_bdz	i_hero	f_hero	i_coca	f_coca	i_canna	f_canna	i_anfet	f_anfet	i_caic	i_trat	ingprev	cocabase	ttoing	dxprev	imc	resultado
1	G	Norte	Hombre	NA	A	A	NA	NA	B	C	A	NA	NA	NA	NA	NA	NA	D	A	D	B	NA	NA	NA	F	D	NA	G	NA	C	MAL
~~2~~	~~H~~	~~Getafe~~	~~Hombre~~	~~B~~	~~A~~	~~D~~	~~B~~	~~B~~	~~B~~	~~D~~	~~F~~	~~E~~	~~A~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~D~~	~~D~~	~~D~~	~~F~~	~~NA~~	~~NA~~	~~A~~	~~D~~	~~A~~	~~B~~	~~NA~~	~~NA~~	~~F~~	~~MAL~~
3	F	Fuenlabrada	Mujer	G	B	A	C	NA	B	C	C	NA	NA	NA	NA	NA	NA	C	C	D	C	NA	NA	A	A	A	B	E	NA	C	MAL
4	H	Tetuan	Hombre	D	A	A	D	A	B	E	A	NA	NA	NA	NA	NA	NA	B	A	E	A	NA	NA	NA	NA	NA	B	F	C	F	MAL
~~5~~	~~G~~	~~Collado Villalba~~	~~Hombre~~	~~O~~	~~B~~	~~A~~	~~NA~~	~~A~~	~~B~~	~~E~~	~~E~~	~~D~~	~~E~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~D~~	~~E~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~H~~	~~A~~	~~B~~	~~W~~	~~NA~~	~~D~~	~~MAL~~
6	F	Torrejón	Hombre	O	A	A	B	B	B	C	B	NA	NA	NA	NA	NA	NA	C	B	NA	NA	NA	NA	NA	F	A	B	B	NA	D	MAL
7	G	Norte	Hombre	O	A	A	C	B	B	NA	NA	NA	NA	NA	NA	NA	NA	B	B	C	B	NA	NA	B	G	B	NA	K	D	E	MAL
8	H	Getafe	Hombre	F	A	A	B	A	B	E	C	E	B	NA	NA	A	A	D	A	D	A	NA	NA	B	F	NA	A	S	NA	D	MAL
9	H	Norte	Hombre	K	B	B	C	A	A	NA	NA	NA	NA	NA	NA	NA	NA	E	C	NA	NA	NA	NA	B	F	B	A	G	E	F	MAL
10	H	Alcala	Hombre	K	B	B	B	A	B	G	F	G	C	NA	NA	NA	NA	F	A	NA	NA	NA	NA	B	H	B	B	L	F	B	MAL
~~11~~	~~F~~	~~Getafe~~	~~Hombre~~	~~O~~	~~A~~	~~A~~	~~A~~	~~NA~~	~~B~~	~~C~~	~~E~~	~~NA~~	~~NA~~	~~B~~	~~A~~	~~B~~	~~F~~	~~C~~	~~A~~	~~C~~	~~A~~	~~A~~	~~F~~	~~A~~	~~F~~	~~C~~	~~B~~	~~H~~	~~NA~~	~~E~~	~~MAL~~
12	H	Cad Vallecas	Hombre	K	C	A	B	B	B	NA	NA	NA	NA	A	A	NA	NA	E	C	NA	NA	NA	NA	NA	NA	C	B	H	G	D	MAL
13	G	Tetuan	Hombre	L	C	B	B	A	B	NA	NA	NA	NA	NA	NA	NA	NA	B	B	NA	NA	NA	NA	B	F	B	B	I	H	D	MAL
14	F	Los Pinares	Hombre	C	A	A	A	A	A	NA	NA	C	A	A	A	NA	NA	B	A	C	A	NA	NA	NA	NA	NA	NA	S	H	D	MAL
15	H	San Fernando de Henares	Hombre	K	B	D	C	B	B	F	A	C	A	NA	NA	NA	NA	A	A	NA	NA	NA	NA	NA	B	NA	B	M	E	C	MAL
16	F	Aranjuez	Hombre	K	C	B	NA	B	B	NA	NA	NA	NA	NA	NA	NA	NA	B	B	C	A	NA	NA	NA	A	NA	B	N	NA	E	MAL
17	H	Vallecas	Hombre	O	A	D	C	A	A	F	A	NA	NA	NA	NA	NA	NA	E	A	E	F	NA	NA	NA	F	NA	B	H	B	D	MAL
~~18~~	~~H~~	~~Norte~~	~~Hombre~~	~~C~~	~~A~~	~~D~~	~~C~~	~~B~~	~~NA~~	~~C~~	~~C~~	~~NA~~	~~NA~~	~~A~~	~~A~~	~~NA~~	~~NA~~	~~C~~	~~C~~	~~E~~	~~F~~	~~NA~~	~~NA~~	~~A~~	~~NA~~	~~C~~	~~B~~	~~J~~	~~NA~~	~~D~~	~~MAL~~
19	H	Este	Hombre	NA	C	C	B	B	A	NA	NA	NA	NA	NA	NA	NA	NA	E	B	E	C	NA	NA	NA	NA	NA	NA	O	NA	D	MAL
20	H	Este	Hombre	E	C	NA	C	B	NA	E	A	NA	NA	NA	NA	NA	NA	F	B	NA	NA	NA	NA	B	NA	B	B	O	NA	D	MAL
21	H	Aranjuez	Hombre	E	C	A	C	B	NA	E	A	NA	NA	A	B	NA	NA	B	A	D	F	NA	NA	A	C	NA	B	S	H	E	MAL
22	F	Getafe	Hombre	A	A	A	C	A	A	A	B	B	A	A	A	NA	NA	A	B	NA	NA	NA	NA	NA	NA	NA	B	T	NA	E	MAL
23	J	Getafe	Mujer	K	C	E	B	C	A	NA	NA	NA	NA	NA	NA	NA	NA	E	E	NA	NA	NA	NA	NA	NA	NA	B	S	NA	D	MAL
24	H	Parla	Hombre	NA	A	A	A	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	C	C	NA	NA	NA	NA	NA	A	A	B	G	NA	F	MAL

~~25~~	~~H~~	~~Leganes~~	~~Hombre~~	~~D~~	~~D~~	~~B~~	~~NA~~	~~B~~	~~A~~	~~F~~	~~A~~	~~F~~	~~A~~	~~NA~~	~~NA~~	~~E~~	~~F~~	~~B~~	~~D~~	~~G~~	~~F~~	~~NA~~	~~NA~~	~~NA~~	~~F~~	~~NA~~	~~A~~	~~L~~	~~NA~~	~~F~~	~~MAL~~
~~26~~	~~E~~	~~M105~~	~~Mujer~~	~~D~~	~~A~~	~~C~~	~~B~~	~~C~~	~~A~~	~~B~~	~~E~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~B~~	~~E~~	~~B~~	~~A~~	~~NA~~	~~NA~~	~~A~~	~~C~~	~~A~~	~~NA~~	~~NA~~	~~C~~	~~D~~	~~MAL~~
27	H	Hortaleza	Hombre	K	A	C	B	B	B	NA	NA	NA	NA	NA	NA	NA	NA	G	A	NA	NA	NA	NA	NA	NA	A	NA	NA	D	MAL
28	C	Parla	Mujer	NA	A	A	B	A	B	B	1	B	A	B	A	NA	NA	A	A	D	A	NA	NA	NA	NA	NA	B	NA	NA	E	MAL
29	H	Getafe	Hombre	D	A	C	A	A	A	E	1	NA	NA	NA	NA	NA	NA	E	C	E	A	NA	NA	NA	G	A	B	NA	NA	D	MAL
30	G	Parla	Hombre	D	C	A	B	A	B	NA	NA	D	B	NA	NA	NA	NA	D	A	E	A	NA	NA	C	G	C	B	A	NA	D	MAL
~~31~~	~~H~~	~~Este~~	~~Hombre~~	~~M~~	~~A~~	~~E~~	~~C~~	~~A~~	~~B~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~D~~	~~E~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~B~~	~~G~~	~~C~~	~~B~~	~~G~~	~~NA~~	~~NA~~	~~MAL~~
32	F	San Blas	Hombre	L	A	A	C	A	B	NA	NA	NA	NA	NA	NA	A	F	A	A	B	C	NA	NA	A	NA	A	B	NA	NA	C	MAL
33	H	Vallecas	Hombre	NA	D	B	B	B	B	F	C	NA	NA	NA	NA	NA	NA	E	A	NA	NA	NA	NA	NA	NA	NA	B	T	NA	F	MAL
34	G	Majadahonda	Hombre	O	A	A	D	B	B	D	E	NA	NA	NA	NA	NA	NA	A	A	D	A	NA	NA	NA	G	C	B	K	NA	D	MAL
35	G	Norte	Hombre	K	B	A	A	B	B	F	A	NA	NA	NA	NA	NA	NA	E	A	F	A	NA	NA	C	G	C	B	NA	NA	D	MAL
36	G	NA	Hombre	C	B	A	C	A	B	E	C	NA	NA	NA	NA	NA	NA	C	E	E	C	C	NA	B	F	B	B	E	NA	E	MAL
~~37~~	~~I~~	~~Vallecas~~	~~Hombre~~	~~NA~~	~~B~~	~~C~~	~~NA~~	~~C~~	~~B~~	~~I~~	~~A~~	~~J~~	~~E~~	~~NA~~	~~NA~~	~~H~~	~~F~~	~~H~~	~~A~~	~~H~~	~~B~~	~~NA~~	~~NA~~	~~A~~	~~I~~	~~NA~~	~~A~~	~~S~~	~~NA~~	~~D~~	~~MAL~~
38	H	Majadahonda	Hombre	K	A	A	B	A	B	E	E	NA	NA	NA	NA	NA	NA	E	A	E	A	NA	NA	NA	NA	NA	B	K	A	D	BIEN
39	G	Sur	Hombre	I	A	A	B	B	B	D	B	C	B	NA	NA	NA	NA	C	C	NA	NA	NA	NA	A	F	A	B	NA	H	F	BIEN
~~40~~	~~G~~	~~Este~~	~~Hombre~~	~~L~~	~~B~~	~~A~~	~~C~~	~~A~~	~~B~~	~~NA~~	~~E~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~D~~	~~D~~	~~NA~~	~~NA~~	~~C~~	~~F~~	~~A~~	~~F~~	~~A~~	~~B~~	~~C~~	~~F~~	~~D~~	~~BIEN~~
41	H	Cad Villaverde	Hombre	D	C	B	C	A	B	NA	NA	NA	NA	NA	NA	NA	NA	F	C	NA	NA	NA	NA	NA	A	A	B	P	H	NA	BIEN
42	H	Alcorcon	Hombre	K	C	A	C	A	B	E	E	NA	NA	NA	NA	NA	NA	D	E	E	E	NA	NA	NA	NA	NA	B	Q	E	D	BIEN
43	H	Cad Vallecas	Hombre	M	D	A	B	A	B	D	A	NA	NA	NA	NA	NA	NA	E	E	NA	NA	NA	NA	NA	B	A	B	U	NA	E	BIEN
44	G	Leganes	Hombre	L	B	A	B	A	B	D	A	NA	NA	NA	NA	NA	NA	C	C	D	A	NA	NA	B	G	B	B	NA	NA	D	BIEN
45	G	Mostoles	Hombre	F	A	A	B	A	B	NA	NA	NA	NA	NA	NA	NA	NA	C	C	NA	NA	NA	NA	NA	E	A	NA	NA	NA	E	BIEN
46	D	Aranjuez	Hombre	J	A	A	B	B	A	B	A	A	B	NA	NA	NA	NA	A	A	NA	NA	NA	NA	NA	NA	NA	B	NA	H	E	BIEN
~~47~~	~~H~~	~~Alcala~~	~~Hombre~~	~~C~~	~~B~~	~~C~~	~~B~~	~~B~~	~~A~~	~~D~~	~~E~~	~~D~~	~~C~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~B~~	~~E~~	~~E~~	~~A~~	~~NA~~	~~NA~~	~~B~~	~~A~~	~~C~~	~~B~~	~~S~~	~~E~~	~~F~~	~~BIEN~~
~~48~~	~~H~~	~~Mostoles~~	~~Mujer~~	~~H~~	~~C~~	~~B~~	~~C~~	~~B~~	~~B~~	~~NA~~	~~NA~~	~~F~~	~~B~~	~~B~~	~~F~~	~~NA~~	~~NA~~	~~B~~	~~B~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~A~~	~~A~~	~~A~~	~~B~~	~~E~~	~~NA~~	~~D~~	~~BIEN~~
49	H	Cad Tetuan	Hombre	D	A	A	B	A	B	E	B	F	A	NA	NA	NA	NA	D	A	NA	NA	NA	NA	NA	A	NA	A	L	NA	D	BIEN
50	H	NA	Hombre	O	D	A	B	B	B	F	A	F	A	NA	NA	NA	NA	F	A	F	C	E	F	NA	F	A	B	Y	NA	E	BIEN
51	H	Alcorcon	Mujer	M	B	A	C	A	B	NA	NA	NA	NA	NA	NA	NA	NA	C	A	NA	NA	NA	NA	B	NA	C	B	T	C	D	BIEN
52	H	M105	Hombre	L	A	A	C	C	B	F	A	F	F	NA	NA	F	F	F	A	F	F	NA	NA	B	I	G	B	B	D	D	BIEN
53	I	Ctd Centro	Hombre	K	B	B	D	A	NA	NA	NA	NA	NA	NA	NA	F	F	C	1	NA	NA	NA	NA	C	G	C	A	N	NA	D	BIEN

Instancia Nueva (tabla de test)

Una vez entrenado nuestro algoritmo Naive Bayes con el conjunto de datos históricos de pacientes tratados antes de su ingreso en la clínica, introducimos una instancia nueva (Nuevo Paciente) en la que se pretende predecir la clase de esta instancia nueva (tratamiento exitoso o fracaso), para poder tomar decisiones sobre su ingreso o derivación a otros centros. Como ejemplo tendremos la siguiente instancia nueva (Tabla 12):

Tabla 12. Instancia nueva para predecir con el algoritmo Naive Bayes

idPaciente	edad	caid	sexo	lugar consumo	hijos	convivencia	estudios	trabajo	antecedentes	i_oh	f_oh	i_tabaco	d_tabaco	i_bdz	f_bdz	i_hero	f_hero	i_coca	f_coca	i_canna	f_canna	i_anfet	f_anfet	i_caic	i_trat	ingprev	cocabase	ttoing	dxprev	imc	resultado
111	H	Mostoles	Hombre	E	B	C	B	B	B	C	B	NA	NA	NA	NA	NA	NA	C	G	NA	NA	NA	NA	NA	G	A	B	L	G	E	?

Para un mejor entendimiento del análisis predictivo con el algoritmo Naïves Bayes de R, se han desarrollado dos casos: (1) aplicación en una tabla de entrenamiento con datos completos, (2) aplicación en una tabla de entrenamiento con datos libre de outliers. Los detalles se exponen a continuación (en rojo se muestran las salidas de ejecución en el entorno R).

4.4.1 Caso 1: Aplicación en la Base de Datos Original

Para este caso se ha utilizado la totalidad de los datos históricos de la base de datos de pacientes drogodependientes. Para ello previamente se ha realizado la clasificación de los datos como se muestra en la Tabla 10 con el propósito de realizar el test de entrenamiento del algoritmo Naive Bayes. Así mismo se utiliza una instancia nueva como se muestra en la Tabla 12 con el propósito de predecir la clase de esta nueva instancia.

- Como primer paso se instala el paquete **"e1071"** de **R.**

```
install.packages("e1071", dependencies = TRUE)
library(e1071)
```

- Se crean las tablas (o data frame) **HistoricoPacientesCaso1 (Tabla 10)** y **Nuevos_Pacientes**

```
HistoricoPacientesCaso1 <- read.csv
("C:/Users/HistoricoPacientesCaso1.csv")
```
tabla de entrenamiento

```
NuevosPacientes <- read.csv ("C:/Users/NuevoPaciente.csv")
```
tabla de test

- Se crean Probabilidades con **HistoricoPacientesCaso1** y se asigna una predicción a **NuevosPacientes**

```
Probabilidades <- naiveBayes (resultado ~., data=
HistoricoPacientesCaso1 [-1])
```
crea tabla Probabilidades

```
Prediccion <- predict (Probabilidades, NuevosPacientes [, -
32]);
```
crea predicción a datos nuevos

- Se añade la columna de Predicción a la tabla NuevosPacientes

```
NuevosPacientes$Prediccion_de_Resultado <- Prediccion
```

- Finalmente se guarda la tabla de NuevosPacientes con la predicción:

```
write.csv(NuevosPacientes,"C:/Users/Predict_NuevosPacientes
.csv")
```

```
table(Prediccion, NuevosPacientes[,33])
```

#El -32 se debe a que la variable dependiente, **Prediccion_de_Resultado**, es el número de columna 33.

```
Prediccion BIEN MAL
    BIEN    0   0
    MAL     0   1
```

```
print(Probabilidades)
```

Brinda información del algoritmo de clasificación **Naive Bayes**

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
     BIEN       MAL
0.3018868 0.6981132

Conditional probabilities:
      edad
Y               C          D          E          F          G
H          I          J
  BIEN 0.00000000 0.06250000 0.00000000 0.00000000 0.25000000 0.
62500000 0.06250000 0.00000000
  MAL  0.02702703 0.00000000 0.02702703 0.18918919 0.21621622 0.
48648649 0.02702703 0.02702703
```

… (Puede verse la salida completa en el Anexo 10-A)

```
Prediccion
```

Brinda información de la predicción obtenida al correr el algoritmo Naive Bayes

```
[1] MAL
Levels: BIEN MAL
```

```
plot(Prediccion)
```

Nos muestra gráficamente la predicción obtenida.

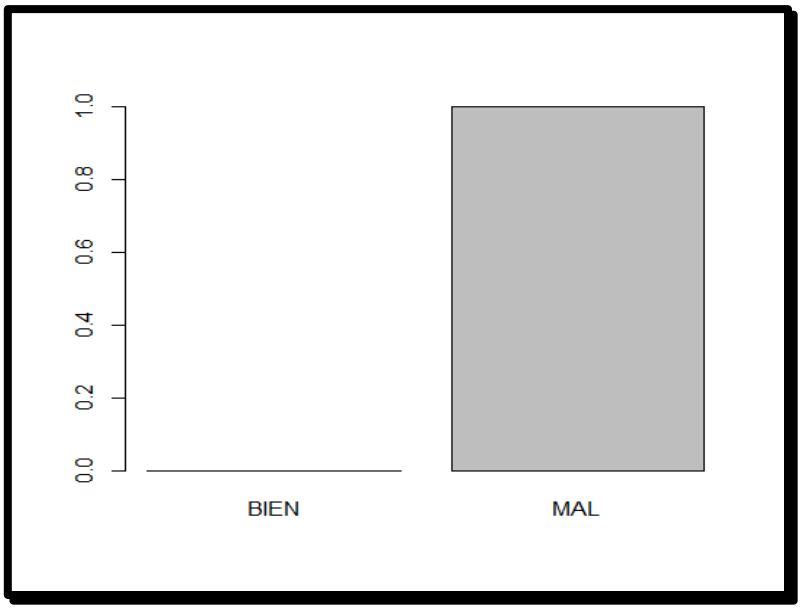


Figura 41.- Gráfico de la predicción con R en el caso 1.

En la Figura 41, se muestra gráficamente los resultados de la predicción obtenida con la ejecución del algoritmo Naive Bayes del paquete *e1071* de R. Cabe mencionar que en este caso se ha aplicado previamente un pre-procesado de los datos de los pacientes drogodependientes. Así mismo, no se ha realizado una limpieza de los outliers detectados en el Capítulo 3.

4.4.2 Caso 2: Aplicación en la Base de Datos libre de Outliers

Este caso es similar al Caso 1 con la diferencia que se ha realizado una limpieza de datos (Data Cleaning) de la base de datos de los pacientes drogodependientes, eliminando todos los pacientes que se han encontrado como posibles outliers. Estos outliers han sido detectados con el algoritmo DOCNM con patrones no monótonos y las librerías "Outliers" y "Rcmdr" de R como se puede observar en la Tabla 9. Así mismo se utiliza una instancia nueva como se muestra en la Tabla 10 con el propósito de predecir la clase de esta instancia.

Cabe mencionar que la predicción obtenida por el algoritmo de Naive Bayes en este caso, es buena y fiable, como resultado de la eliminación de los outliers, ya que los outliers aumentan el error en la predicción.

- Para el Caso 1 ya se instaló el paquete **"e1071"** de R, para este caso solo llamaremos a dicha librería.

```
library(e1071)
```

- Se crean las tablas (o data frame) **HistoricoPacientesCaso2 (anexo 8)** y **Nuevos_Pacientes**

```
HistoricoPacientesCaso2 <-read.csv
("C:/Users/HistoricoPacientesCaso2.csv")
```
tabla de entrenamiento

```
NuevosPacientes <- read.csv ("C:/Users/NuevoPaciente.csv")
```

tabla de test

- Se Crean Probabilidades con el **HistoricoPacientesCaso2** y asigna predicción a **NuevosPacientes**

```
Probabilidades <- naiveBayes (resultado ~., data=
HistoricoPacientesCaso2 [-1])
```
crea tabla Probabilidades

```
Prediccion <- predict (Probabilidades, NuevosPacientes [, -
32]);
```
crea predicción a datos nuevos

- Se añade la columna de Prediccion a la tabla NuevosPacientes

```
NuevosPacientes$Prediccion_de_Resultado <- Prediccion
```

- Finalmente se guarda la tabla de NuevosPacientes con la predicción:

```
write.csv(NuevosPacientes,"C:/Users/Predict_NuevosPacientes
.csv")
```

```
table(Prediccion, NuevosPacientes[,33])
```
#El -32 se debe a que la variable dependiente, **`Prediccion_de_Resultado`**, es en número de columna 33.

Prediccion BIEN MAL

BIEN 0 0
MAL 0 1

```
print(Probabilidades)
```

Brinda información del algoritmo de clasificación **Naive Bayes**

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
     BIEN       MAL
0.3095238 0.6904762

Conditional probabilities:
```

```
       edad
 Y               C          D          F          G          H
 I          J
   BIEN 0.00000000 0.07692308 0.00000000 0.23076923 0.61538462 0.
 07692308 0.00000000
  MAL  0.03448276 0.00000000 0.20689655 0.24137931 0.48275862 0.00
000000 0.03448276
```

… (Puede verse la salida completa en el Anexo 10-B)

```
Prediccion
```

Brinda información de la predicción obtenida al correr el algoritmo Naive Bayes

```
[1] MAL
Levels: BIEN MAL
```

```
plot(Prediccion)
```

Nos muestra gráficamente la predicción obtenida.

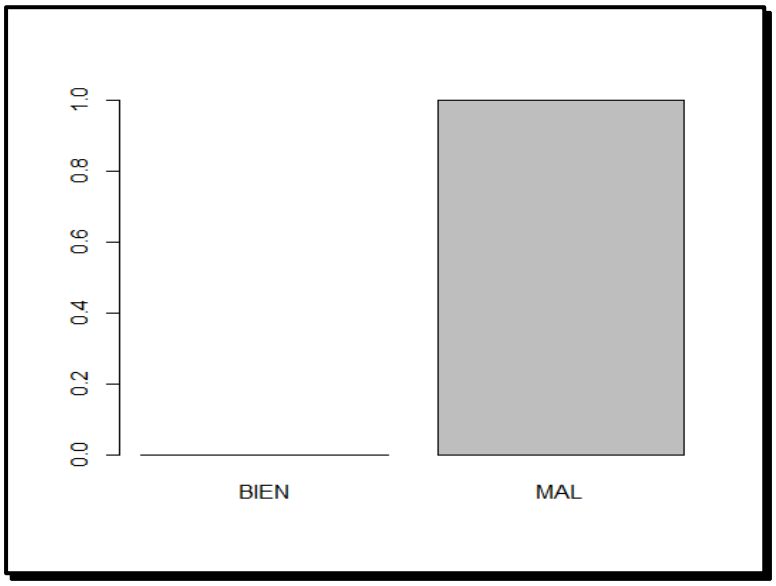


Figura 42.- Gráfico de la predicción con R en el caso 2.

En la Figura 42, se muestran gráficamente los resultados de la predicción como en el Caso 1.

Para el Caso 2 el resultado ha sido similar al Caso 1, ya que los datos históricos de los pacientes drogodependientes en su mayoría tienen como resultado “Mal”. Para ello se propone como trabajo a futuro incluir más datos históricos de pacientes drogodependientes con resultados “Bien” con el propósito de mejorar el test de entrenamiento del algoritmo Naive Bayes. Cabe mencionar que en este caso se ha eliminación los outliers detectados en los tres métodos descritos en el Capítulo 3.

Capítulo 5: Conclusiones y Trabajo a Futuro

Este trabajo está dedicado principalmente a la implementación de algoritmos de limpieza e integración de datos como paso preparatorio para un análisis predictivo. Los procedimientos de limpieza se han basado en la transformación de los campos y datos contenidos en la base de datos original obteniendo un formato coherente para todos los datos. Además, se ha realizado una clasificación de los datos para permitir la ejecución de algoritmos de detección de valores atípicos que puedan conducir a una base de datos más limpia y fiable. Tras la limpieza e integración, los datos que conforman la información tienen calidad suficiente para ser analizados mediante procesos tradicionales.

Para la detección de los valores atípicos se han implementado tres métodos, uno de ellos propio y dos más escogidos de la literatura. Los tres métodos se han comparado y los resultados han dado validez al nuevo método desarrollado que se ha denominado DOCNM (Detección de Outliers mediante Cadenas No Monótonas).

Las pruebas de los algoritmos y procedimientos implementados se han realizado sobre una base de datos real de pacientes drogodependientes. Los datos en estudio no incluyen datos de carácter personal por lo que no se han requerido procedimientos de anonimización ni permisos especiales. El resultado es una herramienta de apoyo para los doctores a cargo de decidir si un paciente entra o no en tratamiento, admitiendo al paciente o derivándolo a otro centro donde pueda tener más probabilidad de éxito su tratamiento.

Como consecuencia de este trabajo han sido aceptadas dos publicaciones: (1) V. Hugo Mariscal, V. López, D. Urgelés, Detección de outliers mediante secuencias no monótonas, VIII Jornadas de Usuarios de R, Albacete, 2016. (2) V. López, D. Urgelés, V. Hugo Mariscal, J. C. Anchiraico, Integración de datos masivos como input de análisis predictivos: Trastornos de bipolaridad y drogodependencia. XXXVI Congreso Nacional de Estadística e Investigación Operativa, Toledo, 2016.

Como trabajo futuro se propone la inclusión de más cantidad de datos procedentes de otros centros o datos históricos que puedan enriquecer el estudio. Los algoritmos implementados pueden ser fácilmente paralelizables en bases de datos distribuidas, lo cual es interesante en la aplicación a los centros asociados en la Orden Hospitalaria San Juan de Dios, que sería el siguiente paso a realizar. Para mejorar la ejecución y el uso de los algoritmos desarrollados también se propone la implementación de una librería R que contenga todas las funcionalidades realizadas, incluyendo la codificación R de algoritmos como KMP y DOCNM que no están disponibles en la actualidad.

Referencias Bibliográficas

[1] Díaz del Mazo L, Vicente Botta B, Arza Lahens M, Moráguez Perelló G, Ferrer González S. "Drogodependencia: un problema de salud contemporáneo" [artículo en línea]. MEDISAN 2008;12(2). [Accessed: 01-Jun-2016].

[2] "Clínica Nuestra Señora de la Paz Madrid." [Online]. Available: http://www.nuestraseñoradelapaz.es/?q=CAIC. [Accessed: 01-Jun-2016].

[3] "OMS | Hay que mejorar el acceso de los drogodependientes a la atención sanitaria", *WHO*. [En línea]. Disponible en: http://www.who.int/mediacentre/news/notes/2012/drug_use_20120626/es/. [Accedido: 25-jul-2016].

[4] "The European Bioinformatics Institute", [Online]. Available: http://www.ebi.ac.uk/. [Accessed: 01-Jun-2016].

[5] "International Society for Computational Biology – ISCB", [Online]. Available: http://www.iscb.org/. [Accessed: 01-Jun-2016].

[6] Cordero, J. (2013). Base de datos en R. Análisis gráfico y estadístico de valores atípicos y ausentes (Tech. Rep. Master in Informatics Research) Eprints Universidad Complutense, Madrid.

[7] Farias, G., Santos, M., López, V. (2010). Making decisions on brain tumour diagnosis by soft computing techniques, Soft Computing, Vol 14 (12), pp. 1287-1296.

[8] Martinez, J. (2013). BioSeq: Una librería para Bioinformática en R (Tech. Rep. Master in Informatics Research) Eprints Universidad Complutense, Madrid.

[9] Sampedro, J. (2012). Aplicaciones de Bioestadística y Bioinformática con R (Tech. Rep. Master Ingeniería Matemática) Eprints Universidad Complutense, Madrid.

[10] Sanchez, O. (2014). Algoritmo de programación dinámica con R para resolver problemas de alineamiento de secuencias (Tech. Rep. Master in Informatics Research) Eprints Universidad Complutense, Madrid.

[11] Valverde, G. (2014). Aplicación de técnicas de Optimización y Big Data al problema de búsqueda de homologías en bases de datos biológicas. (Master in Mathematics, Statistic and Operative Investigation Research) Eprints Universidad Complutense, Madrid.

[12] M. Lenzerini. (2002) "Data integration: A theoretical perspective", en Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 233–246.

[13] Stonebraker, M., Bruckner, D., Ilyas, I. F., Beskales, G., Cherniack, M. Zdonik, S. B., & Xu, S. (2013). Data Curation at Scale: The Data Tamer System. In CIDR.

[14] S. Kandel, A. Paepcke, J. Hellerstein, y J. Heer. (2015) "Wrangler: Interactive visual specification of data transformation scripts", en *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 3363–3372.

[15] Llamocca, P. (2016). “Integración y Visualización de Datos Abiertos Medioambientales” (Tech. Rep. Master in Informatics Research) Eprints Universidad Complutense, Madrid.

[16] C. Leys, C. Ley, O. Klein, P. Bernard, y L. Licata. (2013) “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median”, *Journal of Experimental Social Psychology*, vol. 49, n.º 4, pp. 764-766.

[17] Lopez, V., Valverde, G., & Anchiraico, J. “Specification of a CAD Prediction System for Bipolar Disorder”, *Proceedings of the FLINS conference*. 2016, France.

[18] M. Del Moral, P. L. Fernández, L. Ladero, y L. Lizasoain, (1998) “Conceptos fundamentales en drogodependencias”, *Ladero Lizasoain Drog. Madr. ES Medica Panam.*

[19] E. de Jonge y M. van der Loo, (2013) “An introduction to data cleaning with R”, Stat. Neth. Hague, p. 53.

[20] E. Rahm y H. H. Do, “Data cleaning: Problems and current approaches”, *IEEE Data Eng. Bull.*, vol. 23, n.º 4, pp. 3–13, 2000.

[21] P. Villoslada y S. Baranzini. (2012) “Data integration and systems biology approaches for biomarker discovery: Challenges and opportunities for multiple sclerosis”, *Journal of Neuroimmunology*, vol. 248, n.º 1-2, pp. 58-65.

[22] “Data Wrangler”. [En línea]. Disponible en: http://vis.stanford.edu/wrangler/blog/. [Accedido: 19-ago-2016].

[23] S. Kandel, A. Paepcke, J. Hellerstein, y J. Heer, (2011) “Wrangler: Interactive visual specification of data transformation scripts”, en *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3363–3372.

[24] “Wrangler Enterprise”, Trifacta, [En línea]. Disponible en: https://www.trifacta.com/products/wrangler-enterprise/. [Accedido: 19-ago-2016].

[25] J. Shieber, (2014) “Trifacta Raises $25 Million For Its Data Transformation Software”, *TechCrunch*. [En línea]. Disponible en: http://social.techcrunch.com/2014/05/29/trifacta-raises-25-million-for-its-data-transformation-software/. [Accedido: 15-jun-2016].

[26] “OpenRefine”. [En línea]. Disponible en: http://openrefine.org/community. [Accedido: 10-ago-2016].

[27] Cuadra, J., Romero, I., & Terrado, C. (2008). “CRM Inteligente” (Tech. Rep. Grade in Informatics Research) Eprints Universidad Complutense, Madrid.

[28] D. Meyer, Ed., (2015) “e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien”.

[29] Bello, L. (2014) “Regresión Logística”, *Conferencia Facultad Nacional de Salud Pública – Universidad de Antioquia “Hector Abad Gomez”*.

[30] Herranz, J. (2014) “Modelos Predictivos con el paquete caret”, *Conferencia VI Jornadas de Usuarios de R Santiago de Compostela*.

[31] J. L. C. Reche, (2013) “Regresión logística. Tratamiento computacional con R.”.

[32] A. Alía Martín, (2010) “Estudio e implementación de sensores de fuerza 3D con aplicación a manos robóticas”.

[33] S. Fritsch, F. Guenther, y M. F. Guenther, (2012) “Package ‘neuralnet’”, *Train. Neural Netw.*, vol. 1.

[34] J. Velásquez, C. Zambrano, & L. Vélez, (2011) “Un paquete para la predicción de series de tiempo usando redes neuronales autorregresivas”.

[35] “Constitución Española. - BOE-A-1978-31229-consolidado.pdf.” [Online]. Available: https://www.boe.es/buscar/pdf/1978/BOE-A-1978-31229-consolidado.pdf. [Accessed: 16-Apr-2016].

[36] L. Org, T. I. D. Generales, and T. Ii, PROTECCIÓN DE DATOS DE CARÁCTER PERSONAL. *Regula la Protección de Datos de Carácter*, vol. 1999. 2011.

[37] X. O. source semantic web C. for X.- http://www.ximdex.com, “Agencia Española de Protección de Datos.”. [En línea]. Disponible en: http://www.agpd.es/portalwebAGPD/index-ides-idphp.php. [Accedido: 19-mar-2016].

[38] BOE, “Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica,” *Boletín Of. del Estado*, vol. 274, pp. 40126–40132, 2002.

[39] S. Rajesh, S. Prathima, y L. S. S. Reddy, “Unusual pattern detection in dna database using kmp algorithm”, *International Journal of Computer Applications (0975-8887) Volume*, 2010.

[40] Cormen. T., Leiserson. C., Rivest. Ronald., & Stein. C, “Introduction to algorithms”, *Third Edition, ISBN 978-0-262-03384-8*, 2009.

[41] César Guevara, Matilde Santos & Victoria López, “Negative Selection and Knuth Morris Pratt Algorithm for Anomaly Detection”, Revista IEEE América Latina Volume: 14, Issue: 3, Date: March 2016.

[42] Lukasz Komsta, “outliers: Tests for outliers,” *24-01*, 2011. [Online]. Available:https://cran.r-project.org/web/packages/outliers/index.html. [Accessed: 10-Jun-2016].

[43] J. Fox, “Getting started with the R commander: a basic-statistics graphical user interface to R”, *Journal of statistical software*, vol. 14, n.º 9, pp. 1–42, 2005.

[44] J. Fox, “*Iniciación a R Commander”*, 2008.

[45] “FEIR 40: Modelos de Regresión”. [En línea]. Disponible en: http://www.um.es/ae/FEIR/40/#casos-atipicos-y-residuos. [Accedido: 16-jun-2016].

[46] Rish, Irina (2001). “*An empirical study of the naive Bayes classifier*”. In IJCAI Workshop on Empirical Methods in AI.

[47] D. Meyer, Ed., “Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien”. 2015.

[48] V. López, “Probability and Statistics”, en *Encyclopedia of Sciences and Religions*, A. L. C. Runehov y L. Oviedo, Eds. Springer Netherlands, 2013, pp. 1844-1849.

[49] Bayes, Thomas (1763). «An Essay towards solving a Problem in the Doctrine of Chances.». Philosophical Transactions of the Royal Society of London 53: 370-418. doi:10.1098/rstl.1763.0053.

Anexo 1: Algoritmo DOCNM en R

```
texto<- scan ("DATAKMP/SECUEN.txt", what = 'character')
patron<- scan ("DATAKMP/PATRONSECUEN.txt", what =
'character')

PreKMP<-function(patron) {
  m<-nchar(patron)
  prefix <- numeric (0)
  prefix [1] <-0
  a<-0
  for (b in 2:m) {
    while (a>0 && substr (patron, a+1, a+1)! =
substr(patron,b,b))
    {
      a<-prefix[a]
    }
    if (substr (patron, a+1, a+1) == substr(patron,b,b))
    {
      a<-a+1
    }
    prefix[b]<-a
  }
  return (prefix)
}


KMP<-function(texto,patron){
  n<-nchar(texto)
  m<-nchar(patron)
  PRE<-c(PreKMP(patron))
  q<-0
  for (i in 1: n) {
    while (q>0 && substr (patron, q+1, q+1)! =
substr(texto,i,i))
    {
      q<-PRE[q]
    }
    if (substr (patron, q+1, q+1) == substr(texto,i,i))
    {
      q<-q+1
    }
    if(q==m) {
      cat ("\n Patrón encontrado en la posición: ", (i-m +1))
      q<-PRE[q]
    }
  }
}

PreKMP(patron)
KMP(texto,patron)
```

Anexo 2: Componentes del Algoritmo KMP

La Función Prefijo II.- La función de prefijo, Π de un patrón encapsula el conocimiento acerca de cómo coincide con el patrón en contra de los cambios de sí mismo. Esta información se puede utilizar para evitar cambios inútiles del patrón de "P". En otras palabras, esto permite evitar dar marcha atrás en la cadena "S".

Siguiendo el pseudocódigo de la función de prefijo, (Prefix fucnction):

Imput: Patrón "**P**" de longitud '**m**'
Output: Tabla **Prefix** [**1…, m**]

PREFIX – Fucnction(P)
m <- length[**P**] // Longitud del Patrón
Prefix [1] <- 0
k <- 0
for **q** <- **2** to **m** // Escanear el patrón de **P** [**2, …, m**] de izquierda a derecha
while **k** > 0 and **P** [**k + 1**]! = **P**[**q**]
k <- Prefix **[k + 1]** // El patrón de **Pk** no es el sufijo de **Pq**
if **P** [**k + 1**] = **P**[**q**]
k <- **k+1** // El más largo **Pk** prefijo es también un adecuado sufijo de **Pq**
Prefix[**q**] <- **k**
return **Prefix**

Ejemplo del cálculo del Prefix:

Cálculo de **Prefix** para el siguiente patrón de '**P**'

P	**a**	**b**	**a**	**b**	**a**	**c**	**a**

Inicialmente**: m** = length[**P**] = 7, **Prefix [1]= 0, k** = **0**

PASO 1: q = 2
k = 0
Prefix [2]= 0

q	**1**	**2**	**3**	**4**	**5**	**6**	**7**
P	**a**	**b**	**a**	**b**	**a**	**c**	**a**
Prefix	**0**	**0**					

PASO 2: q = 3
k = 0
Prefix [3]= 1

q	**1**	**2**	**3**	**4**	**5**	**6**	**7**
P	**a**	**b**	**a**	**b**	**a**	**c**	**a**
Prefix	**0**	**0**	**1**				

PASO 3: q = 4
k = 1
Prefix [4]= 2

q	**1**	**2**	**3**	**4**	**5**	**6**	**7**
P	**a**	**b**	**a**	**b**	**a**	**c**	**a**
Prefix	**0**	**0**	**1**	**2**			

PASO 4: q = 5
k = 2
Prefix [5]= 3

q	**1**	**2**	**3**	**4**	**5**	**6**	**7**
P	**a**	**b**	**a**	**b**	**a**	**c**	**a**
Prefix	**0**	**0**	**1**	**2**	**3**		

PASO 5: q = 6
k = 0
Prefix [6]= 1

q	1	2	3	4	5	6	7
P	a	b	a	b	a	c	a
Prefix	0	0	1	2	3	1	

PASO 6: q = 7
k = 1
Prefix [7]= 1

q	1	2	3	4	5	6	7
P	a	b	a	b	a	c	a
Prefix	0	0	1	2	3	1	1

Después de 6 veces de iteración, el cálculo de la función prefijo está completa.

q	1	2	3	4	5	6	7
P	a	b	a	b	a	c	a
Prefix	0	0	1	2	3	1	1

La Función Matcher KMP. - El Matcher KMP, con el patrón de **'P'**, cadena **'T'** y la función de prefijo '**PREFIX**' como entrada, encuentra una coincidencia de **'P'** en **'T'**. El pseudocódigo siguiente calcula el componente coincidente del algoritmo KMP:

Input: Patrón "**P**" de longitud "**m**" y texto "**T"** de longitud "**n**"
Output: Lista de todos los números "**S**", tal que "**P"** ocurre según el desplazamiento "**S"** en "**T".**

KMP – Matcher - Function(T, P)
n <- length[**T**] // Longitud del Texto
m <- length[**P**] // Longitud del Patrón

KMP<- PREFIX - Function (P)
q <- 0 // Numero de caracteres iguales
for **i** <- **1** to **n** // Escanear de T de izquierda a derecha
while **q** > 0 and **P** [**i + 1**]! = **T**[**i**]
q <- **KMP** [**q + 1**] // Siguiente carácter no coincide
if **P** [**q + 1**] = **P**[**i**]
q <- **q+1** // Siguiente carácter coincide
if **q=m**
print **("Patrón encontrado en la posición:", (i-m +1))**
q <- **KMP[q]//** Busca el próximo emparejado

Nota: KMP encuentra cada aparición de una **'P'** en **'T'**. Por eso KMP no termina en el paso **12**, sino que busca resto de **'T'** de las más ocurrencias de **'P'**.

Ejemplo del cálculo de Matcher KMP:

Dada una cadena **'T'** y el patrón de **"P"** de la siguiente manera:

T	b	a	c	b	a	b	a	b	a	b	a	c	a	c	a

P	a	b	a	b	a	c	a

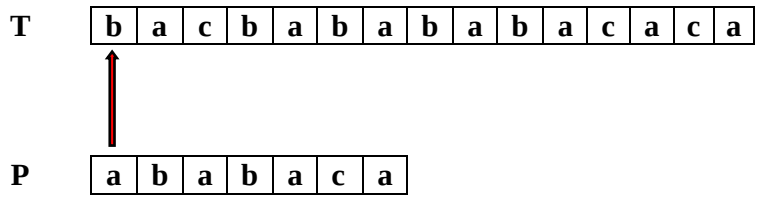
Vamos a ejecutar el algoritmo KMP para saber si “P" se produce en ‘T'.

Para ‘P' la función de prefijo, PREFIX se calculó previamente y es como sigue:

q	**1**	**2**	**3**	**4**	**5**	**6**	**7**
P	**a**	**b**	**a**	**b**	**a**	**c**	**a**
Prefix	**0**	**0**	**1**	**2**	**3**	**1**	**1**

Inicializamos: **n** = Tamaño de **T** = 15;
m = Tamaño de **P** = 7

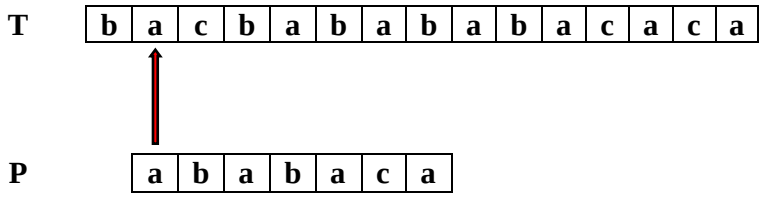
PASO 1: **i** = 1, **q** = 0
comparando **P** [**1**] con **T** [**1**]



P [1] no coincide con **T [1].** **'P'** se desplazará una posición hacia la derecha**.**

PASO 2: **i** = 2, **q** = 0

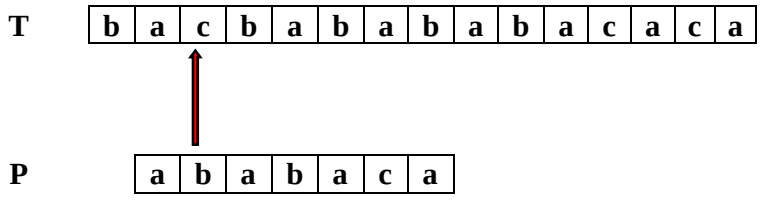
comparando **P** [**1**] con **T** [**2**]



P [1] coincide con **T [2].** Dado que hay una coincidencia**, ‘P'** no se desplazará**.**

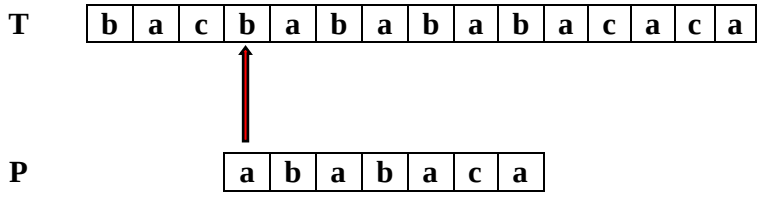
PASO 3: **i** = 3, **q** = 1

comparando **P** [**2**] con **T** [**3**], **P** [**2**] no coincide con **T** [**3**]



Dar marcha atrás en **P,** comparando **P [1] y T [3].**

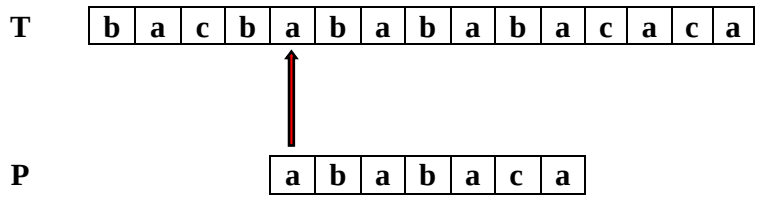
PASO 4: **i** = 4, **q** = 0
comparando **P** [**1**] con **T** [**4**], **P** [**1**] no coincide con **T** [**4**]



P [1] no coincide con **T [4].**, **'P'** se desplazará una posición hacia la derecha**.**

PASO 5: i = 5, **q** = 0

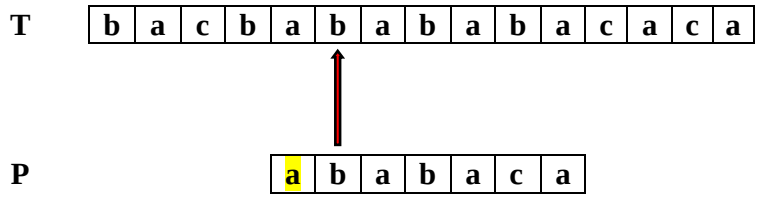
comparando **P [1]** con **T [5], P [1]** coincide con **T [5]**



P [1] coincide con **T [5].** Dado que hay una coincidencia**, 'P'** no se desplazará**.**

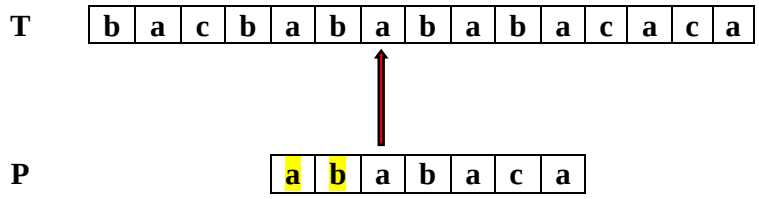
PASO 6: i = 6, **q** = 1

comparando **P** [**2**] con **T** [**6**], **P** [**2**] coincide con **T** [**6**]



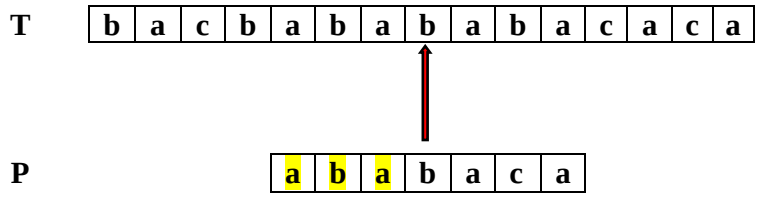
PASO 7: i = 7, **q** = 2

comparando **P** [**3**] con **T** [**7**], **P** [**3**] coincide con **T** [**7**]



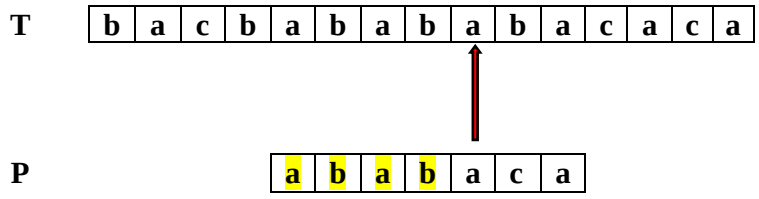
PASO 8: **i** = 8, **q** = 3

comparando **P** [**4**] con **T** [**8**], **P** [**4**] coincide con **T** [**8**]



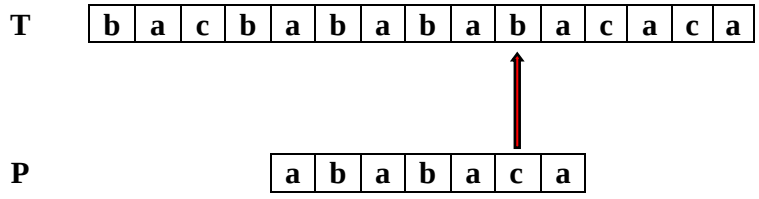
PASO 9: **i** = 9, **q** = 4

comparando **P** [**5**] con **T** [**9**], **P** [**5**] coincide con **T** [**9**]



PASO 10: **i** = 10, **q** = 5

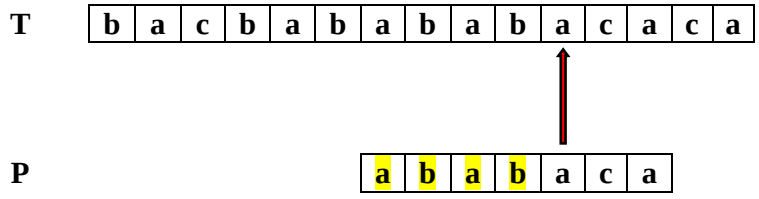
comparando **P** [**6**] con **T** [**10**], **P** [**6**] no coincide con **T** [**10**]



P [6] no coincide con **T [10], retrocediendo en 'P',** comparando **P [4]** con **T [10]** porque después de desajuste **q= PREFIX [5]=3.**

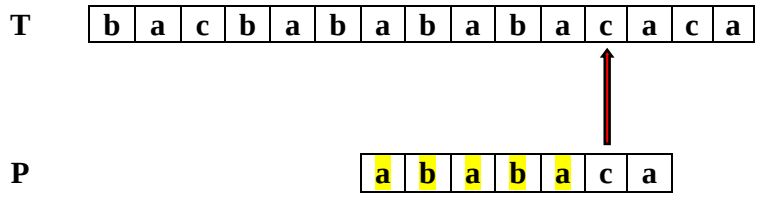
PASO 11: **i** = 11, **q** = 4

comparando **P** [**5**] con **T** [**11**], **P** [**5**] coincide con **T** [**11**]



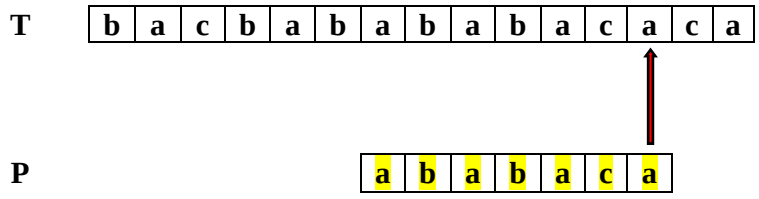
PASO 12: **i** = 12, **q** = 5

comparando **P** [**6**] con **T** [**12**], **P** [**6**] coincide con **T** [**12**]



PASO 13: **i** = 13, **q** = 6

comparando **P** [**7**] con **T** [**13**], **P** [**7**] coincide con **T** [**13**]



El patrón **'P'** ocurre completamente en la cadena **'T'**. El número total de movimientos que tuvo lugar durante el recorrido son: **i - m** = 13 - 7 = 6 movimientos.

Anexo 3: Detección de Outliers con Regresión Lineal Simple

1. Analizamos la normalidad de los datos

Damos inicio con la lectura de los datos del archivo caicv0.csv, de la siguiente manera:

```
drogodependencia<read.table("C:/Users/victor/DatoPreprocesado.csv",
sep = ",", head = TRUE)
drogodependencia
```

Empezamos estudiando la normalidad de la variable expectativa con el test de normalidad (test de shapiro)

```
shapiro.test(drogodependencia$edad)

Shapiro-Wilk normality test

data:  drogodependencia$edad
W = 0.938, p-value = 0.008459

shapiro.test(drogodependencia$i_coca)

Shapiro-Wilk normality test

data:  drogodependencia$i_coca
W = 0.97634, p-value = 0.3717
```

2. Calculamos la correlación entre edad y i_coca (cor.test)

Visto que los datos son normales, realizamos el análisis de correlación

```
cor.test(drogodependencia$edad, drogodependencia$i_coca)

Pearson's product-moment correlation
data: drogodependencia $edad and drogodependencia$i_coca
t = 4.2162, df = 51, p-value = 0.0001017
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.2760330 0.6846124
sample estimates:
cor    0.5083954
```

3. Gráfico de Dispersión

y representamos los puntos (Figura 43)

```
plot(drogodependencia$edad, drogodependencia$i_coca, pch = 20, xlab
= "Edad", ylab = "i_coca", main = "Diagrama de dispersión", cex.main
= 0.95)
```

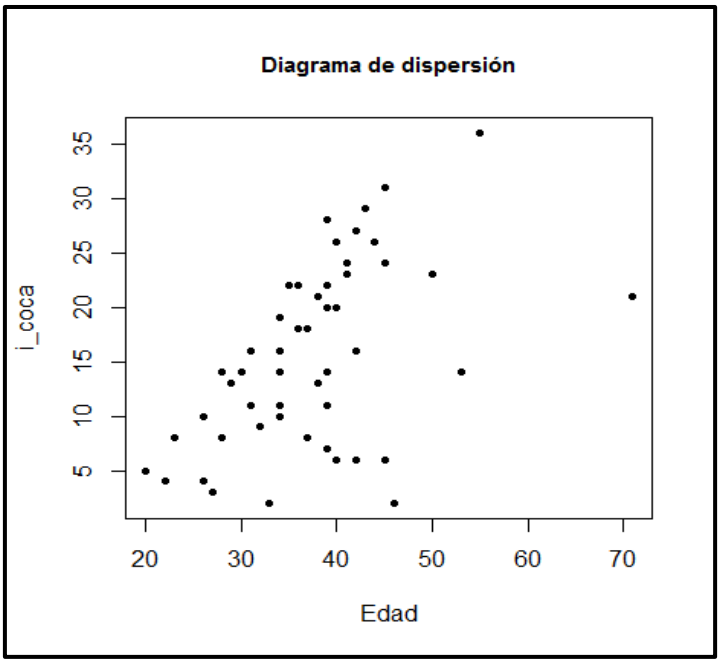


Figura 43. Diagrama de dispersión

La correlación entre ambas variables es significativa con un p-valor menor a 0.05 y se trata de una relación inversa y alta (-0.7851), según crece la edad disminuyen las i_coca.

4. Ajuste de modelo

Una vez visto que existe relación entre las variables pasamos a realizar el ajuste del modelo. Para ello usamos la función lm () que toma la forma

lm (dependiente ~ predictora(s), data = dataFrame, na.action = "acción")

donde **na.action** es opcional, puede ser útil si tenemos valores perdidos.

5. Creación de Modelo modelDROGO

Creamos el objeto modelDROGO que contiene todos los resultados del ajuste.

```
modelDROGO <- lm (i_coca ~ edad, data = drogodependencia)
summary(modelDROGO)

Call:
lm(formula = i_coca ~ edad, data = drogodependencia)
Residuals:
Min      1Q  Median      3Q     Max
-17.756  -4.072   1.335   5.064  11.948
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.2057     4.3315  -0.509 0.612795
edad          0.4774     0.1132   4.216 0.000102 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.122 on 51 degrees of freedom
```

Multiple R-squared: 0.2585, Adjusted R-squared: 0.2439
F-statistic: 17.78 on 1 and 51 DF, p-value: 0.0001017

La parte 'Residuals' nos da la diferencia entre los valores experimentales y ajustados por el modelo. Las estimaciones de los coeficientes del modelo se proporcionan junto con las sus desviaciones estándar ('error estándar'), un t-valor y la probabilidad de la hipótesis nula de que los coeficientes tengan valor de cero. En este caso, por ejemplo, hay evidencia de que ambos coeficientes son significativamente diferentes de cero.

En la parte inferior de la tabla de la Figura 44, se encuentra la desviación sobre la recta regresión (error estándar srsr o residual), el coeficiente de correlación y el resultado del test F sobre la hipótesis nula de que los MSreg/MSres es 1.

```
plot(drogo$edad, drogo$i_coca, pch = 20, xlab = "Edad", ylab = "i_coca")

abline(modelDROGO)
```

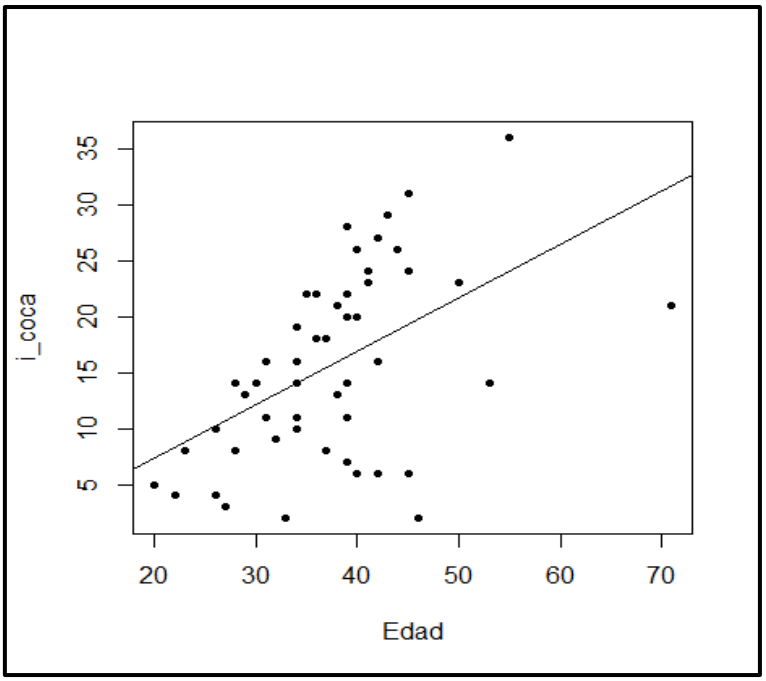


Figura 44. Diagrama de la desviación de la recta regresión

6. Estimuladores puntuales, errores estándar y p-valores asociados

En primer lugar, deseamos obtener los estimadores puntuales, errores estándar y p-valores asociados con cada coeficiente

```
summary(modelDROGO)$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.2056857	4.3315108	-0.5092186	0.6127953079
edad	0.4774184	0.1132343	4.2161992	0.0001016568

El resultado del ajuste es
(4.3315) (0.1132) i_coca = -2.2056 0.4774 * Edad

donde los valores entre paréntesis indican los errores estándar de cada coeficiente. Además, puesto que los p-valores asociados son inferiores a 0.05, podemos concluir que:

Existen evidencias estadísticas suficientes para considerar que hay una relación lineal entre Edad y i_coca. Dicha relación es negativa cuando aumenta la Edad del paciente aumentan las i_coca. Además, vemos que por cada grado que aumenta la Edad del paciente, aumentan las i_coca en 0.50.

El error estándar residual estimado (s) es de 5.898. Este valor es muy importante, es un medidor de la calidad (precisión) del modelo. Además, nos vamos a basar en él para calcular los intervalos de confianza para los coeficientes del modelo. Se calcula haciendo la raíz cuadrada de la media de la suma de cuadrados de los residuos (MSR).

7. Intervalos de confianza

Los intervalos de confianza (IC) complementan la información que proporcionan los contrastes de hipótesis a la hora de expresar el grado de incertidumbre en nuestras estimaciones.

Obtenemos los correspondientes intervalos de confianza para cada parámetro del modelo con nivel significación al 95%.

confint(modelDROGO, level = 0.95)

```
                2.5 %     97.5 %
(Intercept) -6.1081989 12.2478090
edad         0.2665499  0.7464125
```

como el intervalo contiene al cero, podemos afirmar la hipótesis nula de que H0: β0=β1=0H0: β0=β1=0.

Interpretamos los intervalos: con una probabilidad del 95%, la ordenada en el origen del modelo, β0, se encuentra en el intervalo (-6.10, 12.24), mientras que el efecto asociado con la Edad se encuentra en el intervalo (0.26, 0.74).

8. Valores atípicos

Un valor atípico es aquel que difiere sustancialmente de la tendencia general de los datos. Estos valores atípicos pueden perjudicar el modelo ya que afectan a los coeficientes de regresión estimados.

En el caso de observar valores atípicos los pasos a seguir son:

- ✓ Descartar que sea un error.
- ✓ Analizar si es un caso influyente.
- ✓ En caso de ser influyente calcular las rectas de regresión incluyéndolo y excluyéndolo, y elegir la que mejor se adapte al problema y a las observaciones futuras.

Para el estudio de los valores atípico vamos a usar los residuos estandarizados, los residuos divididos por una estimación de su error estándar. Existen unas reglas generales:

I. Residuos estandarizados con un valor absoluto mayor de **3.29** (redondearemos a **3**) son causa de preocupación ya que es improbable que en una muestra media un valor tan grande ocurra por azar.

II. Si más del **1%** de los valores muéstrales tienen residuos estandarizados con un valor absoluto mayor de **2.58** (podemos decir **2.5**) hay evidencias de que el nivel de error en nuestro modelo es inaceptable (ajuste pobre del modelo a los datos).

III. Si más del **5%** de los casos tienen residuos estandarizados con un valor absoluto mayor de **1.96** (usamos **2** por conveniencia) entonces vuelven a haber indicios de que el modelo es una pobre representación de los datos reales.

9. Análisis con outliersTest de la librería (car)

Continuamos con un análisis más analítico:

```
library(car)
outlierTest(modelDROGO, cutoff = 0.05, n.max = 10, order = TRUE)
```

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
49 -2.693245         0.009605      0.50907
```

```
influencePlot(modelDROGO, id.n = 2)
```

```
        StudRes       Hat     CookD
47 -1.948098 0.03399401 0.2516060
49 -2.693245 0.03815742 0.3579993
52  1.805346 0.09837660 0.4126353
53 -1.844503 0.30653746 0.8474229
```

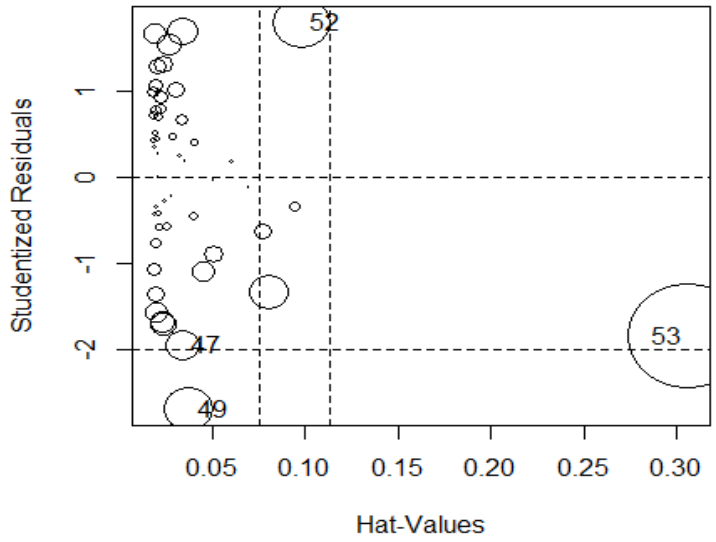


Figura 45. Diagrama de outliersTest

El test y en la Figura 45 nos indican que la observación número 53 es un valor atípico. Las observaciones 47, 49, 52 y 53 que vemos en la Figura 46 son medidas influyentes para ver si llegan a ser atípicos dibujamos el gráfico de las distancias de Cook (J.Faraway, 2009).

```
cook <- cooks.distance(modelDROGO)
labels <- rownames(drogodependencia)
library(faraway)
halfnorm(cook, 3, labs = labels, ylab = "Distancia de Cook")
```

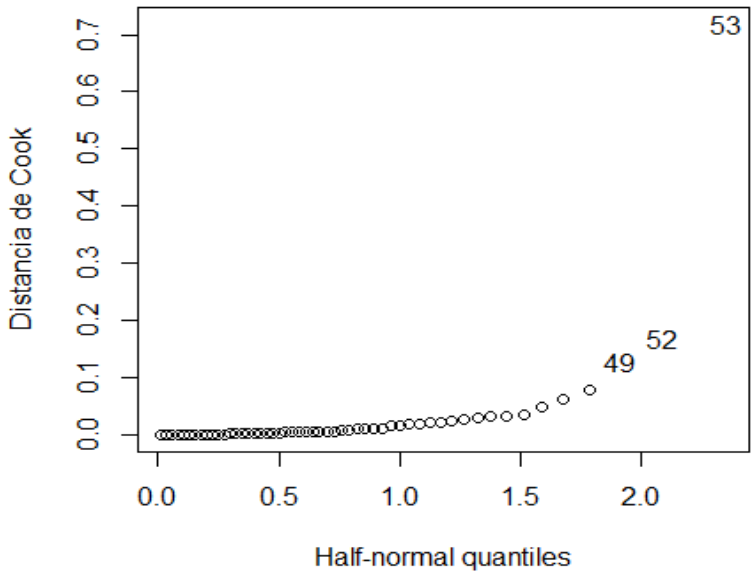


Figura 46. Gráfico de la distancia de Cook

Anexo 4: Detección de Outliers con RCommander

Procedemos con el modelo lineal, ya que su sencillez favorece la interpretación de los coeficientes (Figura 47).

Estadísticos ➥Ajuste de modelos➥Modelo lineal

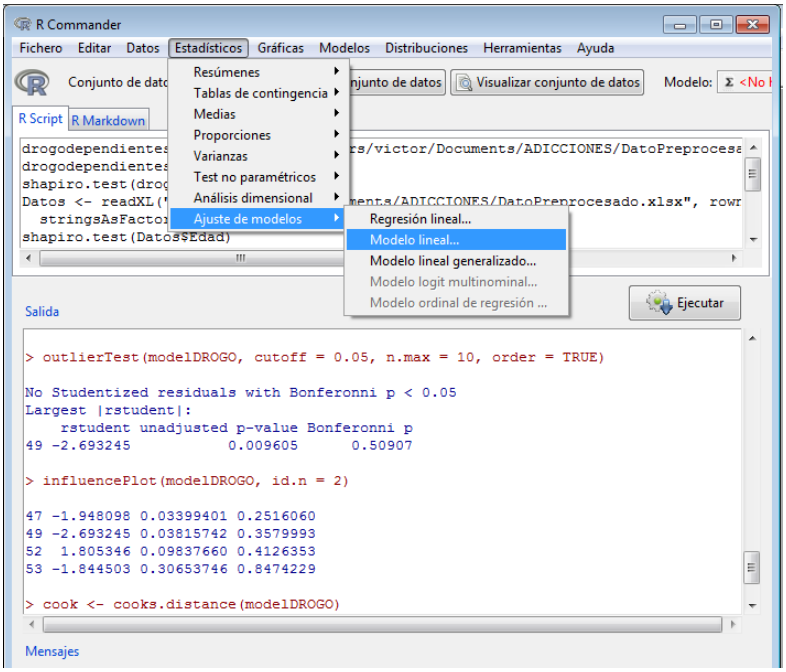


Figura 47. Ajuste del modelo con RCommander

Nombre del modelo: **Modelo4** (Figura 48).

➥Formula del…**edad~i_coca**

➥Aceptar

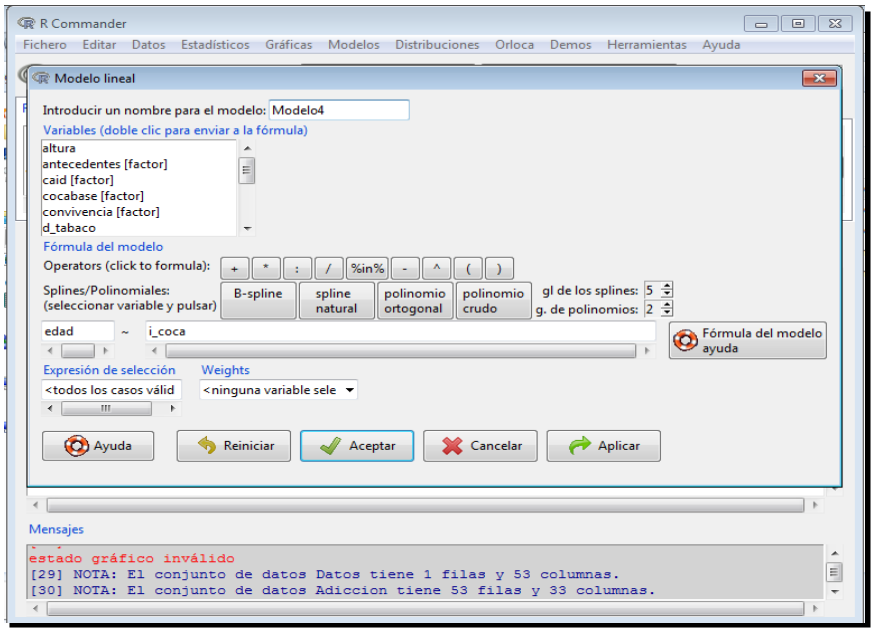


Figura 48. Modelo Lineal con RCommander

> **Modelo1<- lm(edad ~ i_coca, data=Drogodependencia)**
> summary(Modelo1)

```
Residuals:
    Min      1Q  Median      3Q     Max
-11.534  -4.986  -1.572   3.091  30.804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8268     2.2561  12.777  < 2e-16 ***
i_coca1       0.5414     0.1284   4.216 0.000102 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.584 on 51 degrees of freedom
Multiple R-squared:  0.2585,   Adjusted R-squared:  0.2439
F-statistic: 17.78 on 1 and 51 DF,  p-value: 0.0001017
```

El test de valores atípicos de Bonferroni indica la presencia de observaciones atípicas (Figura 49).

Modelos
➥Diagnósticos numéricos
➥Test de valores atípicos de Bonferroni…

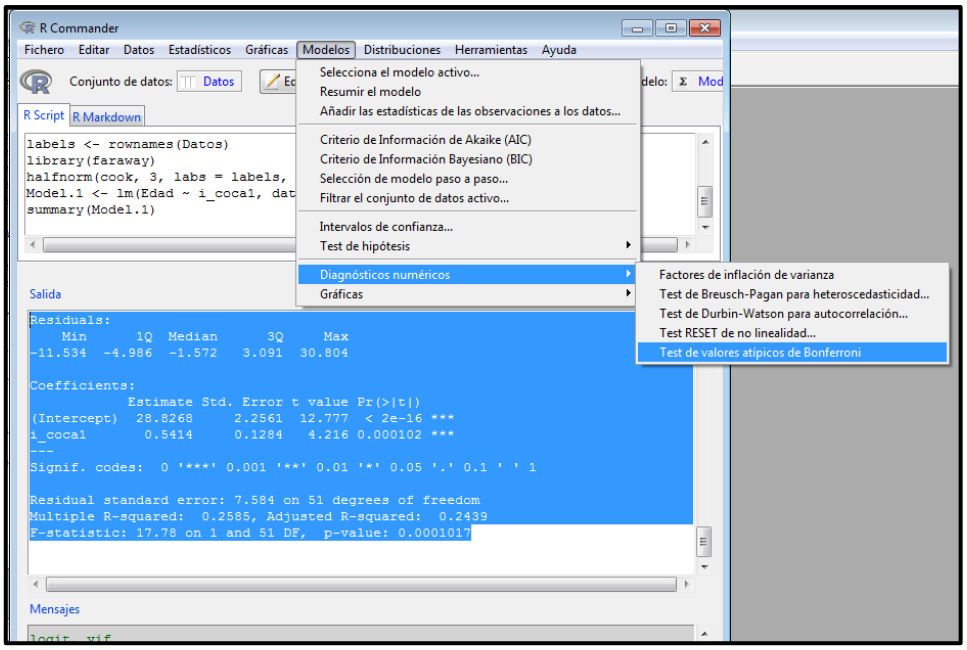


Figura 49. Modelos con RCommander

> **outlierTest(Modelo1)**

```
rstudent unadjusted p-value Bonferonni p
53 4.990774         7.6738e-06   0.00040671
```

Observation: 53

El p-valor es menor que α e implica que hay observaciones atípicas: el número 5.

Este dibujo se consigue transformando la escala de los ejes:

Gráficas (Figura 50)

➥Matriz de diagrama de dispersión

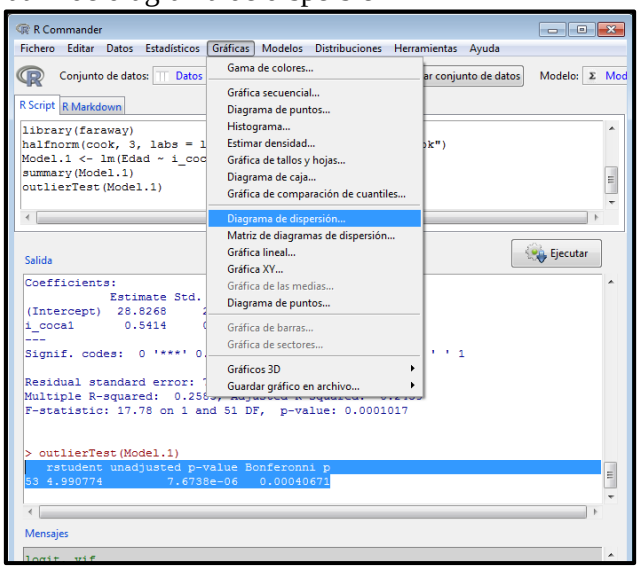


Figura 50.- Diagrama de dispersión con RCommander

Seleccionamos i**_coca** y **edad** (Figura 51)

➥Marcamos **Log eje-x**

➥Aceptar

> scatterplot(edad~i_coca, reg.line=lm, smooth=TRUE, spread=FALSE, id.method='mahal', + id.n = 3, boxplots='xy', span=0.5, ellipse=FALSE, levels=c(.5, .9), data= Drogodependencia)

15 49 53

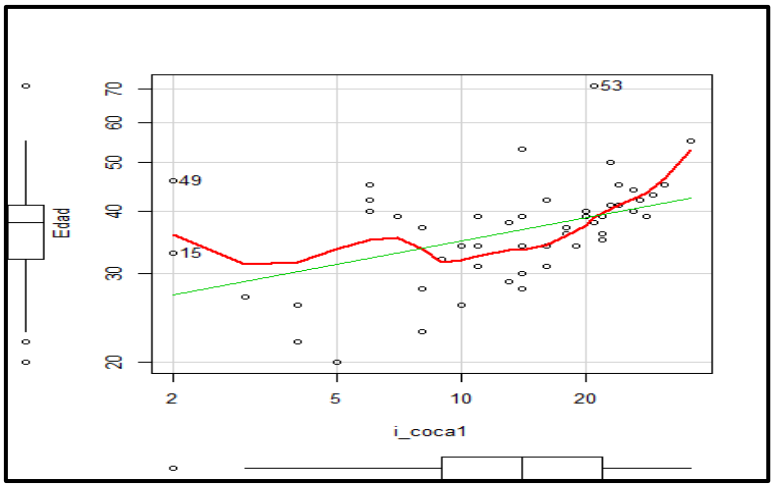


Figura 51. Scatterplot de *edad* y *i_coca*, con RCommander

Anexo 5: Generalidades y Conceptos de Probabilidad

A continuación, en esta sección se explican las generalidades y los conceptos de probabilidad [48] que se necesitan para la comprensión del clasificador bayesiano.

Definición 5.1 Se define la probabilidad de que ocurra un suceso A en espacio muestral E, como una medida en el rango [0..1] que se calcula siguiendo la fórmula de Laplace, asumiendo que los sucesos elementales son equiprobables.

$$\boldsymbol{P(A)} = \frac{\boldsymbol{\#\ casos\ favorables}}{\boldsymbol{\#casos\ posibles}} = \frac{|A|}{|E|}$$

Esta fórmula es consistente con el rango, ya que 0 corresponde a la probabilidad del suceso imposible y 1 corresponde a la probabilidad del suceso total (cuando los casos favorables coinciden con los casos posibles). En este trabajo utilizaremos esta fórmula para estimar las probabilidades sobre las muestras de datos contenidos en nuestro dataset de drogodependientes.

Algunos ejemplos de aplicación de la fórmula de Laplace son las siguientes:

- E= {cara, sello}
  Suceso A= "Obtener cara al lanzar una moneda"
  Suceso A = {cara}
  P(A)=1/2=0.5
- E= {(1,1), (2,2), …, (6,6)}
  Suceso B= "Obtener un número mayor que 2 al lanzar un dado"
  Suceso B = {3,4,5,6}
  P(B)=4/6=0.67
- E= {1,2,3,4,5,6}
  Suceso C= "Sumar 2 al lanzar 2 dados"
  Suceso C= {(1,3), (3,1), (2,2)}
  P(C)=1/36=0.0278

La fórmula de Laplace también es útil para calcular la intersección de dos sucesos. Si A y B son dos sucesos, la probabilidad del suceso A ∩ B será:

$$\boldsymbol{P(A \cap B)} = \frac{\boldsymbol{\#\ casos\ favorables\ A \cap B}}{\boldsymbol{\#casos\ posibles}} = \frac{\boldsymbol{|A \cap B|}}{\boldsymbol{|E|}}$$

Definición 5.2 La probabilidad de que ocurra un suceso A condicionado a que ya ha ocurrido un suceso B se denota como P (A/B) y se define mediante la fórmula:

$$\boldsymbol{P(A\ /\ B)} = \frac{\boldsymbol{P(A \cap B)}}{\boldsymbol{P(B)}}$$

Por ejemplo:

- E= {1, 2, 3, 4, 5, 6}
  Suceso A= "Obtener un número distinto de 6 al lanzar un dado"

Suceso A = {1, 2, 3, 4, 5}
Suceso B= “Obtener un número mayor que 2 al lanzar un dado”
Suceso B = {3,4,5,6}
P(A ∩ B) =3/6 =0.5
P(B) =4/6 = 0.67

A partir de la probabilidad condicionada, despejando en la fórmula, se obtiene la fórmula para la probabilidad de la intersección de dos sucesos:

$$\boldsymbol{P(A \cap B) = P(A/B)\, P(B)}$$

Observación: Una partición de E es un conjunto completo de sucesos mutuamente excluyentes.

Teorema 5.1 (Teorema de la Probabilidad Total) Sea E un espacio muestral y sea $\{\boldsymbol{A_1},\ \boldsymbol{A_2}, \dots, \boldsymbol{A_K}\}$ una partición de dicho espacio. Sea B un suceso sobre E. Entonces,

$$\boldsymbol{P(B) = P(B/A_1).P(A_1) + \cdots + P(B/A_k).P(A_k)}$$

$$\boldsymbol{P(B) = \sum_{i=1}^{k} P(B/A_i).P(A_i)}$$

Teorema 5.2 (Teorema de Bayes) [49**]** En las condiciones del teorema anterior,

$$\boldsymbol{P(A_j/B) = \frac{P(B/A_j).P(A_j)}{\sum_{i=1}^{k} P(B/A_i).(A_i)}, \qquad siendo\ P(B) \neq 0}$$

Donde el denominador es la probabilidad del suceso B, P(B), según el teorema anterior y que debe ser distinta de cero para no anular el denominador y producir incoherencia en la fórmula.

Para ilustrar la aplicación de estos teoremas damos un ejemplo a continuación. Supongamos que una fábrica de enlatados produce 5000 envases diarios. La máquina A produce 3000 de estos envases, de los que se sabe que el 2% son defectuosos y la máquina B produce los 2000 restantes de los que se sabe que el 4% son defectuosos. Nos puede interesar determinar: a. La probabilidad de que un envase elegido al azar sea defectuoso; y b. Supuesto que se ha encontrado un envase defectuoso nos podemos preguntar qué probabilidad hay de que proceda de la maquina A, o también de que proceda de la maquina B.

Las soluciones se muestran a continuación.

a) La probabilidad de que un envase elegido al azar sea defectuoso.

Llamamos D al suceso seleccionar un envase defectuoso y $\overline{D}$ al suceso seleccionar un envase correcto. Los posibles casos se ven en el diagrama de la Figura 52.

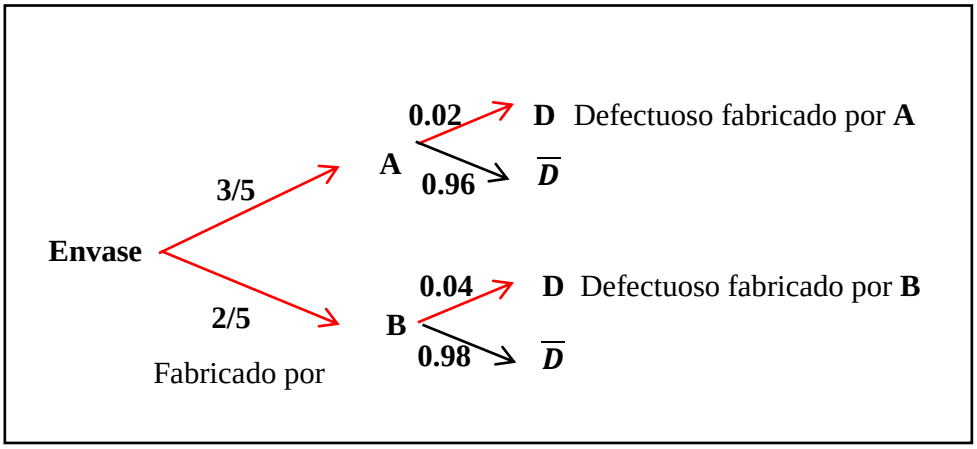


Figura 52.- Diagrama de suceso seleccionar un envase defectuoso.

Se trata de calcular la probabilidad total de que el envase elegido sea defectuoso, que de acuerdo al teorema de la probabilidad total resulta ser:

$$\mathbf{P(D)= P(A).\ P(D/A) + P(B).\ P(D/B) = 0{,}012 + 0{,}016=0{,}028}$$

b) Aplicando el teorema de Bayes la probabilidad de que provenga de la máquina A

$$\boldsymbol{P\left(\frac{A}{D}\right)} = \frac{0{,}012}{0{,}028} = 0{,}4286$$

Y la probabilidad de que provenga de la máquina B

$$\boldsymbol{P\left(\frac{B}{D}\right)} = \frac{0{,}016}{0{,}028} = 0{,}5714$$

Clasificador Naïve Bayes

A partir de esta sección consideramos la existencia de un dataset con k clases diferentes (C1, …, Ck) y n atributos (a1, …, an). Dada una instancia con valores (v1, …, vn), el clasificador bayesiano simple asigna la clase Cα con mayor probabilidad condicional:

$$\boldsymbol{P(C_\alpha |\ a_1 = v_1, \ldots, a_n = v_n)}$$

Esta probabilidad de que la clase sea Cα sabiendo que los atributos han tomado los valores (v1, …, vn) se podría calcular mediante la probabilidad condicionada siguiente:

$$\boldsymbol{P(c_\alpha\ |\overline{a_j = v_j}) = \frac{P(C_\alpha \cap \overline{a_j = v_j})}{P(\overline{a_j = v_j})}}$$

$$\boldsymbol{= \frac{\#instancias\ t.q.\ C_\alpha\ y\ \overline{a_j = v_j}}{\#instancias\ t.q.\ \overline{a_j = v_j}}}$$

Donde se usa la notación

$$\overline{a_j = v_j} = (a_1 = v_1, \ldots, a_n = v_n)$$

Por lo que únicamente tendríamos que contar aquellas instancias en nuestro conjunto de entrenamiento (dataset inicial) que coincidan con la instancia a clasificar (nuevo registro) y asignarle la clase más probable.

Sin embargo, el cálculo directo tiene dos problemas:

- Hay que encontrar aquellas instancias iguales en nuestro conjunto de datos históricos, que puede ser muy grande.
- Quizá no exista ninguna instancia igual en nuestro conjunto de datos históricos, en ese caso no sabremos qué clase elegir como la más probable.

La alternativa propuesta por el clasificador Naive Bayes es utilizar el Teorema de Bayes, que nos permite encontrar la clase Cα con mayor probabilidad:

$$P(c_\alpha \,|\overline{a_j = v_j}) = \frac{P\left(\overline{a_j = v_j}\middle|\, c_\alpha\right) P(c_\alpha)}{P\left(\overline{a_j = v_j}\right)}$$

Como $P(\overline{a_j = v_j})$ valdrá siempre lo mismo, lo que querremos maximizar es $P\left(\overline{a_j = v_j}\middle| c_\alpha\right)P(c_\alpha)$, sin embargo, seguimos necesitando contar el número de instancias en el conjunto de datos con los mismos atributos que la que queremos clasificar. Aquí es donde se hace una suposición simplista (de ahí el nombre del método): los valores de los atributos son independientes unos de otros dentro de cada clase. El concepto de independencia que debemos asumir se define a continuación.

Definición 5.3 Se dice que dos sucesos son independientes si solo sí:

$$P(A \cap B) = P(A).P(B)$$

Teniendo en cuenta la definición de probabilidad condicionada esta definición es equivalente a

$$P\left(\frac{A}{B}\right) = P(A) \Leftrightarrow A, B \; son \; independientes$$

Definición 5.4 (Probabilidades de la intersección de sucesos generales). Sean A1, …, An sucesos cuales quiera sobre el mismo espacio E, se define la probabilidad de la intersección (y se conoce como la regla de la cadena).

$$\boldsymbol{P(A_1 \cap \ldots \cap A_n) = P(A_1).P(A_2/A_1).P(A_3/A_1 \cap A_2).\ldots.P\left(A_n/\bigcap_{i=1}^{n-1} A_i\right)}$$

Como consecuencia de la definición 4.2.1 se obtiene la siguiente propiedad: La probabilidad de la intersección de sucesos independientes es igual al producto de las probabilidades, es decir:

$$P(A_1 \cap A_2 \cap .. \cap A_n) = \prod_{i=1}^{n} P(A_i)$$

Como aplicación a nuestro *dataset* se obtiene lo siguiente:

$$P(C/F_1, \ldots, F_n) = \frac{P(C)\, P(A_1, \ldots, F_n / C)}{P(F_1, \ldots, F_n)}$$

Donde por simplificar, se ha denotado

$$C = C_\alpha \qquad F_i = (a_i = v_i)$$

En la práctica sólo importa el numerador, ya que el denominador no depende de C y los valores de Fi son datos, por lo que el denominador es constante.

El numerador es equivalente a una probabilidad compuesta, que puede ser reescrita como sigue, aplicando repetidamente la definición de probabilidad condicional:

$$P(C, F_1, \ldots, F_n) = P(C)\, P(F_1, \ldots, F_n / C)$$

$$= P(C)\, P(F_1 / C)\, P(F_2, \ldots, F_n / C)$$

$$= P(C)\, P(F_1 / C)\, P(F_2/C, F_1)\, P(F_3, \ldots, F_n/C, F_1, F_2)$$

Se asume que cada Fi es independiente de cualquier otra Fj para j ≠ i. Por lo que la probabilidad compuesta puede expresarse como:

$$P(C_i) \prod_{j=1}^{n} P(F_j/C_i) = P(C)\ P(F_1/C)\, P(F_2/C)\, P(F_3/C) \ldots$$

Todos los parámetros del modelo (por ejemplo, clases a priori y características de las distribuciones de probabilidad) se pueden aproximar con frecuencias relativas del conjunto de entrenamiento. Estas son las estimaciones de máxima verosimilitud de las probabilidades. Una clase a priori se puede calcular asumiendo clases equiprobables (es decir, priori = 1/ número de clases), o mediante el cálculo de una estimación de la probabilidad de clase del conjunto de entrenamiento (es decir, el priori de una clase dada = número de muestras en la clase / número total de muestras.

Para la estimación de los parámetros de la distribución de una característica, se debe asumir una distribución o generar modelos de estadística no paramétrica sobre las características del conjunto de entrenamiento.

Las hipótesis sobre las distribuciones de características son llamadas el modelo de eventos del Clasificador Bayesiano Ingenuo. La distribución multinomial y la distribución de Bernoulli son populares para características discretas como las usadas por ejemplo en la clasificación de documentos (incluyendo el filtrado de spam). Estas hipótesis conducen a dos modelos distintos, que a menudo se confunden. Cuando se trata con los datos continuos, una hipótesis típica es que los valores continuos asociados a cada clase se distribuyen según una Distribución normal.

Por ejemplo, supongamos que los datos de entrenamiento contienen un atributo continuo, a. En primer lugar, segmentar los datos por la clase, y a continuación, calcular la media y la varianza de a en cada clase. Donde μc es la normal de a asociada a la clase Cα, y σ_C^2 es la varianza de a asociada a la clase Cα. Entonces, la densidad de probabilidad de un cierto valor dada una clase, P(ai = vj/Cα), se puede calcular agregando v en la ecuación de una distribución normal con parámetros μc y σ_C^2 . En este caso se podrá hacer la estimación:

$$P(a_j = v_j/C_\alpha) = \frac{1}{\sqrt{2\pi\ \sigma_C^2}} e^{-\frac{(v_j-\mu_C)^2}{2\sigma_C^2}}$$

Otra técnica común para la manipulación de valores continuos es usar binning para discretizar los valores de las características, obteniendo un nuevo conjunto de características de la distribución de Bernoulli. En general, el método de distribución es una mejor opción si hay pocos datos de entrenamiento, o si se conoce la distribución precisa de los datos.

El método de discretización tiende a ser mejor si hay una gran cantidad de datos de entrenamiento, ya que va a aprender para adaptarse a la distribución de los datos. Bayes Ingenuo se utiliza normalmente cuando está disponible una gran cantidad de datos (los modelos más caros computacionalmente pueden lograr una mayor precisión), se prefiere generalmente el método de discretización que el método de distribución.

Definición 5.5 (Construcción de un clasificador del modelo de probabilidad). El clasificador Bayes combina este modelo con una regla de decisión. La primera regla en común, es para recoger la hipótesis del más probable, también conocido como el máximo a posteriori o MAP. El clasificador Bayer (la función classify) se define como:

$$classify(f_1, \dots, f_n) = argmax\ P(C = c) \prod_{i=1}^{n} P(F_i = f_i/C = c).$$

Para ilustrar la aplicación de la clasificación damos un ejemplo a continuación. Supongamos que tengamos que clasificar una persona en hombre o mujer basándonos en las características de sus medidas: peso, altura y número de pie.

Entrenamiento (Entrenamiento previo). Haciendo una distribución Gaussiana extraemos los datos y obtenemos la media y la varianza de cada característica.

En este caso nos encontramos con una distribución equiprobable, es decir que tienen la misma probabilidad. P(hombre)=0.5 y P(mujer)=0.5.

Testing. Ahora recibimos unos datos para ser clasificados como hombre o mujer.

Nos interesa saber la probabilidad a posteriori de los dos casos, según es hombre o mujer.

$$posteriori(hombre) = \frac{P(hombre)\ p(altura/hombre)\ p(peso/hombre)\ p(nrodepie/hombre)}{Evidencia}$$

$$posteriori(mujer) = \frac{P(mujer)\ p(altura/mujer)\ p(peso/mujer)\ p(nrodepie/mujer)}{Evidencia}$$

La evidencia (también denominada constante de normalización) se puede calcular como sigue

$$Evidencia = P(hombre)\ p(altura/hombre)\ p(peso/hombre)\ p(nrodepie/hombre) + P(mujer)\ p(altura/mujer)\ p(peso/mujer)\ p(nrodepie/mujer)$$

En este caso nos encontramos con una distribución equiprobable, es decir que tiene la misma probabilidad. P(hombre)=0.5 y P(mujer)=0.5.

P(hombre) = 0.5

$$\boldsymbol{p(altura/hombre) = \frac{1}{\sqrt{2\pi\sigma^2}}\ exp\left(\frac{-\left((6-\mu)\right)^2}{2\sigma^2}\right) \approx 1.5789}$$

donde μ = 5.855 y σ^2 = 3.5033e − 02 son los parámetros de la distribución normal que han sido determinados previamente en el entrenamiento.

$p(peso/hombre) = 5.9881e - 06$
$p(numero\ de\ pie/hombre) = 1.3112e - 3$
$posteriori(hombre) = el\ producto = 6.1984e - 09$
$P(mujer) = 0.5$
$p(altura/mujer) = 2.2346e - 1$
$p(peso/mujer) = 1.6789e - 2$
$p(numero\ de\ pie/mujer) = 2.8669e - 1$
$posteriori(mujer) = el\ producto = 5.3778e - 04$

En este caso el numerador a posteriori más grande es el de la mujer, por eso determinamos que los datos son de mujer.

Anexo 6: Datos de pacientes con drogodependencias (Archivo Histórico)

Paciente	nacimiento	caid	edad	sexo	Lugar consumo	hijos	convivencia	estudios	trabajo	antecedentes	i_oh	f_oh	i_tabaco	d_tabaco	i_bdz	f_bdz	i_hero	f_hero	i_coca	f_coca	i_canna	f_canna	i_anfet	f_anfet	i_caic	i_trat	ingprev	cocabase	ttoing	dxprev	peso	altura	resultado
1	1994	20	PARLA	Mujer		0	Familia de origen	primarios	Paro	no	9	1	9	1	7	1			5	1	9	1						No			74	167	MAL
2	1993	22	COLLADO VILLALBA	Hombre	Casa	0		primarios	ILT	no	4	1	7	2					4	1	7	3			0	22		No			70	173	BIEN
3	1991	23	m-105	Mujer	Casa, Calle	0	Pareja	primarios	Pensionista	si	8	5							8	5	8	1			0	20	0			TLP	52	159	MAL
4	1988	26	Aranjuez	Hombre	Casa	2	Pareja e hijos		ILT	no									10	2	12	1				25		No	BDZ, Estabilizadores		94	185	MAL
5	1988	26	San Blas	Hombre	Calle	0	Familia de origen	bachiller o fp	Paro	no							0	6	4	1	10	3			0		0	No			62	183	MAL
6	1987	27	getafe	Hombre	Casa, Calle, Bar, Trabajo	0	Familia de origen	bachiller o fp	Paro	si	5	2	6	1	4	1			3	2								No	AD		89	180	MAL
7	1986	28	fuenlabrada	Mujer	Calle, Bar	1	Familia de origen	bachiller o fp		no	14	3							14	3	16	3			0	28	0	No	AP, BDZ, AD		50	171	MAL
8	1986	28	LOS PINARES	Hombre	Casa, Calle, Bar	0	Familia de origen	sin estudios	Paro	si			14	1	3	1			8	1	14	1							BDZ	f90	80	184	MAL
9	1986	29	torrejón	Hombre	Indiferente	0	Familia de origen	primarios	ILT	no	13	2							13	2						22	0	No	AP, BDZ, AD, Estabilizadores		59	164	MAL
10	1984	30	getafe	Hombre	Indiferente	0	Familia de origen	sin estudios		no	14	5			9	1	9	6	14	1	14	1	1	6	0	22	2	No	AP, BDZ, Estabilizadores		76	171	MAL
11	1983	31	COLLADO VILLALBA	Hombre	Indiferente	1	Familia de origen		Paro	no	21	5	17	1					16	5						15	0	No	disulfiram		75	177	MAL
12	1983	31	sur	Hombre	Trabajo, Coche	0	Familia de origen	primarios	ILT	no	17	2	15	2					11	3					0	21	0	No		f90	100	172	BIEN
13	1983	31	este	Hombre	Calle	1	Familia de	bachiller o fp	Paro	no		5							16	4			15	6	0	22	0	No	AP, BDZ, Estabilizado	F4X	67	170	BIEN

							origen																						res, DISULFIRAM				
14	1980	32	Norte	Hombre	Indiferente	0	Familia de origen	bachiller o fp	ILT	no									9	2	13	2			1	20	1		AP, Estabilizadores	F31	100	188	MAL
15	1981	33	Majadahonda	Hombre	Indiferente	0	Familia de origen	universitarios	ILT	no	19	5							2	1	17	1				19	2	No	AP, Estabilizadores		65	179	MAL
16	1981	34	norte	Hombre		0	Familia de origen			no	12	1							16	1	20	2				25	3		BDZ, AD, Estabilizadores		47	163	MAL
17	1981	34	TETUAN	Hombre	Calle	2	Pareja e hijos	primarios	Paro	no									10	2					1	26	1	No	BDZ, AD, strattera	f90	75	179	MAL
18	1981	34	PARLA	Hombre	Casa, Calle	2	Familia de origen	primarios	Paro	no			19	20					19	1	21	1			2	23	2	No	AP, BDZ, AD, Estabilizadores, DISULFIRAM		64	170	MAL
19	1978	34		Hombre	Casa, Calle, Bar	1	Familia de origen	bachiller o fp	Paro	no	22	3							14	5	21	3	12		1	24	1	No	AP, BDZ, AD		80	172	MAL
20	1980	34	LEganes	Hombre	Calle	1	Familia de origen	primarios	Paro	no	16	1							14	3	16	1			1	22	1	No			61	174	BIEN
21	1980	34	mostoles	Hombre	Casa, Coche	0	Familia de origen	primarios	Paro	no									11	3						29	0				90	189	BIEN
22	1989	35	NORTE	Hombre	Casa	1	Familia de origen	sin estudios	ILT	no	28	1							22	1	27	1			2	23	2	No			70	183	MAL
23	1978	36	GETAFE	Hombre	Casa, Coche	0	Familia de origen	primarios	Paro	no	21	3	21	2			1	1	18	1	20	1			1	29		Si	BDZ		72	174	MAL
24	1979	36	Getafe	Hombre	Casa, Calle	0	Pareja	sin estudios	Paro	si	24	1							22	3	24	1				21	0	No			65	170	MAL
25	1976	37	GETAFE	Hombre	Casa, Bar, Coche	0	Solo	primarios	ILT	no	20	6	22	1					18	4	20	6			0	33	0	No			100	170	MAL
26	1971	37	alcala	Hombre	Casa, Calle, Bar	1	Pareja	primarios	ILT	si	16	5	16	3					8	5	6	1			1	37	2	No	BDZ	F32	129	180	BIEN
27	1977	38	MAJADAHONDA	Hombre	Casa	0	Familia de origen	primarios	Paro	no	25	5							21	1	21	1						No	AP, Estabilizadores	F2X, epilepsia mioclónica juvenil	184	84	BIEN
28	1978	38	CAD VALLECAS	Hombre	Bar	3	Familia de origen	primarios	Paro	no	20	1							21	5						36	0	No	Estabilizadores		94	178	BIEN

29	1975	38	Alcorcón	Mujer	Bar	1	Familia de origen	bachiller o fp	Paro	no									13	1					1		2	No	AD	TLP	52	166	BIEN
30	1975	39	Norte	Hombre	Casa, Calle, Bar	0	Solo	bachiller o fp	ILT		11	3			0	1			11	3	24	6			0		2	No	AP, AD, Estabilizadores		78	178	MAL
31	1975	39	Este	Hombre		2	Pareja	primarios	ILT	si									22	2	25	3							AP, BDZ		75	176	MAL
32	1975	39	aranjuez	Hombre	Casa, Bar	2	Familia de origen	bachiller o fp	ILT		24	1			5	2			7	1	19	6			0	36		No	BDZ	f90	82	176	MAL
33	1975	39	PARLA	Hombre		0	Familia de origen	sin estudios											14	3						39	0	No	BDZ, AD, Estabilizadores		104	174	MAL
34	1975	39	CAD villaverde	Hombre	Casa, Calle	2	Pareja e hijos	bachiller o fp	Paro	no									28	3						38	0	No	Estabilizadores, strattera	f90			BIEN
35	1975	39	Alcorcón	Hombre	Casa	2	Familia de origen	bachiller o fp	Paro	no	22	5							20	5	23	5						No	AP, AD	F32	78	183	BIEN
36	1974	40	TETUAN	Hombre	Casa, Calle	0	Familia de origen	universitarios	Paro	no	25	1							6	1	23	1						No	AD, Estabilizadores, DISULFIRAM	TLP	99	179	MAL
37	1975	40	este	Hombre	Casa, Bar	2		bachiller o fp	ILT		23	1							26	2					1		1	No	AP, BDZ		75	176	MAL
38	1974	40	CAD TETUAN	Hombre	Casa, Calle	0	Familia de origen	primarios	Paro	no	25	2	26	1					20	1						40		Si	BDZ, AD		78	183	BIEN
39	1974	41	vallecas	Hombre	Indiferente	0	Solo	bachiller o fp	Paro	si	29	1							23	1	24	6				33		No	AP, BDZ, Estabilizadores	F31, TLP	70	173	MAL
40	1973	41	vallecas	Hombre		3	Pareja e hijos	primarios	ILT	no	28	3							24	1								No	AD		104	175	MAL
41	1972	42	ESTE	Hombre	Bar	0	Otros	bachiller o fp	Paro	no									16	5					1	30	2	No	BDZ, AD, Estabilizadores				MAL
42	1972	42	MOSTOLES	Mujer	Casa, Trabajo	2	Pareja e hijos	bachiller o fp	ILT	no			26	2	6	6			6	2					0	41	0	No	AP, BDZ, AD		73	173	BIEN
43	1972	42		Hombre	Indiferente	3	Familia de origen	primarios	ILT	no	29	1	29	1					27	1	29	3	23	6		32	0	No	BDZ, DISULFIRAM		73	168	BIEN
44	1971	43	alcala	Hombre	Casa	1	Pareja e hijos	primarios	Paro	no	32	6	32	3					29	1					1	24	1	No	BDZ, AD	F4X	181	53	MAL
45	1969	44	M105	Hombre	Calle	0	Familia de origen	bachiller o fp	Pensionista	no	28	1	30	6			26	6	26	1	28	6			1	23	6	No	AP, BDZ, AD, Estabilizadores	F31	168	86	BIEN
46	1970	45	NORTE	Hombre	Casa	1	Pareja e hijos	bachiller o fp	Paro	si									24	3					1	38	1	Si	BDZ, AD, Estabilizadores	F32	100	182	MAL

47	1969	45	leganes	Hombre	Casa, Calle	3	Pareja e hijos		ILT	si	29	1	29	1			22	6	6	4	32	6				35		Si	BDZ, AD		95	171	MAL
48	1994	45	hortaleza	Hombre	Casa	0	Pareja	primarios	ILT	no									31	1								Si			50	155	MAL
49	1968	46	S. Fernando de Henares	Hombre	Casa	1	Solo	bachiller o fp	ILT	no	28	1	12	1					2	1						44		No	AD, Estabilizadores	F32	55	175	MAL
50	1963	50	cad vallecas	Hombre	Casa	2	Familia de origen	primarios	ILT	no					1	1			23	3							2	No	AP, BDZ, Estabilizadores	otros tp	63	176	MAL
51	1961	53	CTD CENTRO	Hombre	Casa	1	Pareja e hijos	universitarios	Paro								26	6	14	1					2	39	2	Si	BDZ, Estabilizadores			66	BIEN
52	1960	55	Vallecas	Hombre		1	Pareja		Pensionista	no	41	1	41	5			36	6	36	1	39	2			0	29		Si	BDZ		63	167	MAL
53	1944	71	GETAFE	Mujer	Casa	2	Otros	primarios	Pensionista	si									21	5								No	BDZ		53	150	MAL

Anexo 7: Cantidad outliers encontrados con el algoritmo DOCNM

Posición	Edad	i_oh	f_oh	i_tabaco	d_tabaco	i_bdz	f_bdz	i_hero	f_hero	i_coca	f_coca	i_canna	f_canna	i_anfet	f_anfet	Total
1	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	23	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
4	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	28	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
8	28	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
9	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	30	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
11	31	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
12	31	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
13	31	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
14	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	33	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
16	34	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
17	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
18	34	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
19	34	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
20	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	36	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
25	37	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
26	37	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
27	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

28	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	38	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
30	39	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
31	39	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
32	39	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
33	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	39	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
35	39	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
36	40	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2
37	40	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
38	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	42	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
42	42	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
43	42	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
44	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	45	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2
48	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	46	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
50	50	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
51	53	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
52	55	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
53	71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Anexo 8: Cantidad de outliers encontrados con la librería “outliers” de R

Posición	Edad	i_oh	f_oh	i_tabaco	d_tabaco	i_bdz	f_bdz	i_hero	f_hero	i_coca	f_coca	i_canna	f_canna	i_anfet	f_anfet	Total
1	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	23	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
4	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	30	0	0	0	0	1	0	0	0	0	0	0	0	1	1	3
11	31	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
12	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	31	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
14	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	34	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
20	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	36	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
24	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	37	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1
26	37	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
27	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

28	38	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
29	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	39	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
31	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	39	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
33	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	39	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
36	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	41	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
40	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	42	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
42	42	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
43	42	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
44	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	44	0	1	0	1	0	0	0	0	0	0	0	1	0	0	3
46	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	45	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
48	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52	55	1	0	1	1	0	0	1	0	1	0	1	0	0	0	6
53	71	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1

Cuadro a detalle de outliers encontrados con el algoritmo DOCNM

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
Edad	20	22	23	26	26	27	28	28	29	30	31	31	31	32	33	34	34	34	34	34	34	35	36	36	37	37	38	38	38	39	39	39	39	39	39	40	40	40	41	41	42	42	42	43	44	45	45	45	46	50	53	55	71
Total Outliers	0	0	1	0	0	0	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	2	1	0	0	0	1	1	1	0	0	0	2	0	1	1	1	1	0

Cuadro a detalle de outliers encontrados con la librería “outliers” de R

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
Edad	20	22	23	26	26	27	28	28	29	30	31	31	31	32	33	34	34	34	34	34	34	35	36	36	37	37	38	38	38	39	39	39	39	39	39	40	40	40	41	41	42	42	42	43	44	45	45	45	46	50	53	55	71
Total Outliers	0	0	1	0	0	0	0	0	0	3	1	0	1	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1	0	1	0	1	0	0	1	0	0	0	1	0	1	1	1	0	3	0	1	0	0	0	0	6	1

Anexo 9: Outliers encontrados con las librerías de R y RCommader

Posición	Edad	i_oh	f_oh	i_tabaco	d_tabaco	i_bdz	f_bdz	i_hero	f_hero	i_coca	f_coca	i_canna	f_canna	i_anfet	f_anfet	Total
1	20	1	1	1	1	0	0	0	0	0	0	1	1	0	0	6
2	22	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
3	23	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
4	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	26	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2
6	27	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
7	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	30	0	0	0	0	1	0	1	0	0	0	0	0	1	0	3
11	31	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
12	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	31	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
14	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	35	1	0	0	0	0	0	0	0	0	0	1	0	0	0	2
23	36	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2
24	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	37	0	1	0	0	0	0	0	0	0	0	0	1	0	0	2
26	37	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
27	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

28	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	39	1	0	0	0	1	0	0	0	0	0	0	0	0	0	2
31	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	42	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
43	42	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
44	43	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
45	44	0	0	0	1	0	0	1	0	0	0	0	0	0	0	2
46	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	45	0	0	0	0	0	0	0	0	1	0	0	1	0	0	2
48	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	46	0	0	1	0	0	0	0	0	1	0	0	0	0	0	2
50	50	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2
51	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52	55	1	1	1	1	0	0	1	1	1	1	1	1	0	0	10
53	71	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2

Cuadro a detalle de outliers encontrados con la librería "*Rcmdr*" de R

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
Edad	20	22	23	26	26	27	28	28	29	30	31	31	31	32	33	34	34	34	34	34	34	35	36	36	37	37	38	38	38	39	39	39	39	39	39	39	40	40	40	41	41	42	42	42	43	44	45	45	45	46	50	53	55	71
Total Outliers	6	1	1	0	2	1	0	0	0	3	1	0	1	0	0	0	0	0	0	0	0	2	2	0	2	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	2	0	2	2	0	10	2

Cabe mencionar que para un mejor análisis se recomienda la eliminación de los candidatos posibles a outliers de ese modo se obtiene una base de datos más fiable y por ende su resultado. El total de outlier detectado por cada edad del paciente drogodependiente se muestran en detalle en los Anexo 6, 7 y 8, de los cuales estos pacientes drogodependientes tienen una coincidencia en los tres métodos: 23, 30, 31(1), 31(3), 37(1), 37(2), 39(1), 42(2), 45(2) y la 55.

Anexo 10-A: Resultado de predicción con R

Finalmente se guarda la tabla de NuevosPacientes con la predicción:

```
write.csv(NuevosPacientes,"C:/Users/Predict_NuevosPacientes
.csv")
```

```
table(Prediccion, NuevosPacientes[,33])
```

#El -32 se debe a que la variable dependiente, Prediccion_de_resultado, es en número de columna 33.

Prediccion BIEN MAL
BIEN 0 0
MAL 0 1

```
print(Probabilidades)
```

Brinda información del algoritmo de clasificación **Naive Bayes**

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
     BIEN       MAL
0.3018868 0.6981132

Conditional probabilities:
      edad
Y               C          D          E          F          G
H          I          J
  BIEN 0.00000000 0.06250000 0.00000000 0.00000000 0.25000000 0.
62500000 0.06250000 0.00000000
  MAL  0.02702703 0.00000000 0.02702703 0.18918919 0.21621622 0.
48648649 0.02702703 0.02702703

      caid
Y          Alcala   Alcorcon   Aranjuez Cad Tetuan Cad Vallecas
Cad Villaverde
  BIEN 0.06666667 0.13333333 0.06666667 0.06666667   0.06666667
0.06666667
  MAL  0.02777778 0.00000000 0.05555556 0.00000000   0.02777778
0.00000000
      caid
Y      Collado Villalba Ctd Centro       Este Fuenlabrada     Ge
tafe  Hortaleza    Leganes
  BIEN       0.00000000 0.06666667 0.06666667  0.00000000 0.0000
0000 0.00000000 0.06666667
  MAL        0.02777778 0.00000000 0.08333333  0.02777778 0.1666
6667 0.02777778 0.02777778
      caid
Y      Los Pinares       M105 Majadahonda   Mostoles      Norte
Parla   San Blas
  BIEN  0.00000000 0.06666667  0.06666667 0.13333333 0.00000000
0.00000000 0.00000000
  MAL   0.02777778 0.02777778  0.02777778 0.00000000 0.13888889
0.08333333 0.02777778
```

```
      caid
Y      San Fernando de Henares        Sur     Tetuan  TorrejÃ³n
Vallecas
  BIEN              0.00000000 0.06666667 0.00000000 0.00000000
0.00000000
  MAL               0.02777778 0.00000000 0.05555556 0.02777778
0.08333333

      sexo
Y         Hombre     Mujer
  BIEN 0.8750000 0.1250000
  MAL  0.8918919 0.1081081

      lugarconsumo
Y               A          B          C          D          E
F          G          H
  BIEN 0.00000000 0.00000000 0.06250000 0.12500000 0.00000000 0.
06250000 0.00000000 0.06250000
  MAL  0.03225806 0.03225806 0.09677419 0.16129032 0.06451613 0.
03225806 0.03225806 0.00000000
      lugarconsumo
Y               I          J          K          L          M
O
  BIEN 0.06250000 0.06250000 0.18750000 0.18750000 0.12500000 0.
06250000
  MAL  0.00000000 0.00000000 0.25806452 0.06451613 0.03225806 0.
19354839

      hijos
Y               A          B          C          D
  BIEN 0.37500000 0.31250000 0.18750000 0.12500000
  MAL  0.51351351 0.21621622 0.21621622 0.05405405

      convivencia
Y               A          B          C          D          E
  BIEN 0.75000000 0.18750000 0.06250000 0.00000000 0.00000000
  MAL  0.52777778 0.16666667 0.13888889 0.11111111 0.05555556

      estudios
Y            A       B       C       D
  BIEN 0.00000 0.56250 0.37500 0.06250
  MAL  0.15625 0.40625 0.37500 0.06250

      trabajo
Y               A          B          C
  BIEN 0.62500000 0.31250000 0.06250000
  MAL  0.45454545 0.45454545 0.09090909

      antecedentes
Y              A         B
  BIEN 0.1333333 0.8666667
  MAL  0.2727273 0.7272727

      i_oh
Y               A          B          C          D          E
F          G          I
  BIEN 0.00000000 0.10000000 0.00000000 0.40000000 0.30000000 0.
20000000 0.00000000 0.00000000
  MAL  0.04166667 0.08333333 0.20833333 0.08333333 0.29166667 0.
20833333 0.04166667 0.04166667
```

```
      f_oh
Y               1          A          B          C          E
F
  BIEN 0.00000000 0.45454545 0.18181818 0.00000000 0.36363636 0.
00000000
  MAL  0.08333333 0.37500000 0.08333333 0.20833333 0.16666667 0.
08333333

      i_tabaco
Y               A          B          C          D          E
F          G          J
  BIEN 0.14285714 0.00000000 0.14285714 0.14285714 0.00000000 0.
57142857 0.00000000 0.00000000
  MAL  0.00000000 0.18181818 0.18181818 0.18181818 0.18181818 0.
09090909 0.09090909 0.09090909

      d_tabaco
Y               A          B          C          E          F
  BIEN 0.28571429 0.42857143 0.14285714 0.00000000 0.14285714
  MAL  0.54545455 0.18181818 0.09090909 0.18181818 0.00000000

      i_bdz
Y              A         B
  BIEN 0.0000000 1.0000000
  MAL  0.7142857 0.2857143

      f_bdz
Y              A         B         F
  BIEN 0.0000000 0.0000000 1.0000000
  MAL  0.8571429 0.1428571 0.0000000

      i_hero
Y        A   B   E   F   H
  BIEN 0.0 0.0 0.0 1.0 0.0
  MAL  0.4 0.2 0.2 0.0 0.2

      f_hero
Y        A   F
  BIEN 0.0 1.0
  MAL  0.2 0.8

      i_coca
Y               A          B          C          D          E
F          G          H
  BIEN 0.06250000 0.12500000 0.31250000 0.18750000 0.12500000 0.
18750000 0.00000000 0.00000000
  MAL  0.13513514 0.21621622 0.16216216 0.16216216 0.21621622 0.
05405405 0.02702703 0.02702703

      f_coca
Y               1          A          B          C          D
E
  BIEN 0.06250000 0.37500000 0.06250000 0.25000000 0.06250000 0.
18750000
  MAL  0.00000000 0.45945946 0.18918919 0.16216216 0.05405405 0.
13513514

      i_canna
Y               B          C          D          E          F
G          H
```

```
  BIEN 0.00000000 0.00000000 0.16666667 0.50000000 0.33333333 0.
00000000 0.00000000
  MAL  0.08695652 0.17391304 0.30434783 0.30434783 0.04347826 0.
04347826 0.04347826

      f_canna
Y              A         B         C         E         F
  BIEN 0.5000000 0.0000000 0.1666667 0.1666667 0.1666667
  MAL  0.4782609 0.1304348 0.1739130 0.0000000 0.2173913

      i_anfet
Y        A   C   E
  BIEN 0.0 0.5 0.5
  MAL  0.5 0.5 0.0

      f_anfet
Y      FALSE
  BIEN     1
  MAL      1

      i_caic
Y              A         B         C
  BIEN 0.3750000 0.5000000 0.1250000
  MAL  0.4444444 0.4444444 0.1111111

      i_trat
Y               A          B          C          D          E
F          G          H
  BIEN 0.33333333 0.08333333 0.00000000 0.00000000 0.08333333 0.
25000000 0.16666667 0.00000000
  MAL  0.12000000 0.04000000 0.08000000 0.04000000 0.00000000 0.
36000000 0.24000000 0.08000000
      i_trat
Y               I
  BIEN 0.08333333
  MAL  0.04000000

      ingprev
Y               A          B          C          D          G
  BIEN 0.58333333 0.08333333 0.25000000 0.00000000 0.08333333
  MAL  0.36363636 0.27272727 0.31818182 0.04545455 0.00000000

      cocabase
Y              A         B
  BIEN 0.1333333 0.8666667
  MAL  0.1562500 0.8437500

      ttoing
Y               A          B          C          E          F
G          H          I
  BIEN 0.00000000 0.08333333 0.08333333 0.08333333 0.00000000 0.
00000000 0.00000000 0.00000000
  MAL  0.03333333 0.03333333 0.00000000 0.06666667 0.03333333 0.
13333333 0.10000000 0.03333333
      ttoing
Y               J          K          L          M          N
O          P          Q
  BIEN 0.00000000 0.08333333 0.08333333 0.00000000 0.08333333 0.
00000000 0.08333333 0.08333333
  MAL  0.03333333 0.06666667 0.06666667 0.03333333 0.03333333 0.
06666667 0.00000000 0.00000000
```

```
      ttoing
Y               S          T          U          W          Y
  BIEN 0.08333333 0.08333333 0.08333333 0.00000000 0.08333333
  MAL  0.16666667 0.06666667 0.00000000 0.03333333 0.00000000

      dxprev
Y               A          B          C          D          E
F          G          H
  BIEN 0.11111111 0.00000000 0.11111111 0.11111111 0.22222222 0.
11111111 0.00000000 0.33333333
  MAL  0.00000000 0.09090909 0.18181818 0.09090909 0.18181818 0.
09090909 0.09090909 0.27272727

      imc
Y               B          C          D          E          F
  BIEN 0.00000000 0.00000000 0.60000000 0.26666667 0.13333333
  MAL  0.02777778 0.11111111 0.50000000 0.19444444 0.16666667
```

Anexo 10-B: Resultado de predicción con R

Finalmente se guarda la tabla de NuevosPacientes con la predicción:

```
write.csv(NuevosPacientes,"C:/Users/Predict_NuevosPacientes
.csv")
```

```
table(Prediccion, NuevosPacientes[,33])
```

#El -32 se debe a que la variable dependiente, Prediccion_de_resultado, es en número de columna 33.

Prediccion BIEN MAL

BIEN 0 0
MAL 0 1

```
print(Probabilidades)
```

Brinda información del algoritmo de clasificación **Naïve Bayes**

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
     BIEN       MAL
0.3095238 0.6904762

Conditional probabilities:
      edad
Y              C          D          F          G          H
I          J
  BIEN 0.00000000 0.07692308 0.00000000 0.23076923 0.61538462 0.
07692308 0.00000000
  MAL  0.03448276 0.00000000 0.20689655 0.24137931 0.48275862 0.
00000000 0.03448276

      caid
Y          Alcala   Alcorcon   Aranjuez Cad Tetuan Cad Vallecas
Cad Villaverde Ctd Centro
  BIEN 0.00000000 0.16666667 0.08333333 0.08333333   0.08333333
0.08333333 0.08333333
  MAL  0.03571429 0.00000000 0.07142857 0.00000000   0.03571429
0.00000000 0.00000000
      caid
Y            Este Fuenlabrada     Getafe  Hortaleza    Leganes L
os Pinares       M105
  BIEN 0.00000000  0.00000000 0.00000000 0.00000000 0.08333333
0.00000000 0.08333333
  MAL  0.07142857  0.03571429 0.14285714 0.03571429 0.00000000
0.03571429 0.00000000
      caid
Y      Majadahonda   Mostoles      Norte      Parla   San Blas S
an Fernando de Henares
```

```
  BIEN  0.08333333 0.08333333 0.00000000 0.00000000 0.00000000
0.00000000
  MAL   0.03571429 0.00000000 0.14285714 0.10714286 0.03571429
0.03571429
      caid
Y             Sur     Tetuan   Torrejon   Vallecas
  BIEN 0.08333333 0.00000000 0.00000000 0.00000000
  MAL  0.00000000 0.07142857 0.03571429 0.07142857

      sexo
Y          Hombre      Mujer
  BIEN 0.92307692 0.07692308
  MAL  0.89655172 0.10344828

      lugarconsumo
Y               A          C          D          E          F
G          I          J
  BIEN 0.00000000 0.00000000 0.15384615 0.00000000 0.07692308 0.
00000000 0.07692308 0.07692308
  MAL  0.04166667 0.08333333 0.12500000 0.08333333 0.04166667 0.
04166667 0.00000000 0.00000000
      lugarconsumo
Y               K          L          M          O
  BIEN 0.23076923 0.15384615 0.15384615 0.07692308
  MAL  0.33333333 0.08333333 0.00000000 0.16666667

      hijos
Y               A          B          C          D
  BIEN 0.46153846 0.23076923 0.15384615 0.15384615
  MAL  0.48275862 0.20689655 0.27586207 0.03448276

      convivencia
Y               A          B          C          D          E
  BIEN 0.84615385 0.15384615 0.00000000 0.00000000 0.00000000
  MAL  0.60714286 0.17857143 0.10714286 0.07142857 0.03571429

      estudios
Y               A          B          C          D
  BIEN 0.00000000 0.61538462 0.30769231 0.07692308
  MAL  0.14814815 0.40740741 0.37037037 0.07407407

      trabajo
Y               A          B          C
  BIEN 0.69230769 0.23076923 0.07692308
  MAL  0.50000000 0.46153846 0.03846154

      antecedentes
Y               A          B
  BIEN 0.08333333 0.91666667
  MAL  0.26923077 0.73076923

      i_oh
Y               A          B          C          D          E
F          G
  BIEN 0.00000000 0.11111111 0.00000000 0.33333333 0.33333333 0.
22222222 0.00000000
  MAL  0.05882353 0.05882353 0.17647059 0.05882353 0.35294118 0.
23529412 0.05882353
```

```
      f_oh
Y              1          A          B          C          E
F
  BIEN 0.00000000 0.55555556 0.22222222 0.00000000 0.22222222 0.
00000000
  MAL  0.11764706 0.41176471 0.11764706 0.23529412 0.05882353 0.
05882353

      i_tabaco
Y             A         B         C         D         E
F         G
  BIEN 0.2000000 0.0000000 0.2000000 0.0000000 0.0000000 0.60000
00 0.0000000
  MAL  0.0000000 0.2857143 0.2857143 0.1428571 0.1428571 0.00000
00 0.1428571

      d_tabaco
Y             A         B         C         F
  BIEN 0.4000000 0.4000000 0.0000000 0.2000000
  MAL  0.5714286 0.2857143 0.1428571 0.0000000

      i_bdz
Y        A   B
  BIEN
  MAL  0.8 0.2

      f_bdz
Y        A   B
  BIEN
  MAL  0.8 0.2

      i_hero
Y      A F
  BIEN 0 1
  MAL  1 0

      f_hero
Y        A   F
  BIEN 0.0 1.0
  MAL  0.5 0.5

      i_coca
Y              A          B          C          D          E
F          G
  BIEN 0.07692308 0.00000000 0.38461538 0.15384615 0.15384615 0.
23076923 0.00000000
  MAL  0.17241379 0.20689655 0.13793103 0.10344828 0.27586207 0.
06896552 0.03448276

      f_coca
Y              1          A          B          C          E
  BIEN 0.07692308 0.46153846 0.00000000 0.30769231 0.15384615
  MAL  0.00000000 0.51724138 0.24137931 0.17241379 0.06896552

      i_canna
Y              B          C          D          E          F
  BIEN 0.00000000 0.00000000 0.20000000 0.40000000 0.40000000
  MAL  0.05882353 0.17647059 0.35294118 0.35294118 0.05882353
```

```
f_canna
Y              A         B         C         E         F
  BIEN 0.4000000 0.0000000 0.2000000 0.2000000 0.2000000
  MAL  0.5294118 0.1176471 0.2352941 0.0000000 0.1176471

      i_anfet
Y      C E
  BIEN 0 1
  MAL  1 0

      f_anfet
Y      FALSE
  BIEN     1
  MAL

      i_caic
Y              A         B         C
  BIEN 0.2000000 0.6000000 0.2000000
  MAL  0.2500000 0.5833333 0.1666667

      i_trat
Y               A          B          C          E          F
G          H          I
  BIEN 0.22222222 0.11111111 0.00000000 0.11111111 0.22222222 0.
22222222 0.00000000 0.11111111
  MAL  0.16666667 0.05555556 0.05555556 0.00000000 0.38888889 0.
27777778 0.05555556 0.00000000

      ingprev
Y              A         B         C         D         G
  BIEN 0.5555556 0.1111111 0.2222222 0.0000000 0.1111111
  MAL  0.3125000 0.3750000 0.2500000 0.0625000 0.0000000

      cocabase
Y              A         B
  BIEN 0.1666667 0.8333333
  MAL  0.1200000 0.8800000

      ttoing
Y              A          B          E          F          G
H          I          K
  BIEN 0.00000000 0.11111111 0.00000000 0.00000000 0.00000000 0.
00000000 0.00000000 0.11111111
  MAL  0.04166667 0.04166667 0.08333333 0.04166667 0.12500000 0.
08333333 0.04166667 0.08333333
      ttoing
Y              L          M          N          O          P
Q          S          T
  BIEN 0.11111111 0.00000000 0.11111111 0.00000000 0.11111111 0.
11111111 0.00000000 0.11111111
  MAL  0.04166667 0.04166667 0.04166667 0.08333333 0.00000000 0.
00000000 0.16666667 0.08333333
      ttoing
Y              U          Y
  BIEN 0.11111111 0.11111111
  MAL  0.00000000 0.00000000
```

```
dxprev
Y              A         B         C         D         E
F         G         H
  BIEN 0.1428571 0.0000000 0.1428571 0.1428571 0.1428571 0.00000
00 0.0000000 0.4285714
  MAL  0.0000000 0.1000000 0.1000000 0.1000000 0.2000000 0.10000
00 0.1000000 0.3000000

      imc
Y               B          C          D          E          F
  BIEN 0.00000000 0.00000000 0.58333333 0.33333333 0.08333333
  MAL  0.03448276 0.13793103 0.48275862 0.20689655 0.13793103
```