UNIVERSIDAD
COMPLUTENSE
MADRID

Proyecto de Innovación

Convocatoria 2019/2020

Nº de proyecto 351

Desarrollo de algoritmos predictivos por inteligencia artificial (Deep-learning) para asegurar el éxito del alumno.

Antonio López Farré

Facultad de Medicina

Departamento de Medicina

1. Objetivos propuestos en la presentación del proyecto

1.- Diseñar un algoritmo predictivo por Deep-learning para identificar al comienzo del curso académico a aquellos alumnos/as con riesgo de no superar en el grado de Nutrición y Dietética Humana las asignaturas de patología médica aplicada (3er curso), epidemiología y salud pública (4º curso) y nutrición y alimentación en el paciente quirúrgico (3er y 4º curso).
2.- Implementar esta iniciativa como parte del proceso de mejora docente en el grado de Nutrición y Dietética Humana de la Facultad de Medicina.

Esta iniciativa pone en valor:
1.- Para la Facultad de Medicina, la realización de actividades que mejorará la docencia particularmente para aquellos alumnos/as que tengan mayor dificultad en superar las asignaturas, ya que se podrían predecir al comienzo del curso la necesidad de estos alumnos/as de mayor atención por parte del profesorado.
2.- Esta iniciativa de crear un algoritmo por Deep-learning para el grado de Nutrición y Dietética Humana sería un modelo exportable para asignaturas de otros grados de las Facultades de la Universidad Complutense de Madrid, e incluso para otras universidades.

2. Objetivos alcanzados

En este proyecto se ha diseñado un conjunto de predictores de desempeño académico de estudiantes por asignatura concreta utilizando en los resultados de los cursos anteriores y las variables demográficas. Un problema que se ha arrastrado a lo largo de las distintas ejecuciones es la escasez o ausencia de casos con suspensos, siendo necesario introducir una serie de cotas a la hora de entrenar los modelos.
Entre las distintas pruebas realizadas se obtuvieron los siguientes modelos:

Predicción de la asignatura de Patología Médica Aplicada tomando las notas de Fisiología y de acceso a la universidad:

Caso 1

Para este primer caso se han generado una serie de modelos para identificar aquellos alumnos que van a sacar una nota superior a 5.

Número de individuos con nota superior a 5: 59

Total de individuos: 126

Resultados:

- Las variables principales en las que se apoyó los modelos fueron principalmente la nota de fisiología,el cuestionario de patología médica aplicada, la edad y la nota de acceso a la universidad.

- Los resultados obtenidos rondan una tasa de clasificación del 65%, llegando a alcanzar un 78% con GridSearchCV (svm)

Caso 2

En este caso la variable a predecir son alumnos con una nota superior a 6.

Número de individuos con nota superior a 6: 29

Total de individuos: 126

Resultados:

- Las variables principales, coinciden en gran medida con las nombradas en el caso 1, apareciendo entre las primeras posiciones el número de horas de trabajo semanales.
- En éste caso los resultados mejoraron hasta alcanzar una media del 70% llegando a alcanzar un 75% con Gradient Boosting y Random Forest

Predicción de la asignatura de Epidemiología y Salud Pública tomando las notas de Patología Médica Aplicada, Fisiología, Dietoterapia y de acceso a la universidad:

Caso1

En este caso la variable a predecir son alumnos con una nota igual o superior a 7.

Número de individuos con nota igual o superior a 7: 6

Total de individuos: 66

Como se puede observar, existe un número muy pequeño de casos con alumnos con una nota igual o superior a 7, es por ello que los modelos tienen una tasa de exactitud muy buena. Estos resultados pueden parecer buenos, pero en realidad no lo son ya que los modelos identifican que la mayoría de la población no obtiene una nota igual o superior a 7, y por lo tanto clasificando a todos los alumnos como que no van a conseguir esta nota, se consigue una tasa de acierto muy elevada.

Resultados:

- Las variables principales en las que se apoyó el modelo fueron principalmente la nota de Patología Médica Aplicada y el número de horas que trabaja a la semana.

Caso2

En este caso la variable a predecir son alumnos con una nota superior a 8.

Número de individuos con nota superior a 8: 39

Total de individuos: 66

En este caso tenemos dos poblaciones de individuos más igualadas y se obtienen resultados más fiables.

Resultados:

- Las variables principales en las que se apoyan los modelos fueron principalmente la nota de Dietoterapia, Personas en el hogar, residencia, nota de Acceso a la Universidad, nacionalidad, nota Fisiología, número de hora que trabaja a la semana y el género.
- Los resultados obtenidos rondan una tasa de clasificación del 70%, llegando a alcanzar un 76% con KNeighborsClassifier.

4. Metodología empleada en el proyecto

La metodología de este proyecto requiere el trabajo de un equipo multidisciplinar, constituido por:

Profesores de Facultad de Medicina (Departamentos de Medicina, Salud Pública y Materno-Infantil, Cirugía, Fisiología)

Profesores de Facultad de Informática (Departamento de Arquitectura de Computadores y Automática)

Así como de personal de administración y servicios (PAS) y estudiantes.

Para desarrollar el algoritmo predictivo por Deep-learning se utilizarán una serie de “inputs o neuronas” de variables demográficas como género, edad, nacionalidad, lugar de residencia, tipo de residencia (familiar, colegio mayor, piso con compañeros, etc.), número de personas con la que convive, si es activo laboralmente y número de horas trabajadas, así como de una serie de variables académicas (calificaciones de acceso a la universidad, calificaciones en prácticas/seminarios y en exámenes parciales y finales de asignaturas previas especialmente relevantes para la asignatura sobre la que se realizará el algoritmo predictivo). Para recoger la información de estas variables los estudiantes contestarán un sencillo cuestionario.

Este diseño experimental se utilizará para desarrollar un algoritmo predictivo por Deep-learning en las siguientes asignaturas del grado de Nutrición y Dietética Humana:

1.- Algoritmo predictivo de éxito para la asignatura de Patología Médica Aplicada de 3er curso, basándose en las variables demográficas y académicas anteriormente citadas. En el análisis de variables académicas se utilizarán las calificaciones obtenidas en la asignatura de Fisiología de 1er curso que aporta aspectos fundamentales para la comprensión de la asignatura Patología Médica Aplicada.

2.- Algoritmo predictivo de éxito para la asignatura de Epidemiología y Salud Pública de 4º curso, basándose en las variables demográficas y académicas anteriormente citadas. En el análisis de variables académicas se utilizarán las calificaciones obtenidas en la asignatura de Patología Médica Aplicada de 3er curso que aporta aspectos fundamentales para la comprensión de la asignatura Epidemiología y Salud Pública.

3.- Algoritmo predictivo de éxito para la asignatura de Nutrición y Alimentación en el paciente quirúrgico de 3er y 4º curso, basándose en las variables demográficas y académicas anteriormente citadas. En el análisis de variables académicas se utilizarán las calificaciones obtenidas en la asignatura de Dietoterapia de 2º curso que aporta aspectos fundamentales para la comprensión de la asignatura Nutrición y Alimentación en el paciente quirúrgico.

El diseño experimental permitirá conocer al principio del curso académico, mediante el algoritmo predictivo generado por Deep-learning, que alumnos tendrán problemas para solventar con éxito las asignaturas anteriormente citadas (Patología Médica Aplicada de

3er curso, Epidemiología y Salud Pública de 4º curso y Nutrición y Alimentación en el paciente quirúrgico de 3er y 4º curso de grado de Nutrición y Dietética Humana) por lo que el profesorado podrá implantar un reforzamiento docente mediante tutorías y talleres/seminarios y actividades virtuales específicas y complementarias. Obviamente, este reforzamiento también será ofertado al resto de alumnos.

Si se concediera el proyecto, en el curso 2019/2020 se desarrollará el proyecto y se obtendrán los algoritmos predictivos por Deep-learning, ya que requiere un autoaprendizaje con los datos obtenidos de los estudiantes en los cuestionarios. La implantación se realizará en el curso académico 2020/2021 y también se implantaría en formación de pos-grado.

5. Recursos humanos

1.- PERSONAL DOCENTE INVESTIGADOR (PDI)

Facultad de Medicina:

- Dpto. Medicina
    - Antonio José López Farré (50704818-F)
    - Miguel Ángel García Fernández (51600788-N)
    - Luis Collado Yurrita (05379605-C)
    - Luis Álvarez Salas-Walther (00670197-T)
- Dpto. Salud Pública y Materno-Infantil
    - José Javier Zamorano León (50985204-T)
    - María Elisa Calle (2703327-E)
- Dpto. Cirugía
    - Manel Giner Nogueras (37660995-J)
- Dpto. Fisiología
    - Vicente Lahera Juliá (51701558-L)
- Facultad de Informática, Dpto. Arquitectura de Computadores y Automática
    - José Ignacio Hidalgo Pérez (50443985-V)

2.- PERSONAL DE ADMINISTRACIÓN Y SERVICIOS (PAS)

- María Victoria Gómez García (02613604-E)
- María Joséfa Fernández López (50813467-G)
- Rocío Milagros Serrano Ruiz-Calderón (00821637-P)

3.- ALUMNOS

- De grado:
    - Marta Hernández Artiles (44739299-K)
    - Silvia Hernández Ramón (70069646-P)
- De formación post-grado (Máster oficial UCM
    - Gal·la Freixer Ballesteros (48040452-F)
    - Khaoula Zekri (Y3099058Y)

María Begoña Larrea Cruz (05416207-Y)

6. Desarrollo de las actividades

Tras un estudio inicial de los datos recibidos, se han realizado una serie de modificaciones orientadas a facilitar su uso en el entorno de trabajo y a su correcta interpretación para la resolución del problema. A continuación se exponen dichas modificaciones.

En primer lugar, se han renombrado las columnas, eliminando algunos elementos como tildes, puntos o espacios con la finalidad de facilitar su uso.

Uno de los puntos a destacar de estos cambios previos a las ejecuciones, es el tratamiento de los casos con valor 99 (Ns/Nc) en alguna de las columnas, principalmente en aquellas relacionadas con las notas de las distintas asignaturas y la calificación de acceso a la universidad. Dependiendo de la posición que ocupe el 99, tenemos dos situaciones:

- 99 en el valor a predecir: En esta situación, la fila con valor 99 en la columna a predecir se eliminará del conjunto de datos.
- 99 en otro valor: Dichos valores se reemplazarán con 0s o 7s (en caso de la calificación de acceso a la universidad).

De forma adicional, se han descartado aquellas variables que en su mayoría carecen de valores.

En esta primera toma de contacto se ha realizado un análisis de variables junto a una plantilla de clasificadores. Entre este conjunto se encuentran:

- Gradient Boosting
- Logistic Regression
- Random Forest
- Redes neuronales (MLPClassifier)
- KNeighborsClassifie
- GridSearchCV (svm)

1.- Algoritmo predictivo de éxito para la asignatura de Patología Médica Aplicada de 3er curso, basándose en las variables demográficas y académicas anteriormente citadas. En el análisis de variables académicas se utilizarán las calificaciones obtenidas en la asignatura de Fisiología de 1er curso que aporta aspectos fundamentales para la comprensión de la asignatura Patología Médica Aplicada. Resultados con GridSearch (SVM)

CAO 1: Nota superior a 5 ->1

```
Usuarios Totales-->126
Usuarios Entrenamiento-->94
Usuarios Test-->32
```

```
59 casos de 1
67 casos de 0
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Correct classification rate: 0.78125
              precision    recall  f1-score   support

         0.0       0.86      0.71      0.77        17
         1.0       0.72      0.87      0.79        15

    accuracy                           0.78        32
   macro avg       0.79      0.79      0.78        32
weighted avg       0.79      0.78      0.78        32
```

	tn	fp	fn	tp
0	12	5	2	13

CASO 2. Nota superior a 6 -> 1

```
Usuarios Totales-->126
Usuarios Entrenamiento-->94
Usuarios Test-->32


29 casos de 1
97 casos de 0
```

Resultados con RANDOM FOREST Regresor

```
Correct classification rate: 0.75
              precision    recall  f1-score   support

         0.0       0.77      0.91      0.83        22
         1.0       0.67      0.40      0.50        10

    accuracy                           0.75        32
   macro avg       0.72      0.65      0.67        32
weighted avg       0.74      0.75      0.73        32
```

	tn	fp	fn	tp
0	20	2	6	4

2.- Algoritmo predictivo de éxito para la asignatura de Epidemiología y Salud Pública de 4º curso, basándose en las variables demográficas y académicas anteriormente citadas. En el análisis de variables académicas se utilizarán las calificaciones obtenidas en la asignatura de Patología Médica Aplicada de 3er curso que aporta aspectos fundamentales para la comprensión de la asignatura Epidemiología y Salud Pública

Caso 1: dividiendo alumnos a partir de nota >7 = 1

```
Usuarios Totales-->66
Usuarios Entrenamiento-->49
Usuarios Test-->17
```

```
60 casos de 1
6 casos de 0
```

TODOS modelos identifican que la mayoría de la población no obtiene una nota igual o superior a 7, y por lo tanto clasificando a todos los alumnos como que no van a conseguir esta nota, se consigue una tasa de acierto muy elevada pero modelos no aplicables.

Caso dividiendo alumnos a partir de nota >8 = 1

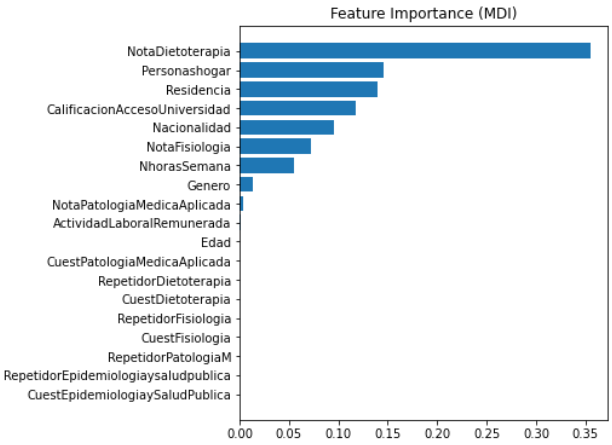
1. KNeighborsClassifier

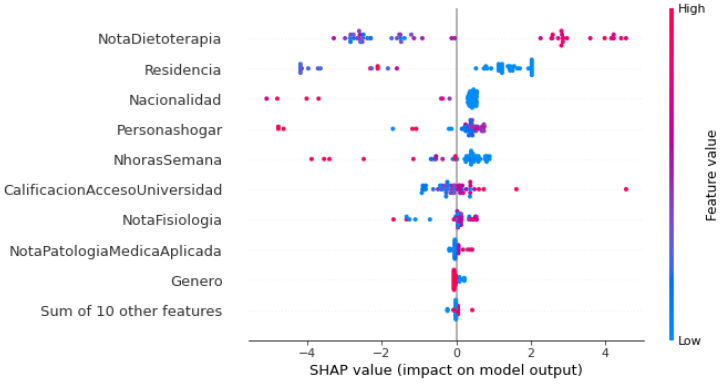
```
Correct classification rate: 0.7647058823529411
              precision    recall  f1-score   support

         0.0       0.67      1.00      0.80         8
         1.0       1.00      0.56      0.71         9

    accuracy                           0.76        17
   macro avg       0.83      0.78      0.76        17
weighted avg       0.84      0.76      0.75        17
```

	tn	fp	fn	tp
0	8	0	4	5





Nota: Se ha creado un enlace para uso del algoritmo http://147.96.81.95:8081/ (Ver anexo 2)

7. Anexos

Anexo I

CUESTIONARIO PARA ESTUDIANTES. CURSO 2019-20

1.- NOMBRE Y APELLIDOS:

_____

2.- EDAD: _____ años

3.- GÉNERO: Mujer ☐ Hombre ☐

4.- NACIONALIDAD:

_____

5.- GRADO EN EL QUE ESTÁ ACTUALMENTE MATRICULADO:

_____

6.- SEÑALE LA ASIGNATURA A LA QUE PERTENECE EL/LA PROFESOR/A QUE LE HA FACILITADO EL PRESENTE CUESTIONARIO:

- Fisiología: ☐ Es repetidor de la asignatura SI ☐ NO ☐
- Dietoterapia: ☐ Es repetidor de la asignatura SI ☐ NO ☐
- Patología Médica Aplicada: ☐ Es repetidor de la asignatura SI ☐ NO ☐
- Epidemiología y salud pública: ☐ Es repetidor de la asignatura SI ☐ NO ☐

5.- LUGAR DE RESIDENCIA (marque con una cruz la casilla correspondiente):

- Casa con padres ☐
- Casa con compañeros ☐
- Colegio mayor ☐
- Casa individual ☐
- Otros ☐. Especifíque _____

6.- SI NO VIVE EN UN COLEGIO MAYOR, INDIQUE ELNÚMERO DE PERSONAS CON LAS QUE CONVIVE EN EL LUGAR DE RESIDENCIA

7.- DESEMPEÑA ALGUNA ACTIVIDAD LABORAL REMUNERADA (marque con una cruz la casilla co-rrespondiente):

☐ SI Por favor, especifique el número de horas trabajadas/semana

☐ NO

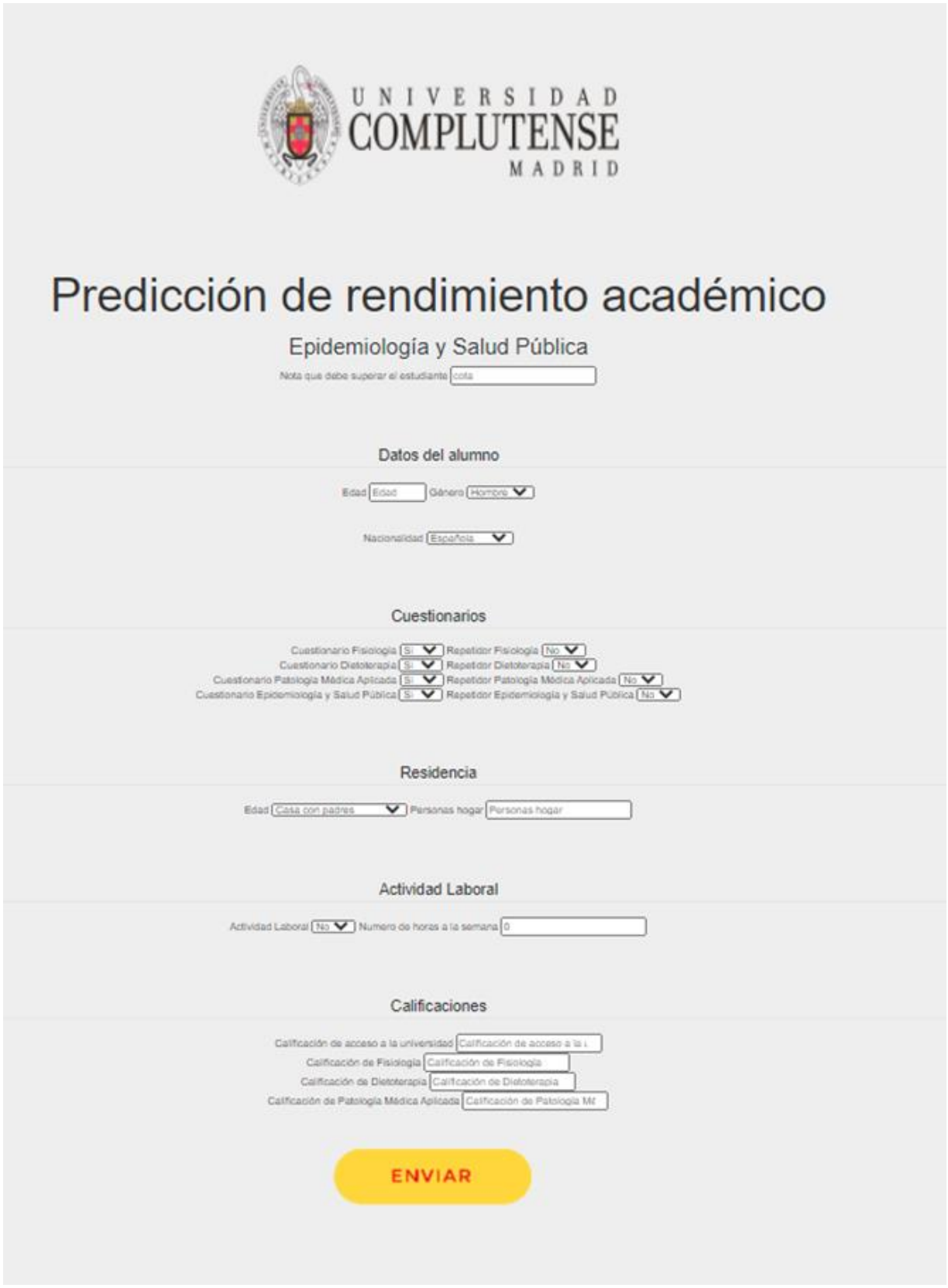
8.- CALIFICACIÓN DE ACCESO UNIVERSIDAD

9.-POR FAVOR, INDIQUE LA CALIFICACIÓN OBTENIDA EN LAS SIGUIENTES

ASIGNATURAS, EN EL CASO DE QUE HAYAN SIDO PREVIAMENTE CURSADAS:

- Fisiología:
- Dietoterapia:
- Patología Médica Aplicada:

Anexo 2


UNIVERSIDAD COMPLUTENSE MADRID
Predicción de rendimiento académico
Epidemiología y Salud Pública
Nota que debe superar el estudiante cota
Datos del alumno
Edad Edad Género Hombre
Nacionalidad Española
Cuestionarios
Cuestionario Fisiología Si Repetidor Fisiología No
Cuestionario Dietoterapia Si Repetidor Dietoterapia No
Cuestionario Patología Médica Aplicada Si Repetidor Patología Médica Aplicada No
Cuestionario Epidemiología y Salud Pública Si Repetidor Epidemiología y Salud Pública No
Residencia
Edad Casa con padres Personas hogar Personas hogar
Actividad Laboral
Actividad Laboral No Numero de horas a la semana 0
Calificaciones
Calificación de acceso a la universidad Calificación de acceso a la u
Calificación de Fisiología Calificación de Fisiología
Calificación de Dietoterapia Calificación de Dietoterapia
Calificación de Patología Médica Aplicada Calificación de Patología Mé
ENVIAR