Reporte de Rendimiento Académico Segundo Final IA

Francisco Gonzalez, Estudiante

I. INTRODUCCIÓN

En el presente trabajo se busca analizar datos obtenidos acerca del rendimiento de los alumnos en las distintas asignaturas de cursos básicos en las carreras de la Facultad de Ingeniería UNA (FIUNA), estos datos se encuentran en formato csv. Se pretende utilizar estos datos para construir modelos predictivos, para conseguir esto se hace uso de las habilidades obtenidas en la cátedra en lo que refiere a machine learning y sus aplicaciones. Estos modelos son muy importantes, pues con ellos se pueden identificar cuáles son las materias que más se le dificulta al alumno y para así poder mejorar el rendimiento del mismo en ellas, además ofrecen una perspectiva generalizada de cómo le irá en la carrera, en especial en el área de cursos básicos.
Para un mejor análisis de los mismos utilizar el lenguaje python implementado en jupyter notebook para la construcción de los modelos y gráficas a utilizar.

II. METODOLOGÍA

Los modelos lineales predicen un objetivo continuo basándose en relaciones lineales entre el objetivo y uno o más predictores. Los modelos lineales son relativamente simples y proporcionan una fórmula matemática fácil de interpretar para la puntuación.

Las propiedades de estos modelos se entienden bien y normalmente pueden crearse con bastante rapidez en comparación con otros tipos de modelos (como redes neuronales o árboles de decisión) del mismo conjunto de datos.

El análisis de la regresión lineal se utiliza para predecir el valor de una variable según el valor de otra. La variable que desea predecir se denomina variable dependiente. La variable que se está utilizando para predecir el valor de la otra variable se denomina variable independiente.

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores, este modelo es aplicado en este trabajo para poder clasificar el tiempo de permanencia en años, también es una función importante en las redes neuronales que son utilizadas para clasificación.

Una red neuronal, es un algoritmo que consiste en simular el comportamiento de un cerebro biológico mediante miles de neuronas artificiales interconectadas que se almacenan en filas llamadas capas, formando miles de conexiones. Existen de varios tipos, pero el utilizado en este trabajo es el del perceptrón multicapa.

El modelo de metodología utilizado es el CRISP-DM, el cual establece fundamentalmente un modelo de proceso que es de naturaleza jerárquica, comenzando por el nivel fase y luego bajando a tareas genéricas, tareas específicas e instancias de proceso.

III.DATOS

Los datos a ser utilizados están en formato csv, pero utilizando las librerías de python se pueden visualizar. Luego de construir una tabla con los datos obtenidos se procede a la limpieza del mismo, para este caso se necesitan datos de cursos básicos de los estudiantes, estás fueron anonimizadas previamente. Una vez obtenido el primer dataframe(df) con todos los datos desde primer semestre hasta cuarto semestre de todas las carreras, se procede a trabajar solo con el plan semestral del 2013 pues este tiene mayor cantidad de muestras. A partir de este df se obtienen los ids de los alumnos para poder obtener más datos sobre los mismos, implementando una función que reciba el df de cursos básicos de cada plan disponible y la cantidad de materias que está posee, entonces la función pasa a buscar por ids si estos ya concluyeron cb, el promedio general del primer semestre, la cantidad de materias recursadas del primer semestre, y las notas por cada materia correspondiente al semestre ya citado, la función retorna un nuevo df con todos estos datos para cada carrera. Luego se realizan la unión de las mismas para el df final y la extracción de qué datos se utilizan como características en los modelos propuestos, también se añadió a la carrera como característica categórica.

IV. ARQUITECTURA

A partir de la tabla con los datos filtrados de cursos básicos se construye la matriz de correlaciones entre las variables para determinar cuáles características tiene mayor correlación,
luego de extraer estas características se construyen los modelos lineales para obtener los mejores resultados posibles y así poder predecir con mayor exactitud el rendimiento de los alumnos.

Para ello se utilizan los modelos de regresión lineal y logística de la librería sklearn, así como el modelo secuencial y las capas de la librería keras. Con

un conjunto de 689 muestras, entrada con 14 características y uno de salida. Los modelos encargados de la clasificación son el logístico y el de red neuronal perceptrón multicapa.

Para evaluar estos modelos las métricas de evaluación utilizadas son: precisión, f1 score ya que son las que mejor identifican los errores en problemas de clasificación.

V. RESULTADOS

se construyeron 3 modelos:

- modelo 1(regresión lineal): 'tardoCB ~ prom_al1 + prom_c1 + prom_dt + prom_f1 + prom_ga + prom_gd + prom_1er_S + recursadas_s1 + C(carrera)'.
- Modelo 2(regresión logística): 'tardoCB ~ prom_al1 + prom_c1 + prom_dt + prom_f1 + prom_ga + prom_gd + prom_1er_S + recursadas_s1 + C(carrera)'.
- Modelo 3(red neuronal): 'tardoCB ~ prom_al1 + prom_c1 + prom_dt + prom_f1 + prom_ga + prom_gd + prom_1er_S + recursadas_s1 + C(carrera)'

todos tratando de predecir la variable ‘tardoCB’, que es el tiempo en años que tarda un estudiante en culminar cursos básicos. utilizando las siguientes variables:

- recursadas_s1: total de materias que recurso en el primer semestre.
- prom_c1: calificación promedio en calculo 1.
- prom_al1: calificación promedio en álgebra lineal 1.
- prom_f1: calificación promedio en física 1.
- prom_ga: calificación promedio en geometria analitica.

- prom_gd: calificación promedio en geometría descriptiva.
- prom_dt: calificación promedio en dibujo técnico.
- prom_1er_S: calificación promedio en el primer semestre.
- C(carrera): variable categórica correspondiente a la carrera.

luego se obtuvo la siguiente tabla de desempeños de los modelos:

1- modelo 1:

	precision	recall	f1-score	support
1.0	0.0000	0.0000	0.0000	0
2.0	0.8056	0.5179	0.6304	56
3.0	0.4559	0.7561	0.5688	41
4.0	0.4074	0.4583	0.4314	24
5.0	0.6000	0.2727	0.3750	11
6.0	0.0000	0.0000	0.0000	5
7.0	0.0000	0.0000	0.0000	1
accuracy			0.5362	138
macro avg	0.3241	0.2864	0.2865	138
weighted avg	0.5810	0.5362	0.5297	138

2- modelo 2:

	precision	recall	f1-score	support
2	0.6618	0.8036	0.7258	56
3	0.3478	0.3902	0.3678	41
4	0.2941	0.2083	0.2439	24
5	0.6667	0.1818	0.2857	11
6	0.2500	0.2000	0.2222	5
7	0.0000	0.0000	0.0000	1
accuracy			0.5000	138
macro avg	0.3701	0.2973	0.3076	138
weighted avg	0.4852	0.5000	0.4771	138

3- modelo 3:

	precision	recall	f1-score	support
0	0.5526	0.8077	0.6563	52
1	0.3250	0.3421	0.3333	38
2	0.5000	0.1154	0.1875	26
3	0.5000	0.3636	0.4211	22
accuracy			0.4783	138
macro avg	0.4694	0.4072	0.3995	138
weighted avg	0.4716	0.4783	0.4415	138

El modelo de regresión lineal presenta una precisión del 53 %, seguido del modelo de regresión logística con una precisión de 50 % y el modelo de redes neuronales con una precisión del 48 %.
La precisión de este último aumenta bastante si se establecen más épocas pero esto produce overfitting y ya no es posible tener una buena predicción, es posible mejorar el modelo añadiendo más capas con más neuronas o cambiando las funciones de activación de las capas ocultas.

Lo que también se puede observar es que materias como calculo 1, algebra lineal 1, geometria analitica y física 1, poseen correlaciones negativas relativamente altas con el tiempo que se tarda en culminar cb, Es decir, si el alumno tiene puntuaciones altas en estas materias y recurso menos veces, su estadía en cb es menor. Por lo tanto mejorando la calidad de enseñanza en estas materias se tendrán mejoras importantes en el desempeño general.