Universidad Nacional de Colombia

Facultad de Ingeniería
Departamento de Sistemas e Industrial

Introducción a los sistemas inteligentes
Proyecto final

Jose Luis Moreno Hernandez
David Alexander Zambrano Bohorquez

Bogotá D.C.
2022

1. *Business understanding.*

A partir de el conjunto de datos “Students performance in exams” de las notas de estudiantes en el cual se muestran las calificaciones de varios estudiantes en tres pruebas distintas (matemáticas, lectura y escritura) a partir de diferentes datos de contexto del estudiante tales como género, etnia, nivel de educación de los padres,tipo de almuerzo, entre otros. A partir de este se propone crear un modelo de aprendizaje de máquina con el cual se pueda predecir cómo le irá a un estudiante en las tres pruebas teniendo en cuenta estas características.

Se van a tener en cuenta los siguientes objetivos para evaluar:

- Definir una tendencia a partir de las principales características que permitirían obtener a un estudiante una buena calificación en las diferentes pruebas
- Permitir predecir el posible resultado de las pruebas de un estudiante a partir de ciertas características definidas a partir de su preparación o contexto previo a la presentación de las pruebas.

Se considerará al estudio como un éxito si

- La precisión en la predicción del modelo es mayor a 70% con los datos de prueba que son proporcionados
- A partir de datos aleatorios que definan el caso de un estudiante al azar, se siga la tendencia definida por las predicciones y los datos definidos para el modelo

Datos: Los datos que se van a utilizar son datos ficticios que están hechos especialmente para ciencia de datos generados por Royce Kimmons y compilados dentro de kaggle. Estos se componen de 1000 datos con diferente información requerida para su evaluación.

Riesgos: Es posible que debido a la naturaleza de los datos, estos no tengan una tendencia que se represente dentro de la vida real, de manera que algunas de las conclusiones que se puedan obtener no sean aplicables totalmente a un caso similar de la vida real. Para mitigar ese riesgo se podrían visualizar datos reales y compararlos con los propuestos para ver qué tan buena sería la evaluación con estos.

Como objetivos dentro de la minería de datos se tiene:

- Utilización de datos otorgados por Royce Kimmons para crear un modelo que permita predecir los resultados de diferentes tipos de exámenes a partir de datos externos y de contexto de un usuario.

2. *Data understanding.*

Cada registro del conjunto de datos consta de cinco datos cualitativos de cada estudiante y tres resultados obtenidos en las pruebas. Entre los datos cualitativos se encuentran:

- Género: Ya sea masculino o femenino.
- Raza o etnia: Hay cinco grupos distintos, los cuales van numerados desde la A hasta la E.
- Nivel de educación de los padres: Se tiene en cuenta el nivel de educación de los padres esperando que influya en el resultado de los hijos. Dichos niveles de educación pueden ser licenciatura (bachelor’s degree), educación superior incompleta (some college), maestría (master’s degree), título asociado

(associate's degree), secundaria (high school) y secundaria incompleta (some high school).

- Almuerzo: Tipo de almuerzo que reciben los estudiantes. Puede ser estándar o gratuito.
- Curso de preparación para el test: Si completó o no algún curso de preparación para las pruebas.

Los resultados obtenidos en las pruebas de matemáticas, escritura y lectura toman valores entre cero y cien donde un mayor valor significa que existió un mejor desempeño

El total de datos que se va a utilizar va a ser de 1000, debido a la facilidad y rapidez de procesamiento de estos y la información que puede brindar es suficiente para la meta que se ha propuesto. Los tipos de datos son en su mayoría de tipo texto que definen el tipo de una característica entre algunas posibles elecciones. Todos estos datos tienen diferentes definiciones y no son conflictivos dentro de las entradas del modelo.

Por otro lado se considera que la calidad de los datos es buena, ya que no existe ningún problema de datos perdidos, errores de datos o mediciones debido a la naturaleza artificial del conjunto de datos utilizado.

3. Data preparation.

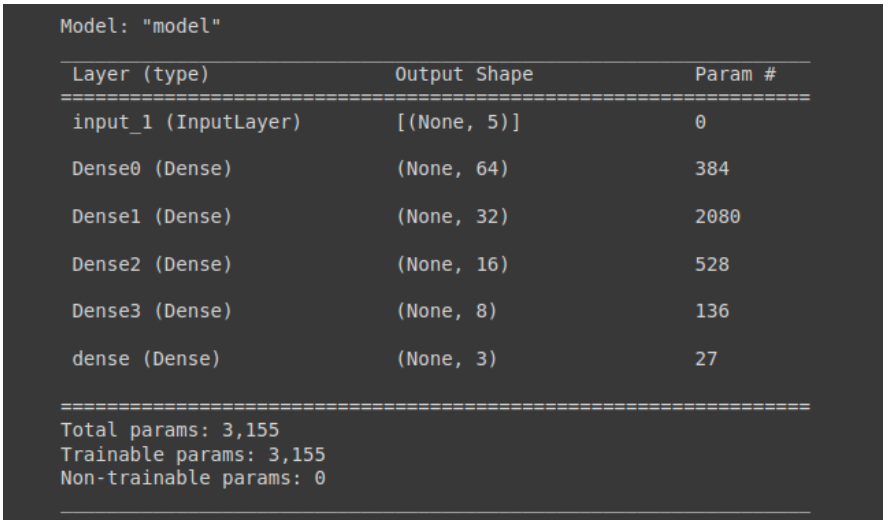
La preparación de los datos fue llevada a cabo en dos etapas, cada una enfocada en un modelo en específico. La primera etapa se enfocó en preparar los datos para alimentar una red neuronal con la finalidad de poder predecir la nota que cada estudiante iba a obtener en las pruebas mientras que la segunda etapa se enfocó en preparar los datos para alimentar clasificadores de Naive Bayes para poder predecir si un estudiante pasa o no cada prueba usando un umbral de 60 puntos.

1. Primera etapa: La primera etapa reemplaza los valores cualitativos por valores cuantitativos asignándoles un entero distinto a cada valor único de cada columna a manera de enumeración, posteriormente se separa el conjunto de datos en dos subconjuntos distintos, un subconjunto "X" con las características de los estudiantes y un subconjunto o target "y" con los resultados obtenidos en las pruebas por cada estudiante. Una vez teniendo ambos subconjuntos, ambos se separan en subconjuntos para entrenamiento y validación.
2. Segunda etapa: Como se busca hacer una clasificación binaria en esta etapa, se necesita modificar el valor numérico de cada prueba de cada estudiante de tal manera que se sepa si este pasó o no el examen. Para lo anterior se parte del conjunto de datos de la primera etapa en el que los elementos cualitativos se reemplazaron por valores cuantitativos y se procede a reemplazar los resultados de los tres exámenes de tal manera que si este es mayor o igual a 60, el estudiante pasó y se reemplaza el valor de su nota por un 1, en el caso contrario se reemplaza por un 0. De igual manera se separa los datos en dos subconjuntos "X" e "y" y posteriormente en datos de entrenamiento y de prueba.

4. Modeling.

Como se mencionó anteriormente, en este proyectos se utilizaron dos modelos distintos, una red neuronal y un clasificador naive bayes.

1. Red neuronal: La red neuronal construida consta de una entrada de 5 parámetros equivalentes a las 5 cualidades de cada estudiante y 3 neuronas de salida equivalentes a los resultados en las tres pruebas de cada estudiante. Como función de activación se utilizó leaky relu en cada capa, la función de utilizada fue mse y el optimizador por defecto fue adam. A cada capa oculta se le añadió inicialización de pesos con media 1 y desviación estándar 0. El número de capas y de neurona no se consideró tan relevante ya que variar dichos parámetros no generó variaciones considerables en el resultado del modelo. A continuación se muestra el resumen del modelo utilizado:

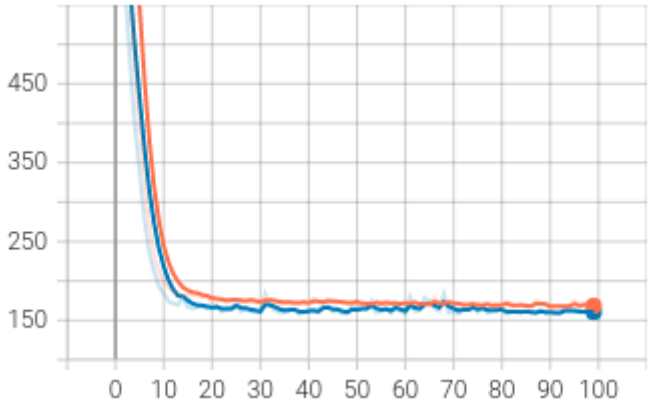


```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 5)]               0

 Dense0 (Dense)              (None, 64)                384

 Dense1 (Dense)              (None, 32)                2080

 Dense2 (Dense)              (None, 16)                528

 Dense3 (Dense)              (None, 8)                 136

 dense (Dense)               (None, 3)                 27

=================================================================
Total params: 3,155
Trainable params: 3,155
Non-trainable params: 0
_________________________________________________________________
```

2. Clasificador naive Bayes: A diferencia del modelo anterior, en este caso se requirieron tres clasificadores en total, uno para cada prueba ya que se consideró una manera adecuada de producir múltiples salidas, por lo tanto cada clasificador se entrenó usando el mismo subconjunto de "X" (X_train) pero diferentes valores como target, más específicamente una columna del target original para cada clasificador.

5. *Evaluation.*

Una vez creada la red neuronal, se procedió a entrenarla con los datos de entrenamiento. Desde su entrenamiento se pudo apreciar que la realización del ejercicio de predecir los resultados de los estudiantes en las pruebas no iba a ser muy efectiva analizando el error tanto de entrenamiento como de validación.

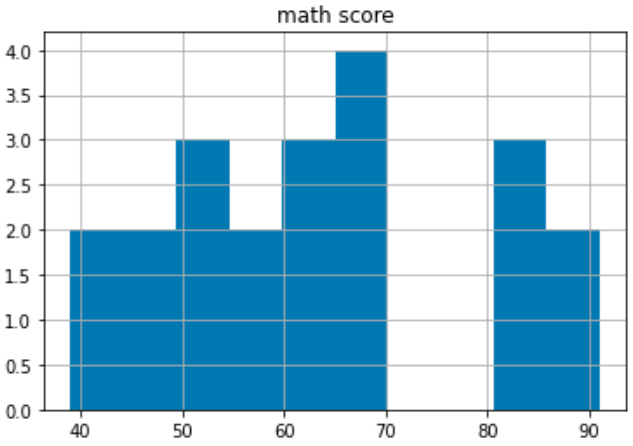


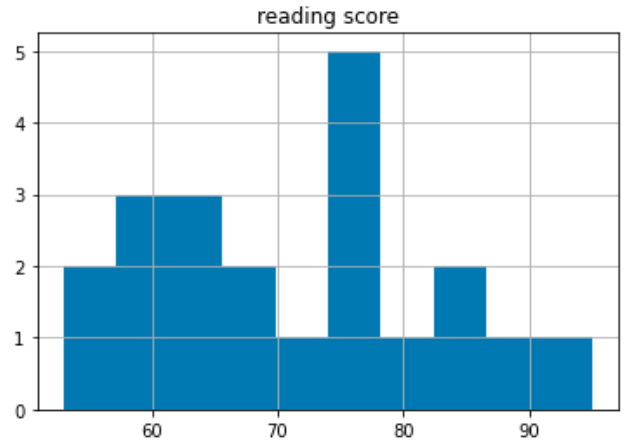
Como se puede observar en la gráfica anterior generada por tensorboard, tanto el error de entrenamiento (naranja) como el error de validación (azul) decrecen de manera considerable en las primeras iteraciones y convergen a un valor superior a 150, teniendo en cuenta que la pérdida se calcula mediante el error cuadrático medio, se obtuvo valores bastante grandes los cuales indican que las predicciones pueden no ser bastante acertadas.
Al generar predicciones con los datos de prueba para el modelo, se obtuvo el siguiente resultado:

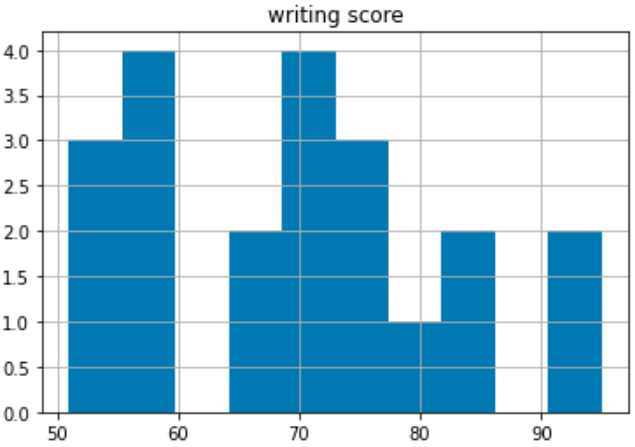
```
Diferencia promedio:  [13.20757358 12.83900308 12.17353687]
Math accuracy = 2.500%
Reading accuracy = 4.500%
Writing accuracy = 3.500%
```

La red neuronal solo acertó el 2.5% de las predicciones en las pruebas de matemáticas, el 4.5% en las pruebas de lectura y el 3.5% en las pruebas de escritura y la diferencia promedio entre cada predicción y su valor verdadero estuvo por encima de 12 puntos.

Lo anterior indica que el planteamiento de redes neuronales para predecir el resultado de cada estudiante no es buena idea, sin embargo, la culpa no es del todo de las redes neuronales, sino de la manera planteada para predecir el desempeño de cada estudiante. La razón radica en que pueden haber distintos estudiantes con exactamente las mismas características pero con puntajes bastante diferentes, por ejemplo, para el caso en el que el estudiante es de género femenino, su raza es del grupo C, sus padres tienen título de asociado, recibe almuerzo estándar y no recibió ningún curso de preparación para el examen, hay 21 registros en total, a continuación se muestran los histogramas de los puntajes obtenidos en cada prueba por los 21 estudiantes.







En los histogramas se puede apreciar que para los tres exámenes distintos, los 21 estudiantes obtuvieron notas en un rango de valores bastante grande, abarcando desde valores cercanos a 50 hasta valores superiores a 90 por lo que tratar de predecir la nota exacta en este conjunto de datos es una tarea bastante compleja, se necesitarían más cualidades para poder caracterizar mejor a los estudiantes y generar predicciones más exactas.

Debido a la poca exactitud que tendría la red neuronal al predecir el valor final de los exámenes, se ha hecho necesario cambiar la finalidad del ejercicio y enfocarlo a la predicción de pasar o no cada uno de los exámenes.

Para la resolución de esta tarea se utilizaron los clasificadores Naive Bayes mencionados anteriormente. En total se usó uno para cada examen y se entrenaron con los resultados de sus respectivos exámenes.
Una vez entrenados los clasificadores se calculó el accuracy de cada uno con los datos de prueba:

```
math classifier accuracy:  0.67
reading classifier accuracy:  0.735
writing classifier accuracy:  0.765
```

Se tiene que cada clasificador obtuvo un buen accuracy superior al planteado como objetivo del proyecto el cual es un 70% a excepción del correspondiente a la prueba de matemáticas, sin embargo vamos a analizar los resultados más a fondo. A continuación se presentan las matrices de confusión de los tres clasificadores.

```
math test
                      Valor    Predicho
---------  -------  -------  ----------
           -        Reprobó  Aprobó
Valor      Reprobó  19       50
Verdadero  Aprobó   16       115

reading test
                      Valor    Predicho
---------  -------  -------  ----------
           -        Reprobó  Aprobó
Valor      Reprobó  13       37
Verdadero  Aprobó   22       128

writing test
                      Valor    Predicho
---------  -------  -------  ----------
           -        Reprobó  Aprobó
Valor      Reprobó  16       37
Verdadero  Aprobó   19       128
```

En la matriz de confusión, la clase “Aprobó” se considera la clase positiva y la clase “Reprobó” se considera la clase negativa, con ello se puede proceder a calcular el recall, la precisión y la especificidad de cada clasificador. Para el recall se tiene lo siguiente:

```
Math recall  0.8778625954198473
Reading recall  0.8533333333333334
Writing recall  0.8707482993197279
```

Lo cual indica que el clasificador es bueno reconociendo los casos en los que el estudiante aprobó la materia. Para la precisión se tiene lo siguiente:

```
Math precition  0.696969696969697
Reading precition  0.7757575757575758
Writing precition  0.7757575757575758
```

Los valores de precisión obtenidos indican que los clasificadores tienen buena habilidad para no clasificar estudiantes como aprobados cuando reprobaron. Y por último, los valores obtenidos de especificidad fueron los siguientes:

```
Math especifity  0.2753623188405797
Reading especifity  0.26
Writing especifity  0.3018867924528302
```

En todos los clasificadores fue bastante bajo el valor de la especificidad, lo cual indica que el clasificador tiende a ser malo reconociendo los casos en los que los estudiantes reprueban.

A simple vista pareciera como que no tiene sentido que la precisión sea grande mientras que la especificidad es pequeña, pero en este caso si lo tiene ya que la cantidad de estudiantes aprobados es mayor que la de estudiantes reprobados, por lo tanto los falsos positivos no van a ser muy significativos en la precisión generando valores grandes. Si la cantidad de estudiantes reprobados fuera mayor, la precisión tendería a decrecer.
De las métricas se puede concluir que los clasificadores son muy malos para reconocer los casos en los que los estudiantes reprueban, esto se debe a que como tal en el conjunto de datos hay una mayor cantidad de estudiantes aprobados que reprobados, lo cual se presta para que si un clasificador clasificase a todos los estudiantes como aprobados, obtendría un buen accuracy. A esto se le suma el hecho de como se vio anteriormente, hay bastantes estudiantes con las mismas características, lo que hará que cuando llegue un nuevo estudiante con esas características, se clasifique con la clase que más aparece en ese grupo de estudiantes, la cual normalmente es 1 debido a su dominancia en el conjunto de datos, haciendo que sea difícil clasificar estudiantes como reprobados.

Conclusiones:

Debido a la naturaleza del conjunto de datos, es complicado generar predicciones para saber cómo se desempeñará un estudiante debido a la variación en los resultados de las pruebas de un grupo de estudiantes con las mismas características.
Se puede concluir que en este tipo de situaciones, es complicado y se incurre en errores bastante grandes al tratar de generar un ajuste de curvas, por lo cual se debe segmentar los datos en clases en las cuales caigan gran parte del número de datos de tal manera que no se pierda bastante información y se puedan generar predicciones que sean bastante dicientes.

Link video explicativo del poster:

https://www.youtube.com/watch?v=oslbMopLjxY