PROYECTO 1 (data cleaning + MLlib)

ALEJANDRO AYALA GIL
ESTEBAN CARDONA GIL
JUAN CAMILO GOMEZ MUÑOZ
JULIAN PAREDES C
TANIA C. OBANDO SUÁREZ

PROCESAMIENTO DE GRANDES VOLÚMENES DE DATOS
CALI, 4 DE OCTUBRE DE 2020

1) **¿Qué se va a predecir?**

Se va a predecir si un estudiante perteneciente a alguno de los colegios mostrados en el Dataset va a aprobar o reprobar el curso de portugués con base, en calificaciones de los estudiantes, características demográficas, sociales y relacionadas con la escuela).

Este problema es un problema de clasificación binaria, en donde las salidas de los modelos representarán dos clases: aprobado (si la salida es 1) y reprobado (si la salida es 0).

2) **Descripción inicial del conjunto de datos.**

El Dataset nos proporciona un conjunto de datos con el desempeño de varios estudiantes en la materia de portugués(por) en dos escuelas portuguesas.

Inicialmente el conjunto de datos contaba con:

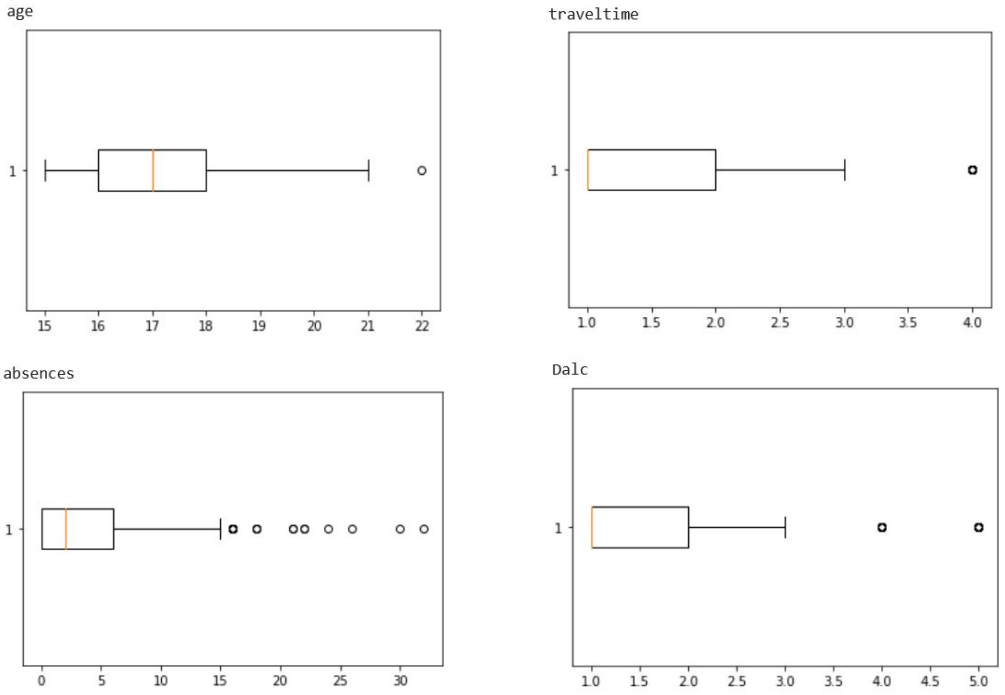
Número de filas o instancias: 649
Número columnas o atributos: 33
Número de datos nulos: 0

Los 33 atributos corresponden respectivamente a la tabla presentada como **anexo 2**. Donde se describen los atributos, los tipos de datos, los valores posibles y las descripciones de los atributos.

3) **Resumen de las transformaciones realizadas**

- Al crear el dataframe en spark todos los atributos quedaron almacenados como tipo string.
- Después se llevaron todas las variables categóricas a variables numéricas sin embargo el tipo de estos seguía siendo String.
- Por lo cual se hizo un casteo para dejar todos los datos como Int para facilitar la manipulación de los datos.
- En el caso de las variables G1,G2 y G3, que representan la nota de portugues obtenida en cada periodo, se estableció un umbral del 60% de la nota máxima para categorizar la materia como aprobada o reprobada. Dicho resultado en este caso corresponde a una nota de 12, por lo cual los estudiantes con una nota estrictamente inferior a esta, se consideran como reprobados, mientras que los que tienen notas iguales a mayores a 12 se consideran como aprobados. Lo anterior con el fin de poder realizar la predicción explicada al inicio del documento.
- Se realizaron diagramas de caja y bigotes para analizar de manera gráfica la dispersión de los datos y encontrar datos atípicos también se analizó de manera numérica datos mínimos, máximos, media, desviación estándar y la moda. Para esta fase de eliminación de registros con base en la información proveída por estas métricas, establecimos que íbamos a borrar hasta un máximo del 10% de los datos, por lo cual los atributos modificados aquí

fueron age, traveltime, absences y Dalc. A continuación se presentan los diagramas de cajas y bigotes obtenidos para estas columnas:



Aquí, eliminamos todos un registro cuyo valor para age era de 22, 16 registros cuyo valor de traveltime correspondía a 4, 11 registros donde el valor de absences era mayor a 17, y 17 registros donde Dalc era igual a 5. Al realizar estos cambios, el tamaño del conjunto de datos obtenido fue de 608 registros.

- ➢ Se analizaron las correlaciones de todas las variables, pero destacan las siguientes:

Correlaciones positivamente fuertes	**Correlaciones positivamente moderadas**	**Correlaciones negativamente moderadas**
G1 y G2 □ 0.778	Medu y Fedu □ 0.649	school y address □ -0.361
G1 y G3 □ 0.738	Walc y Dalc □0.576	traveltime y address □ -0.340
G2 y G3 □ 0.885		failures y G1 □ -0.320
		failures y G2 □ -0.339
		failures y G3 □ -0.367

NOTA: En el **anexo 2** ubicado al final del documento se evidencian todas las correlaciones de manera gráfica a través de un mapa de calor.

4) Descripción final del conjunto de datos.

Después de realizarse las anteriormente mencionadas transformaciones, obtuvimos dos conjuntos de datos para el entrenamiento de los modelos, los cuales tienen las siguientes características:

Dataset final 1
Número de filas o instancias: 548
Número columnas o atributos: 30
Número de datos nulos: 0

Dataset final 2
Número de filas o instancias: 548
Número columnas o atributos: 32
Número de datos nulos: 0

- Para el dataset 1, se eliminaron los atributos G1 y G2, debido a su alta correlación con el atributo de salida G3. Es por esto que cuenta con 31 columnas.
- Para el conjunto de datos 2 se eliminó el atributo G2 debido a la alta correlación con el atributo de salida G3. Es por esto que cuenta con 32 columnas.

La decisión de tener dos conjuntos de datos para el entrenamiento y validación de los modelos fue realizada por una sugerencia que contenía la descripción del conjunto de datos original donde establecía desde un principio la alta correlación entre G1, G2 y G3, al mismo tiempo que decían que para la estimación de G3 era más “difícil” de obtener sin G1 y G2.

5) Comparación de las tres técnicas de aprendizaje automático.

En este proyecto, se usaron las técnicas de clasificación de regresión logística, Random forest, y Máquina de vectores de soporte. Para la evaluación de los modelos, se usaron las métricas de accuracy, precision, recall, f1 score, y el área bajo la curva ROC. A continuación se presenta el desempeño de los mejores modelos obtenidos en cada técnica con el respectivo conjunto de datos usado.

	Accuracy	**Precision**	**Recall**	**F1 score**	**Área bajo la curva ROC**
Regresión logística	0.70950	0.76404	0.68687	0.71032	0.71218
Random	0.72067	0.85507	0.59596	0.71778	0.73548

forest					
Máquina de vectores de soporte	0.77095	0.78431	0.80808	0.77047	0.76654

Desempeño de los modelos usando el dataset 1

	Accuracy	**Precision**	**Recall**	**F1 score**	**Área bajo la curva ROC**
Regresión logística	0.86667	0.875	0.875	0.86667	0.86607
Random forest	0.88484	0.87096	0.92045	0.88451	0.88231
Máquina de vectores de soporte	0.88484	0.86316	0.93181	0.88431	0.8814

Desempeño de los modelos usando el dataset 2

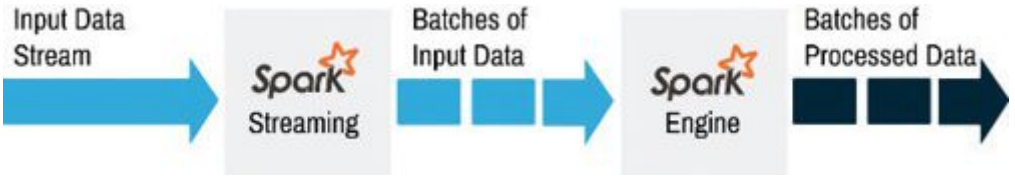
Con base a estos resultados, podemos obtener las siguientes conclusiones:

- ➢ Para el dataset 1 y al usar regresión logística o random forest, existe una ligera tendencia a clasificar mejor la clase reprobados. Sin embargo, este comportamiento no se ve reflejado en el vector de máquinas de soporte
- ➢ Para el dataset 2, existe una ligera tendencia a clasificar mejor la clase aprobados. Sin embargo, podemos apreciar que la tasa de falsos positivos y falsos negativos en el mejor modelo de regresión logística es igual.
- ➢ Los resultados obtenidos con el modelo de la máquina de vectores de soporte y el modelo de random forest para el dataset 2 son demasiado similares.
- ➢ En términos generales, el modelo que mejor se adapta a ambos conjuntos de datos y al problema que queremos resolver es el de máquinas de vectores de soporte.
- ➢ En términos generales, el dataset 2 nos permite obtener una mejor solución al problema. Cabe mencionar que la diferencia entre la solución obtenida con el dataset 1 y el dataset 2 es insignificante. Posiblemente esto se deba a que el dataset 2 modela de una forma más afín la naturaleza del problema que queremos resolver que la del dataset 1.

6) ¿Cómo vamos a utilizar este conjunto de datos para las entregas posteriores?

Teniendo en cuenta los dos proyectos que se desarrollarán en el curso. A partir del dataset que escogimos y las técnicas de machine learning implementadas, se buscará extender las predicciones y análisis realizados previamente mediante el uso de streaming (Spark Streaming) y grafos (Spark GraphX). En cuanto a eso, se ha realizado un acercamiento en el que se detalla el funcionamiento de las distintas herramientas aplicadas a el dataset escogido.

- Spark Streaming es una extensión de la API core de Spark que ofrece procesamiento de datos en streaming de manera escalable, alto rendimiento y tolerancia a fallos. Los datos pueden ser tomados de diferentes fuentes como Apache Kafka, Apache Flume, RabbitMQ, Amazon Kinesis, ZeroMQ o sockets TCP.



A diferencia de este primer proyecto en el cual se realiza el análisis y predicciones de los datos tomando todo el dataset, en Spark Streaming se reciben streams de datos en vivo los cuales se dividen en batches o lotes, que son procesados por el motor de Spark para generar un stream de salida o DStream, el cual es una secuencia de RDDs.

Para procesar dichos datos se llega a utilizar algoritmos complejos de machine learning expresados como funciones de alto nivel como lo son map, reduce, join y window. Una vez procesados, los datos son enviados a archivos en file systems o para dashboards en tiempo real.

En nuestra aplicación en concreto dado que nuestro dataset es estático, se buscará reducir en porciones pequeñas el dataset, limitando la cantidad de datos que recibirá el batch. De tal manera, que se pueda verificar si el funcionamiento en streaming sigue siendo válido con respecto al presentado en este proyecto. Para esto, se continuaría haciendo uso de los factores filtrados en el dataset, probando de manera experimental para el siguiente proyecto que cantidad mínima sería suficiente para que se siga cumpliendo el modelo.

- Spark GraphX es un motor para el análisis de grafos, construido sobre el núcleo de Spark, el cual utiliza los RDDs creados previamente con el fin de permitir a los usuarios crear, transformar y obtener conclusiones a partir de un grafo estructurado.

  GraphX se utiliza para explorar la naturaleza o las propiedades topológicas de un grafo. Además, esta herramienta permite ejecutar algoritmos que puedan encontrar subgrafos conectados y desconectados, realizar ciertos conteos o incluso realizar cálculos para hallar el camino más corto de un punto a otro en un grafo.

  Es por esto que para el último proyecto se propone utilizar GraphX junto con Neo4j, dado que los datos procesados en Spark con Mlib se puede llegar a conectar con

Neo4j para poder visualizar y conformar los grafos que se pueden producir con estos datos, para después poder analizar las propiedades topológicas del grafo haciendo uso de GraphX.

Con respecto al diseño del grafo para la tercera entrega del proyecto, se planea crear el grafo a partir de similitudes que tenga un estudiante con otro, estas similitudes serán por medio de valores que sean relevantes al momento de comparar una persona con otra, estos serían los parámetros que quedarían después de hacer la correlación con el objetivo de que se pueda aportar información distinta a la de otros datos . Una vez creado el grafo se podrá trabajar con el algoritmo de vecinos comunes para poder predecir si un estudiante ganará la materia o perderá con respecto a los datos obtenidos de sus vecinos.

7) Cibergrafía

- **How to Handle Imbalanced Classes in Machine Learning. Recuperado de la web en la siguiente Url: https://elitedatascience.com/imbalanced-classes**
- **What is Considered to Be a "Strong" Correlation?. Recuperado de la web en la siguiente Url: https://www.statology.org/what-is-a-strong-correlation/**
- **Classification and regression. Recuperado de la web en la siguiente Url: https://spark.apache.org/docs/2.2.0/ml-classification-regression.html**
- **Student Performance Data Set. Recuperado de la web en la siguiente Url: https://archive.ics.uci.edu/ml/datasets/Student+Performance#**
- **Un vistazo a Apache Spark Streaming. Recuperado de la web en la siguiente Url: https://sg.com.mx/revista/50/un-vistazo-apache-spark-streaming**
- **Spark Streaming. Recuperado de la web en la siguiente Url: https://bigdatadummy.com/2017/05/12/spark-streaming/**
- **Spark Streaming (procesamiento por lotes y tiempo real). Recuperado de la web en la siguiente Url: https://www.diegocalvo.es/spark-streaming/**
- **Neo4j and Apache Spark. Recuperado de la web en la siguiente Url: https://neo4j.com/developer/apache-spark/**
- **Graph data processing with Neo4j and Apache Spark. Recuperado de la web en la siguiente Url: https://medium.com/neo4j/graph-data-processing-30451b5b576f**

8) Anexos

a) Anexo 1

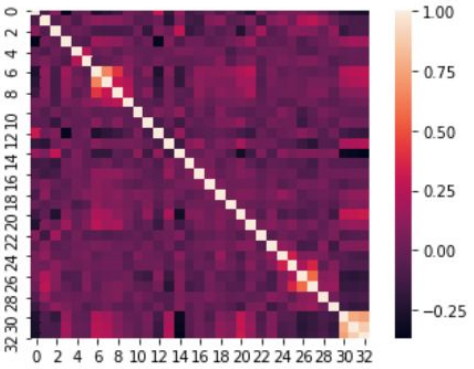
Número del	Atributo	Tipo	Posibles valores	Descripción

atributo				
0	school	binario	GP,MS	Escuela a la que pertenece el estudiante: Gabriel Pereira (GP) o Mousinho da Silveira (MS).
1	sex	binario	F,M	Sexo del estudiante: femenino (F) o masculino (M).
2	age	numérico	[15,22]	Edad del estudiante.
3	address	binario	U,R	Sector donde se ubica la casa del estudiante: urbano (U) o rural (R).
4	famsize	binario	LE3,GT3	Tamaño del núcleo familiar del estudiantes: menor igual a 3 (LE3) o mayor a 3 (GT3).
5	Pstatus	binario	T,A	Estado de cohabitación de los padres: juntos (T) o separados (A).
6	Medu	numérico	[0,4]	Grado de educación de la madre del estudiante: Ninguno (0), Educación primaria hasta 4to grado (1), 5to-9no grado (2), educación secundaria (3) o educación superior (4).
7	Fedu	numérico	[0,4]	Grado de educación del padre del estudiante: Ninguno (0), Educación primaria hasta 4to grado (1), 5to-9no grado (2), educación secundaria (3) o educación superior (4).
8	Mjob	nominal	teacher,health,services, at_home, other	Trabajo de la madre del estudiante: profesora (teacher), relacionado con las ramas de la salud (health), servicios civiles como policía o algún cargo administrativo (services), ama de casa (at_home) u otro (other).
9	Fjob	nominal	teacher,health, services, at_home, other	Trabajo del padre del estudiante: profesor (teacher), relacionado con las ramas de la salud (health), servicios civiles como policía o algún cargo administrativo (services), amo de casa (at_home) u otro (other).

10	reason	nominal	home, reputation, course, other	Razón para elegir la escuela: cercana a la casa (home), reputación del colegio (reputation), preferencia de cursos (course) u otro (other).
11	guardian	nominal	mother, father, other	Tutor o acudiente del estudiante: madre (mother), padre (father) u otro (other)
12	traveltime	numérico	[1,4]	Tiempo de viaje de la casa a la escuela: menor a 15 minutos (1), entre 15 y 30 minutos (2), entre 30 y 60 minutos (3) o mayor de una hora (4).
13	studytime	numérico	[1,4]	Horas de estudio semanales: menor a 2 horas (1), entre 2 y 5 horas (2), entre 5 y 10 horas (3) o superior a 10 horas (4).
14	failures	numérico	[1,4]	Número de clases previamente reprobadas por el estudiantes: n,si $1 \leq n \leq 3$, de lo contrario 4.
15	schoolsup	binario	yes,no	Apoyo educativo adicional: si (yes) o no (no).
16	famsup	binario	yes,no	Apoyo educativo familiar: si (yes) o no (no).
17	paid	binario	yes,no	Clases extra de portugues pagadas: si (yes) o no (no).
18	activities	binario	yes,no	El estudiante practica actividades extracurriculares: si (yes) o no (no).
19	nursery	binario	yes,no	El estudiante asistió a la enfermería: si (yes) o no (no).
20	higher	binario	yes,no	El estudiante quiere tener una educación superior: si (yes) o no (no).
21	internet	binario	yes,no	El estudiante cuenta con acceso a internet: si (yes) o no (no).
22	romantic	binario	yes,no	El estudiante tiene una relación romántica: si (yes) o no (no).
23	famrel	numérico	[1,5]	Calidad de la relación familiar del estudiante en una escala del 1 al 5, siendo 1 muy mala y 5 excelente.

24	freetime	numérico	[1,5]	Tiempo libre del estudiante después de la escuela en una escala del 1 al 5, siendo 1 muy bajo y 5 muy alto.
25	goout	numérico	[1,5]	Salidas con amigos del estudiante en una escala del 1 al 5, siendo 1 muy bajo y 5 muy alto.
26	Dalc	numérico	[1,5]	Consumo de alcohol del estudiante durante la semana en una escala del 1 al 5, siendo 1 muy bajo y 5 muy alto.
27	Walc	numérico	[1,5]	Consumo de alcohol del estudiante durante los fines de semana en una escala del 1 al 5, siendo 1 muy bajo y 5 muy alto.
28	health	numérico	[1,5]	Estado de salud actual del estudiante en una escala del 1 al 5, siendo 1 muy malo y 5 muy bueno.
29	absences	numérico	[0,93]	Número de ausencias escolares del estudiante durante el año lectivo cursado.
30	G1	numérico	[0,20]	Nota final del primer periodo en la asignatura de portuges.
31	G2	numérico	[0,20]	Nota final del segundo periodo en la asignatura de portuges.
32	G3	numérico	[0,20]	Nota final del tercer periodo en la asignatura de portuges.

b) Anexo 2



Mapa de calor de las correlación entre todos los atributos del dataframe. En los ejes se muestran el número de la columna relacionada con el atributo.