GA & MISK ACADEMY
DATA SCIENCE COURSE

FINAL PROJECT

April,2020

OVERVIEW

1. Project Background and Description

In this Machine Learning project, I will build a model to predict student performance of tow secondary Portuguese schools. Tow datasets are provided regarding the performance in two subjects: Mathematics (mat) and Portuguese language (por). The data was collected using school reports and questionnaires.

2. Dataset Source

Dataset available at :
https://archive.ics.uci.edu/ml/datasets/student+performance

3. Data Shape

The data consist of 1044 rows and 33 columns

(1044,33)

4. Data Dictionary

	Variable	Description
1	school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2	sex	student's sex (binary: 'F' - female or 'M' - male)

	Variable	Description
3	age	student's age (numeric: from 15 to 22)
4	address	student's home address type (binary: 'U' - urban or 'R' - rural)
5	famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6	Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7	Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3-secondary education or 4 -higher education)
8	Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2(5th to 9th grade), 3-secondary education or 4-higher education)
9	Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10	Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11	reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	guardian	student's guardian (nominal: 'mother', 'father' or 'other')
13	traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

	Variable	Description
14	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
16	schoolsup	extra educational support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
20	nursery	attended nursery school (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)
24	famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25	freetime	free time after school (numeric: from 1 - very low to 5 - very high)
26	goout	going out with friends (numeric: from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

	Variable	Description
28	Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29	health	current health status (numeric: from 1 - very bad to 5 - very good)
30	absences	number of school absences (numeric: from 0 to 93)
31	G1	first period grade (numeric: from 0 to 20)
31	G2	second period grade (numeric: from 0 to 20)
32	G3	final grade (numeric: from 0 to 20, output target)

5. Implementation Plan

Step 1: Create a new blank notebook

Step 2: Explore and visualize the data

Step 3: Select the significant predictors for the model (features selection).

Step 4: Manipulating and transforming data- if needed (features engineering)

Step 5: Splitting the data

Step 6: Train the model

Step 7: Test the model

Step 8: Evaluate the model

Step 9: Draw conclusions

6. Assumptions of the model

1. Linearity: Y and X must have an approximately linear relationship.
2. Independence: Errors (residuals) εi and εj must be independent of one another for any i≠j
3. Normality: The errors (residuals) follow a Normal distribution with mean 0.
4. Equality of Variances (Homoscedasticity of errors): The errors (residuals) should have a roughly consistent pattern, regardless of the value of X. (There should be no discernable relationship between X and the residuals.)