# Student Performance Prediction and Analysis

Sarthak Maini - 2020576
Jay Saraf - 2020438
Aryan Vohra - 2020557
Hardik Patel - 2020507

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

# Motivation

- Data availability is a challenge for startup to medium size institutes or schools.

- We have used a small dataset by choice to simulate a real mid-sized university/college environment.

- We are focusing our study on the Demographic factors such as illness, financial problems etc that are faced by the students.

# Motivation

- Once all these factors are recognised, the model can be further used to develop a Student Performance Prediction software which can help educational entities in understanding the shortcomings of educational processes.

- Early precaution can be taken in order to improve student academic performance

- Also it can be used to select the right student for a particular task.

# Literature Review

***Prediction of Student’s performance by modelling small dataset size (by Lubna Mahmoud Abu Zohair)***

Research focussed on small to mid-sized universities, which have very less number of student records for analysis

Dataset consists of only 50 graduated students in one master’s program.Attributes include Student’s ID,Age,Bachelor degree Name,etc.

After Data Cleaning only 38 records were remaining.Feature Encoding,Missing Value imputation, Normalization and Feature Selection were performed.

Multiple ML classification algorithms used - MLP-ANN, Naive Bayes,SVM,KNN,LDA used.

SVM gave the best accuracy of 68.4%

## Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor (by H. Al-Shehri et al. )

This research presented two prediction models for student's performance prediction in final examination.

Dataset consists of 395 data samples taken from University of Minho in Portugal.

Label Encoding,Correlation analysis and Feature selection were performed

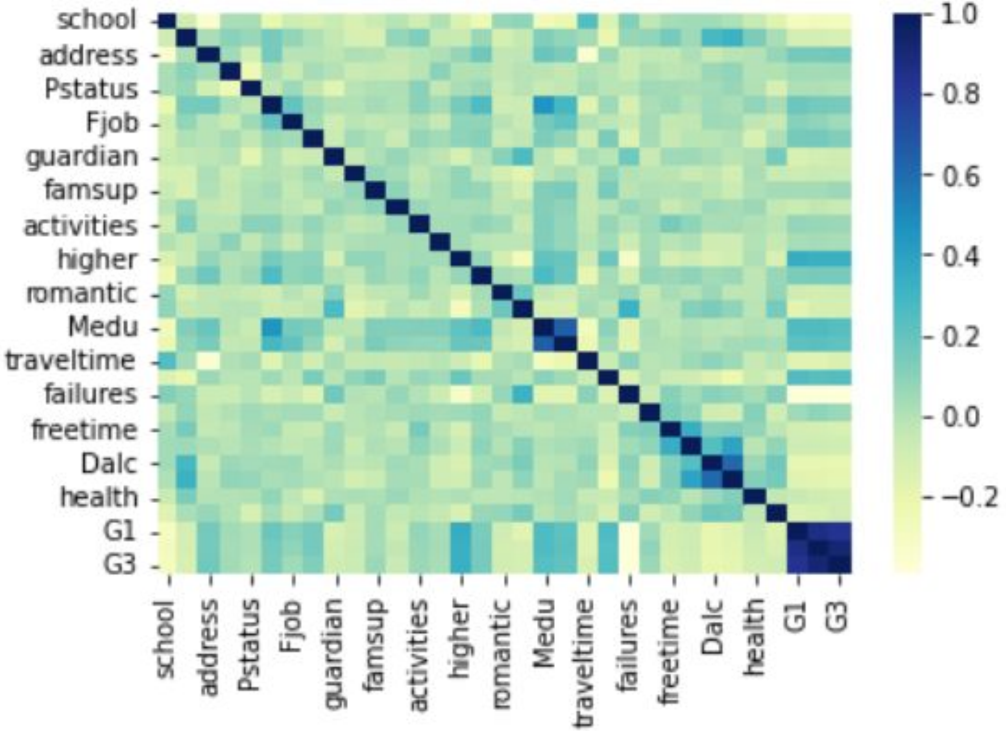
Used a combination of KNN and SVM to predict performance in the final exam.

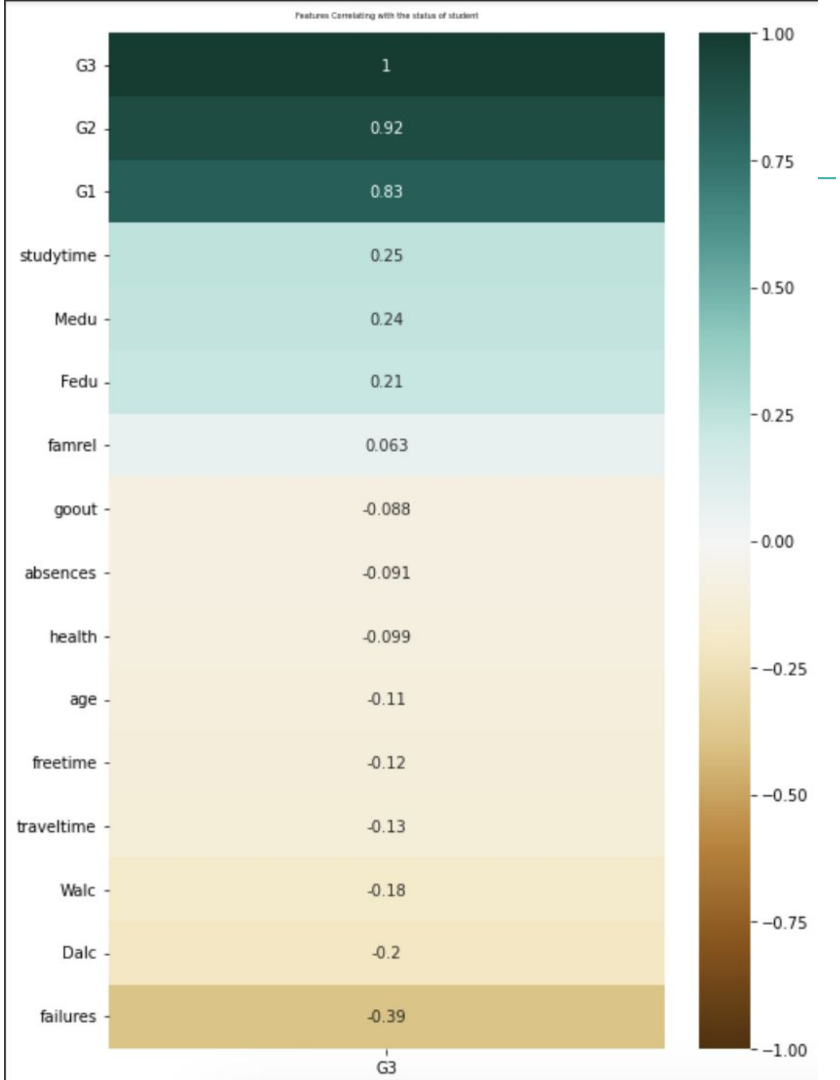
Used different partition ratios and 10-cross validation with KNN and SVM .

Ideal results obtained with 90:10 percentage split for SVM and KNN.
SVM slightly exceeds KNN due to higher correlation coefficient value of 0.96.

# Data Description

- The data was taken from the University of Minho, Portugal.
- It is an analysis of two Portuguese schools centrics towards Portuguese language in the year 2005.
- 4 features are nominal, 13 are binary and 16 are numeric.
- The dataset was formed from the school reports and questionnaires . The grading system is based on the scale of 20.
- G3 being the final grade which a student gets , is our target variable.


school
address
Pstatus
Fjob
guardian
famsup
activities
higher
romantic
Medu
traveltime
failures
freetime
Dalc
health
G1
G3
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

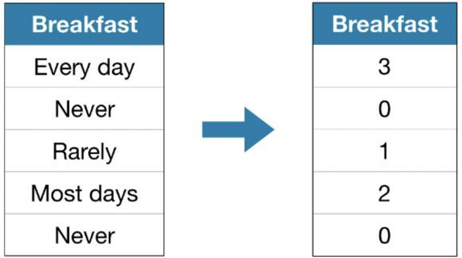

Features Correlating with the status of student
G3 1
G2 0.92
G1 0.83
studytime 0.25
Medu 0.24
Fedu 0.21
famrel 0.063
goout -0.088
absences -0.091
health -0.099
age -0.11
freetime -0.12
traveltime -0.13
Walc -0.18
Dalc -0.2
failures -0.39
G3
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

# Data Preprocessing

Label Encoding - to convert categorical data to numeric form

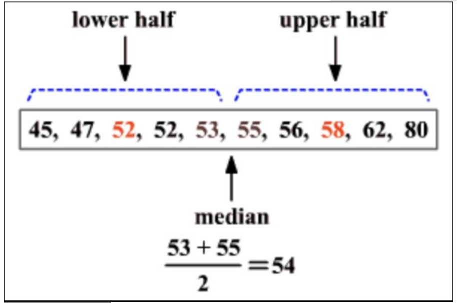
Robust Scaling - based on Interquartile Range or IQR to handle the outliers in data

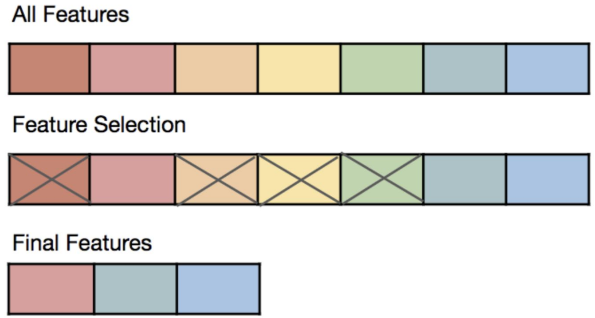
Feature Selection - based on correlation between different attributes and Recursive Feature Elimination



| Breakfast |
|---|
| Every day |
| Never |
| Rarely |
| Most days |
| Never |

| Breakfast |
|---|
| 3 |
| 0 |
| 1 |
| 2 |
| 0 |





# Methodology

- After preprocessing the data, we applied several  models namely Linear Regression, Decision Tree ,Random Forest,MLP ,SVM and KNN on the data.
- Linear regression models the data assuming a linear relationship occurs between the input attributes and the output attribute.
- Decision Tree models the data in a hierarchical tree-like structure.It divides the data based on simple rules ,hence it is very explanatory and interpretable in nature.
- MLP is a fully connected class of feed forward neural networks having the ability to model even the most complex non-linear relations.
- Random Forest is an ensemble-based approach of combining the prediction of several Decision Trees
- K-means clustering is an unsupervised clustering algorithm that takes unlabelled data and tries to find possible clusters in the data
- SVM tries to maximize the marginal distance between the support vectors

# Results

| | | **With Grades** | **Without Grades** |
|---|---|---|---|
| **Linear Regression** | 4-level classification | 0.76 | 0.21 |
| | Regression | 0.89 | 0.17 |
| **Decision Tree** | 4-level classification | 0.91 | 0.72 |
| | Regression | 0.5 | 0.15 |
| **Random Forest** | 4-level classification | **0.92** | **0.91** |
| **MLP** | 4-level classification | 0.82 | 0.61 |
| **SVM** | 4-level classification | 0.87 | 0.69 |
| **K-Means Clustering** | 4-level classification | 0.71 | 0.66 |

# Analysis

Linear regression consistently works better on the dataset when all 20 classes are used giving a performance of 0.9

Decision tree works better on the dataset when 4 classes are used (by class reduction) giving an accuracy of 0.86

Linear regression is less robust to data changes (G1 and G2 removed) (0.17)

Decision trees are more robust to data changes (G1 and G2 removed) (o.72)

Generally not taking into account G1 and G2 (high correlation with G3) results in poorer models ,

Hence it can be derived that models learn best when G1 and G2 / Highly correlated data are present

Class reduction is much more beneficial for Decision trees as compared to Linear regression.

Conclusions:-

For better interpretability we use decision trees and for best performance we make use for the linear regression model. When we have more data it is better to use Linear regression and with less data decision trees give better results.

- We observed that out of all the classifiers , random forests had the best accuracy (0.92). This is because random forests are internally based on decision trees and since decision trees have a really good performance, multiple weak decision trees serve as really good predictors for the test set.
- We can use max voting , taking into account all the predictions of the different decision trees.
- Other models gave worse predictions generally but some models like those SVM still gave a comparable performance (0.87) in cases when all the classes were present.

# Individual Contribution

| *Team Member* | *Tasks* |
|---|---|
| Sarthak | Decision Tree, Data Collection, SVM, Random Forest and literature review |
| Aryan | Data Pre-processing, Decision Tree, SVM, K means clustering and Result Analysis |
| Jay | EDA, Linear Regression, MLP,K means clustering |
| Hardik | Linear Regression, Random Forest, MLP, Feature Selection and Motivation |

Everyone had an equal contribution in the project.

# References

- Feature Selection
- IQR
- Label encoding
- H. Al-Shehri et al., "Student performance prediction using Support Vector Machine and K-Nearest Neighbor," 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017, pp. 1-4, doi: 10.1109/CCECE.2017.7946847.
- Abu Zohair, L. M. (2019). Prediction of Student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education, 16(1), 1-18. https://doi.org/10.1186/s41239-019-0160-3