Inspiring Excellence

CSE422 : Artificial Intelligence
Project Report
Project Idea : Student Performance Prediction

Group No : 01**, Lab Section :** 05**, Spring 2024**	
ID	**Name**
20101197	Abir Ahammed Bhuiyan
18201085	Allama Bakhtiyar Nafis

Introduction

In this project, we aim to analyze and predict student performance in secondary education using machine learning techniques. Our objective is to build predictive machine learning models that can accurately predict or classify students' final year results i.e. pass or fail based on attributes such as father's education, mother's education, internet etc.

Dataset Description & Visualization

The dataset consists of student grades, demographic, social, and school-related features collected from two Portuguese schools. The target variable, G3, represents the final year grade, while G1 and G2 correspond to grades in the 1st and 2nd periods, respectively. Moreover, the dataset actually contains two csv files based on subjects viz Mathematics and Portuguese. For our work we have only used the Mathematics csv file.

The attributes of the dataset are given below,

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)
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)
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)

19. **activities** - extra-curricular activities (binary: yes or no)
20. **nursery** - attended nursery school (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)
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)
32. **G2** - second period grade (numeric: from 0 to 20)
33. **G3** - final grade (numeric: from 0 to 20, output target)

Dataset link: https://archive.ics.uci.edu/dataset/320/student+performance
Dataset Source: The dataset was collected through school reports and questionnaires by Cortez and Silva in 2008.

Data Pre-processing

Feature Engineering: In the dataset we have dropped ‘G1’ and ‘G2’ columns and renamed the ‘G3’ columns as ‘passed’. This ‘passed’ column will be used as our target column. Moreover, we have changed the discrete values of the ‘passed’ column to ‘yes’ or ‘no’ based on the grade/number each student has received.

Label Encoding: In the ‘passed’ column we have employed label encoding and transformed yes or no value to 1 and 0. Moreover, in the dataset for all the binary categorical values we have used label encoding for representing them in number.

Frequency Encoding: Some categorical columns that have more than two values we have use frequency encoding. The reason behind using frequency encoding is one-hot encoding will introduce ‘curse of dimensionality’, to avoid ‘curse of dimensionality’ we have used frequency encoding.

Scaling: Scaling was used for faster convergence of the model training.

Train-Test Split: The whole dataset was divided into train and test dataframe. This was done for evaluating the model performance on unseen data.

ML Models Implementation

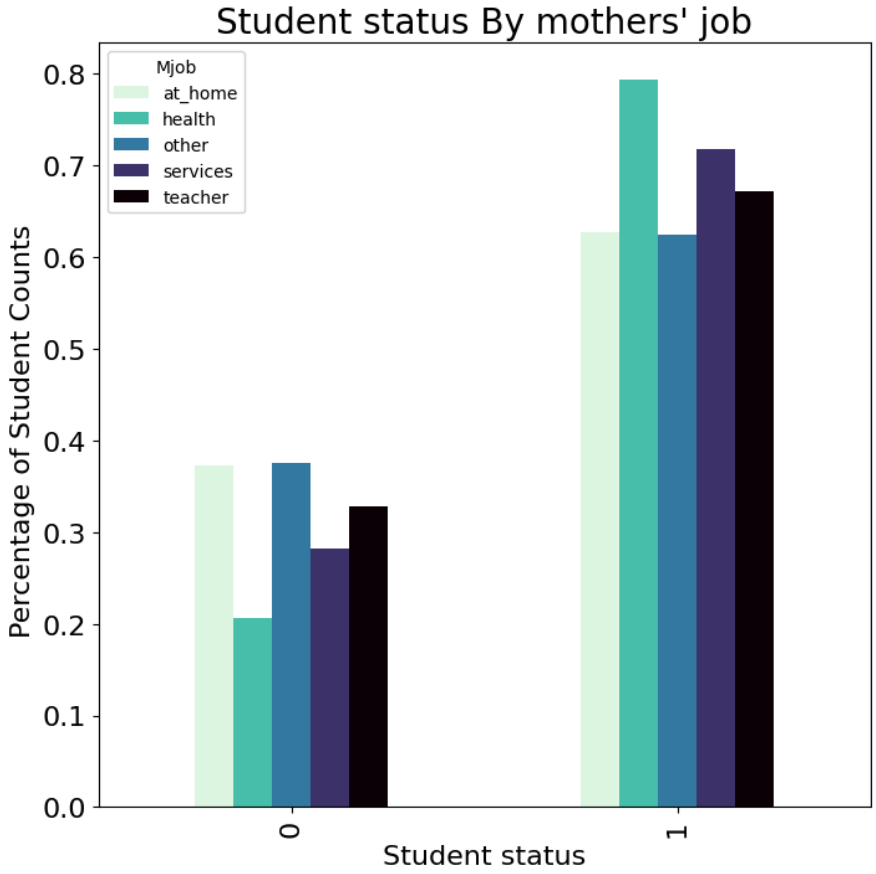
- **KNN (k-Nearest Neighbors Classifier) :** KNN works on the clustering principle, also it places a new datapoint based on the nearest neighbors. In this work we are trying to classify only two values, 'yes' or 'no'. So, clustering the data points and predicting based on distance between the nearest neighbor should work fine.

- **DTC (Decision Tree Classifier) :** In decision tree we deal with a binary tree and in each level decision tree tries to take decision by condition that were learned from the features. In our classification the labels are also binary and because of that reason it was a good choice to use DTC. Because, answering yes no question should be easier for DTC.

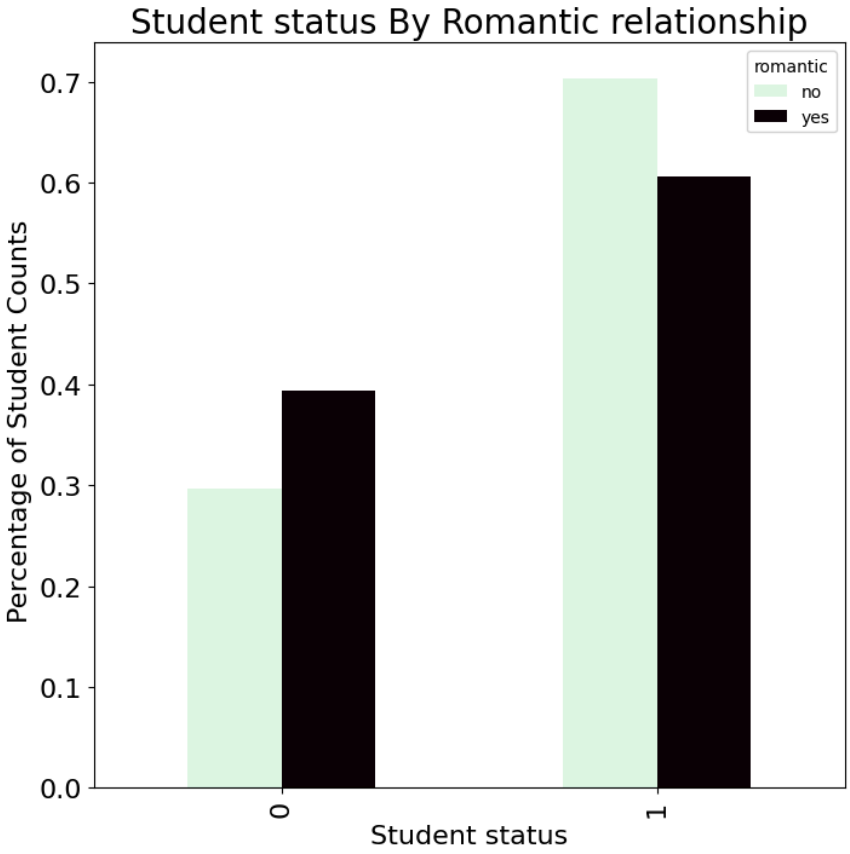
- **SVC (Support Vector Classifier) :** In SVM each object we want to classify is represented as a point in an n-dimensional space the coordinates of this point are its attributes or features. SVM classifies the points by drawing a hyperplane between each class. In our case we are dealing with a binary classification problem so using SVC is an easy and straightforward task.

Results & Analysis

By analyzing the students' performance based on mothers job, we have found that if a student's mother works in the health sector then that student is likely to pass.



Also, if a student is not in a romantic relationship that means the student is more likely to pass.



Among all the models SVC performed better with 72% accuracy on the test dataframe.

Before Scaling:

Model	Accuracy	Precision	Recall	F1-Score
KNN	65%	65%	65%	65%
DTC	61%	65%	61%	62%
SVC	72%	52%	72%	60%

After Scaling:

Model	Accuracy	Precision	Recall	F1-Score
KNN	71%	68%	71%	69%
DTC	62%	65%	62%	63%
SVC	72%	67%	72%	66%

Conclusion:

After the experiment of three models we got better performance in SVC. The other two models KNN, DTC gave suboptimal results, maybe this is due to their underlying algorithms. Moreover, from the analysis it is also apparent that scaling has good impact on the the models performance.

We have learned some new ML techniques and got a deeper understanding of ML algorithms.