CSE/ECE 343: Machine Learning Final Project Report
Title: Student Performance Prediction and Analysis

Sarthak Maini sarthak20576@iiitd.ac.in
Aryan Vohra aryan20557@iiitd.ac.in
Jay Saraf jay20438@iiitd.ac.in
Hardik Patel hardik20507@iiitd.ac.in

1. Abstract

It is extremely important to develop a Student Performance prediction software. It helps educational entities and institutes in understanding the shortcomings of the current educational process and suggests improvements in our current system. We have also purposely used a smaller-to-medium dataset for training out models so as to simulate a real mid-sized university/college environment. In our experimental techniques we have used two major well-known strategies for Training on the dataset, these include decision trees and linear regression. The results show the effectiveness of using such techniques for predicting Student performance. In the end we have mentioned appropriate conclusions drawn from the input which suggests what further improvements can be made by the student in their studying techniques to improve their performance. Empirical studies outcome indicated that Decision Tree in 4-level classification scenario is the most effective in predicting student performance.

2. Introduction

It is important to develop Student Performance prediction software. It helps educational entities and institutes in understanding the shortcomings of the current educational process and suggests improvements in our current system. It can also be used to take early precautions in order to improve students' performance in academic endeavors. Another application of such a system can be to select the right student for a particular task.

We also need to figure out the reasons that lead to poor performance of the students. Factors like extracurricular activities, illnesses, financial problems, and family problems have a major effect on students' performance.
We used a dataset, which was collected from two Portuguese schools in the subject of Portuguese. We first used simple linear regression to predict the students' performance in the range 0 - 20 and then used decision trees for easier data analysis so as to conclude the reason for the students' performance. Here we have developed 8 models.

We used both the techniques in permutations of: -
1) using dataset with grades and without grades
2) regression (0-20) and 4-level classification

Furthermore, we also reviewed two research papers which predict student performance so as to understand techniques used by researchers in similar settings.

3. Literature Survey

One of the challenges that occurred in the task of predicting students' performance is the availability of a large dataset. Abu Zohair, L.M. explored the utilization possibility of small students' data-set size in educational domains [1].

Recently several researchers have started using Educational Data Mining and it has depicted effective performance in identifying the most appropriate prediction method. Ferda Ünal employed Decision Tree, Random Forest, and Naive Bayes to predict the performance of students on the five-level grading version and

binary version of a Portuguese school dataset [2]. Cortez and A. Silva tested four DM methods, i.e. Decision Trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines (SVM) with Three different DM goals (i.e. binary/5-level classification and regression) [3].
Osmanbegovic and Suljic used Naive Bayes, Multilayer Perceptron (MLP) and C4.5 implementation of Decision Tree for prediction on the data collected from first semester students at the University of Tuzla [4].
T. M. Christian and M. Ayub used Naïve Bayes (NB) Tree to predict student's performance during their study [5].
H. Al-Shehri et al. used a combination of SVM and KNN to estimate students' performance in the final exam [6].

From the literature survey above, we observed that very few researchers have explored the possibility of combining two or more models for student performance prediction. Thereby we decided to explore different combinations of models along with testing the performance with standalone models.

4. Dataset

(a) Dataset Description

The dataset is taken from the University of Minho, Portugal. It is an analysis of two Portuguese schools' students in Portuguese language subject in the year 2005.
There are in total 33 features among which 4 features are nominal, 13 features are binary and 16 features are numeric. The dataset was formed from school reports and questionnaires. The grading system is based on a scale of 20, resulting in a score ranging from 0-20, and G3 being the final grade that the student achieves.

Attributes	Description
Sex	Binary (Male or Female)
Age	Ranging from 15 to 22
School	Binary (Gabriel Pereira or Mousinho da Silveira)
Address	Binary
Famsize	Family Size - Binary (> 3 or <3)
Pstatus	Parents cohabitation status
Mjob	5 classes
Fjob	5 classes
Reason	4 classes
Guardian	Ternary
Activities	Binary
Romantic	Binary
Schoolup	Binary
Paid	Binary
Higher	Binary

Internet	binary
Nursery	Binary

(b) Data pre-processing

Before the data can be passed onto the model for training and testing it need to be preprocessed, so as to achieve the best results.
The preprocessing tasks included data cleaning, missing data imputation, label encoding, data normalization, outlier removal and feature selection, so as to make the data ready for analysis.
The data was checked for duplicates and null values. Since the data contained integers as well as categorical values, the categorical values were changed to integer values using label encoding, so as to make the classification algorithms work on the categorical data. The data was then checked for outliers and these were handled using a technique called Interquartile Range or IQR. The data samples lying outside the range defined by the quartiles were removed. We performed IQR using the RobustScaler method of sklearn library which takes care of outliers as well as normalizes the data at the same time. The next preprocessing step is feature selection which is done using RFE (recursive feature elimination) method of sklearn library which generates a ranking of features by recursively considering a smaller set of features.

After preprocessing the data is now ready to be passed onto the model for training and evaluation.

5. Methodology

Data Mining is an emerging field especially in the field of education, particularly called educational data mining. Taking insights from the literature survey done, we try out different models namely Linear Regression and Decision Trees to predict student performance in the final exam.

Linear Regression:
It is a technique to model the relationship between data in a linear way. In other words, Linear Regression tries to fit a line to the data assuming a linear relationship between the input variables and the output variable.
We started out which linear regression before exploring any other model which can give a non-linear boundary so as to check the best performance that can be achieved with a simple linear model.

Decision Trees:
It is an algorithm which models the data in a hierarchical tree-like manner. It splits the data based upon certain rules which can easily be translated to the form of IF ELSE statements which are very easy to understand, making it extremely explainable and interpretable model. Also, it is a non-parametric method meaning that it does take any assumption about data distribution.
Unlike Linear Regression, Decision Tree has the ability to model the non-linearity in data, it is able to capture very complex non-linear decision boundaries. But on the downside, this makes the decision tree prone to overfitting.

Multilayer Perceptron (MLP) :

The multilayer perceptron is a fully connected class of feed forward neural networks. It consists of input, output and hidden layers. Since every neuron in each layer is having its own set of weights, therefore MLP can model even the most complex non-linear relationship in data. It even inherently does feature extraction, which had to be done separately in case of traditional ML algorithms.

Random Forest:

It is an ensemble-based approach of combining the prediction of several Decision Trees. Since individual decision trees typically tend to overfit the data, thereby an ensemble is used so that the model does not overfit and gives better generalization. The prediction which is given by most decision trees is considered as the final prediction corresponding to the Random Forest Classifier.

K-means clustering:

It is an unsupervised clustering algorithm that takes unlabelled data and tries to find possible clusters in the data. It computes centroids, updates the data association and then repeats until the optimal centroid is found that has the maximum inter cluster distance and minimum intra cluster distance.

Support Vector Classifier (SVC):

It is very similar to linear decision boundaries but tries to maximize the marginal distance between the support vectors (points closest to the margin of separation). Hence Support vector classifiers try to find an Optimum Separation Hyperplane with the highest margin to its support vectors. Additionally, if a non-linearly separable data is presented then a kernel function to project the data into higher dimensions is used. Plotting of linear decision boundary with max margin on this new n-dimensional data can be performed in order to validate the results.

6. Results and Analysis

		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

Linear regression consistently gives better results for most cases. It gave the third highest accuracy of 89% on the raw data that was pre-processed.
Decision tree model was more robust to changes in the dataset, even after removing grades (G1 and G2 which had a high correlation with G3 Target variable) the decision tree model still had a decent prediction accuracy (in the case of 4-level classification) and was able to predict student performance with 72% percent accuracy.
In general, "without grades" cases had the lowest accuracy considering the fact that G1 and G2 which were highly correlated with G3 were removed from the dataset.

We could also see that class reduction technique was really helpful in the case of decision tree, reason being that predicting 20 classes with 0-1 loss is too strict for decision trees. However, the same could not be seen in the case of linear regression, as linear regression loses information when class size is reduced and it is not able to make an accurate prediction.

Random Forest Classifier performs the best in both scenarios i.e., the 4-level classification scenario as well as the regression scenario. It gave an accuracy of 92% on the dataset with grades whereas gave an accuracy of 91% on the dataset without the grade-component.

Support vector machine (SVM) performs well, with an accuracy of 87%, when the grade data is present in the dataset. However its performance decreases drastically after the grade-data is removed.

MLP also performed comparable to SVM when the grading data was available, with an accuracy of 82%. However, in all the other cases, the performance of the models was not up to the mark.

7. Conclusion

We concluded that Random Forest is the best when compared on the basis of performance. But since bootstrapping reduces the interpretability of the random forest thereby if interpretability is our utmost priority, then standalone decision trees are better. Also, they are more robust to changes in the actual dataset. Even in cases of little to low availability of data they can perform fairly well.

On the other hand, Linear regressor, MLP and SVM are also quite accurate models when the grade information is there in the dataset as a feature variable.
Hence depending on the amount of information available and the interpretability of the model required we can appropriately choose the desired model.

8. 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

9. Future Work

In future, we wish to look at models that combine decision trees and linear regression as well as those that combine perceptron and decision tree so that the explainability of decision tree can be combined with the accuracy of MLP so that our model is accurate and interpretable at the same time. We are looking to explore other possible

combinations of models as well so as to achieve the best results.

10. References

[1] 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

[2] Ünal, Ferda. (2020). Data Mining for Student Performance Prediction in Education. 10.5772/intechopen.91449.

[3] Using Data Mining To Predict Secondary School Student Performance by Paulo Cortez and Alice Silva

[4] Osmanbegović, Edin & Suljic, Mirza. (2012). DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE. Journal of Economics & Business/Economic Review. 10. 3-12.

[5] T. M. Christian and M. Ayub, "Exploration of classification using NBTree for predicting students' performance," 2014 International Conference on Data and Software Engineering (ICODSE), 2014, pp. 1-6, doi: 10.1109/ICODSE.2014.7062654.

[6] Student performance prediction using Support Vector Machine and K-Nearest Neighbor by H. Al-Shehri et al.