STUDENTS PERFORMANCE IN EXAM

Avani Saju
Msccs(DA)
Dept.Computer Science
Rajagiri College of Social Sciences

Abstract -The "Students Performance in Exam" dataset compiles information on how high school students performed on exams. There are 1000 entries in the dataset, and there are 8 variables: gender, race/ethnicity, parental education level, lunch, test preparation course, math, reading, and writing scores. These scores are used to calculate the Meanscore. The aim of this project is to classify whether the Students are Pass or Not .If the marks are 60 above they are passed else fail . Also to examine the variables that affect student performance on exams and to pinpoint the connections between these variables and exam results. The project aims to build multiple classification models using Orange tool and select the model with the best accuracy out of it and other forms of analytics, including data visualisation, can be done with this dataset.

Index Terms— Exams, score, Performance, writing

1. INTRODUCTION

Exams are an essential component of the learning process for students because they provide them the opportunity to evaluate their skills and make an intuitive connection between those skills and the future. It is a norm for evaluating students at all educational levels, from elementary school students all the way up to university students, that the short-term goal of many students is to achieve good results on tests. There are a lot of students who put in a lot of work, time, and even money in order to earn good results on their exams. It is not enough to just put in a lot of effort in order to get good results on tests like this. You will be influenced by a variety of external elements such as the approach that you take to studying, the habits that you develop throughout the process of studying, as well as the context and settings of the study.

II. LITERATURE REVIEW

[1] The objective of student performance prediction, often known as SPP, is to forecast the grade that a student will earn prior to enrolling in a class or completing an examination.
[2] This study makes a contribution to the early prediction of students who are at a high risk of failing, and it defines the most effective approaches for machine learning.
[3] It has been found through the study of this research that specific characteristics each have their own unique effect, and that eliminating data that has been incorrectly classified can have an impact on the overall findings.
[4] With the development of two different categorization models consisting of two stages, the purpose of this study is to enhance prediction. Some different approaches to machine learning: Decision Tree (J48), Random Forest (RF), Gradient Boosting Method (GBM) and Logistic Regression (LR).
[5] In order to make accurate projections of students' final grades, this article makes use of an educational data mining technique by employing a number of feature selection procedures and machine learning algorithms. The results that were obtained make an effort to determine the influence of various student-related characteristics on the prediction.
[6]The findings of these case studies provide methods for reliably predicting the academic achievement of students, and they compare the accuracy of their predictions with those provided the MI algorithms Index.ANN, XG Boost, Random Forest, and Machine Learning and Artificial Intelligence are the terms used here.
[7]Accuracy, precision, as well as f measure and recall values, are used in the evaluation of the model's performance as a performance assessment metric. The findings of the experiment indicate thatrandom forest performs significantly better than the other models that were examined.
[8]On the basis of the real data gathered by the teaching monitoring big data platform, the

performance of the model is evaluated and validated. The findings demonstrate that the performance of the model's prediction is superior than that of the comparator algorithms with regard to both the failing prediction and the GPA prediction.
[9]The purpose of this study is to present a survey of the numerous strategies that have been put into place for the prediction of student performance. The analysis is presented based on the classification algorithms, the dataset that was applied, the utilised software tools, and the performance measurements.
[10]Based on the findings of this research, the Deep Neural Network (DNN) model should be used to indicate to students the class they are in. This provides the school with knowledge, which enables it to offer a treatment for pupils who may be falling behind in their studies. In terms of accuracy, the deep neural network model that was suggested exceeds other machine learning methods that are already in use, obtaining up to 85.4% accuracy.

III. IMPLEMENTATION

Tool Used- Orange
An open-source data mining and visualisation toolkit is Orange (3.34.0). Interactive data visualisation and exploratory quick qualitative data analysis are both done using it. This platform enables workflow creation for simple data analysis using widgets for the end user.

A) Data description

The dataset used in this project was obtained from the Kaggle website. The details about the attributes given here are used for classifying whether the Students are Pass or Not .If the marks are 60 above they are passed else fail.
This dataset consists of 1000 instances and 10 features. Out of 10 variables in train data, there are 4 numeric variables and 6 are categorical variables. The attributes in the datasets are:

- Gender
- Race/ethnicity
- Parental level of education
- Lunch
- Test preparation course
- Math score
- Reading score
- Writing score
- Score (mean)
- pass_or_not

No	**Attribute**	**Description**	**Type**
1	Gender	Gender of the Students – Male ,Female	Categorical
2	Race/ethnicity	Race/ethnicity of students-Group A,B,C,D,E	Categorical
3	Parental level of education	Education of the parents- High school, some college, some high school, associates degree , Bachelors degree, Masters degree	Categorical
4	Lunch	Wheather the lunch is standard or Free/Reduced	Categorical
5	Test preparation course	Wheather the course is completed or none	categorical
6	Math score	Math score of students(13-100)	Numeric
7	Reading score	Reading score of students (28-100)	Numeric
8	Writing score	Writing score of students (31-100)	Numeric
9	Score	Average of these scores (28.6-98.6)	Numeric
10	pass_or_not	Wheather the students are pass or not	categorical

B) PRE-PROCESSING

In this case, the data set that was in categorical form has been treated as ordinal by doing
Continuize Discrete Variables as a part of the

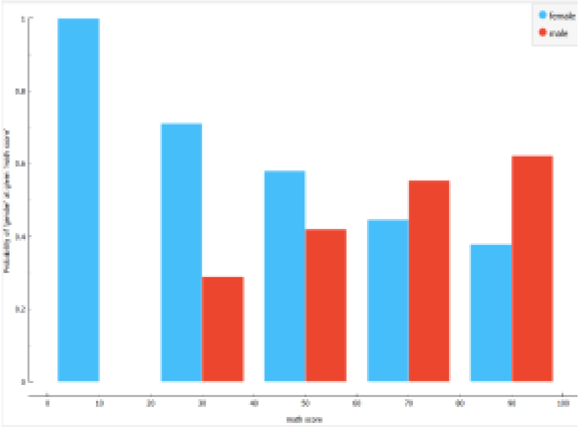
Preprocessing step in order to gain
a better understanding of the model.
Numeric values normalize to interval [0,1] It's done to keep the data safe and to make the database more flexible by getting rid of duplicates and inconsistencies.

C)Data Exploration

To analyse the data and find out more about it, the data needs to be explored.

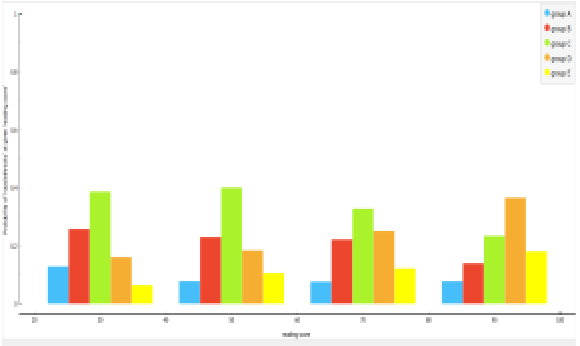
To analyse the data well, visualization techniques are used. The visualization technique mainly used here is Bar plot and distributions.

Gender vs Math score



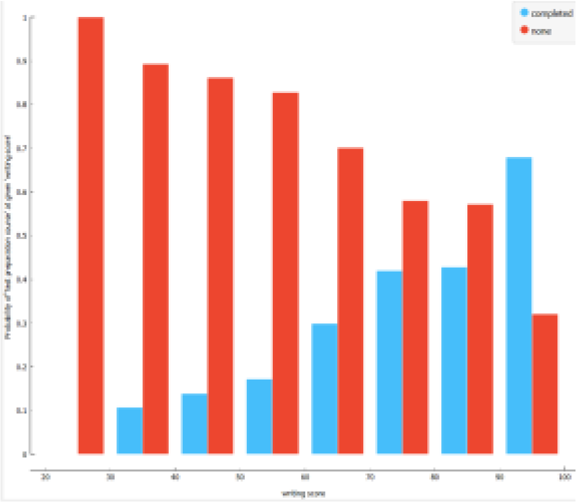
From the above chart Male students has the highest math score than female

Race/ethinicity vs Reading score



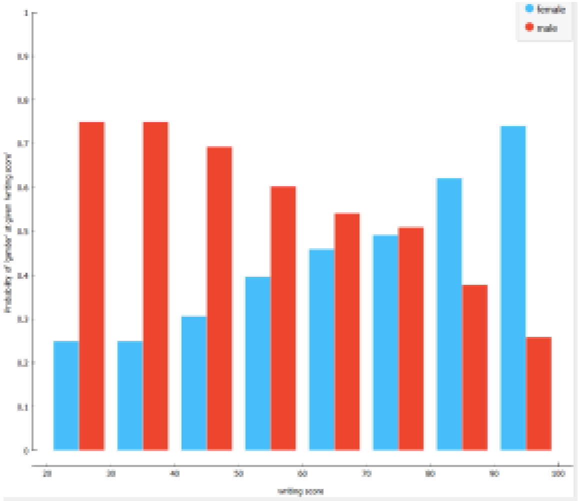
Race/ethinicity group D has highest reading score and the least score obtained by group E.

Test preparation course vs Writing score



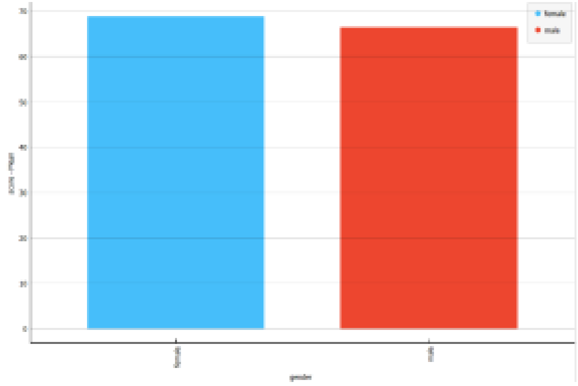
Test preparation course completed students has the highest Writing score than the students not completed the Test preparation course

Gender vs Writing score



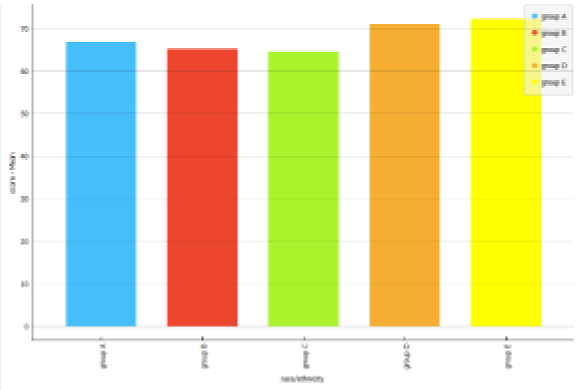
From the above chart Female students has the highest writing score than the male students

Gender vs Mean score



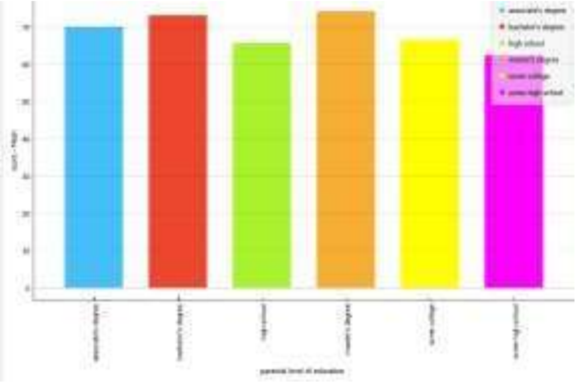
From the above bar plot female students has the highest mean score than male students

Race/ethinicity vs Mean score



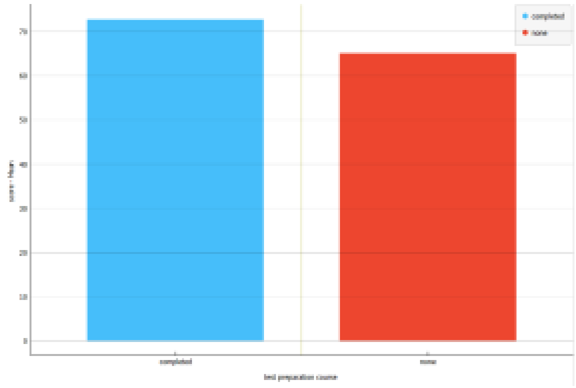
Race/ethinicity Group E has the highest mean score and the least is Group c

Parent level of education vs Mean score



The students whose parents education Masters degree got highest score and the least score obtained by the students whose parents education is high school. This shows that the Parent level of education is impacted on the students studies

Test preparation course vs Mean score



Here shows that the Test preparation course compled students got highest marks.

d) Classification Techniques

The classification techniques used in this project are K-Nearest Neighbour(kNN), Gradient boosting, Random forest , Tree, Naïve Bayes.

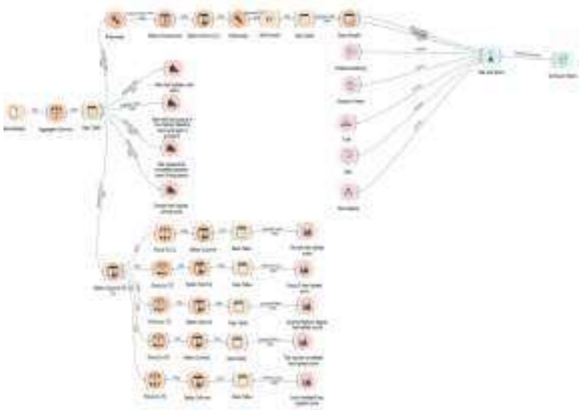
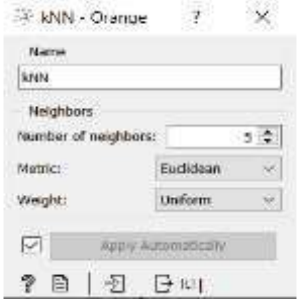


Fig: Model Building of Students performance in exam using Orange

1.K- Nearest Neighbors Algorithm (KNN)

It is a method for supervised machine learning that provides a straightforward answer to the problems of regression as well as classification. When a new sample is received, the method begins by calculating the distance between the new sample and the existing data points. After this is complete, the new sample is then assigned to the group that is shared by the k neighbours who are the nearest.



kNN - Orange

Name

kNN

Neighbors

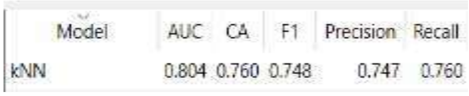
Number of neighbors: 5

Metric: Euclidean

Weight: Uniform

☑ Apply Automatically

The model evaluation value

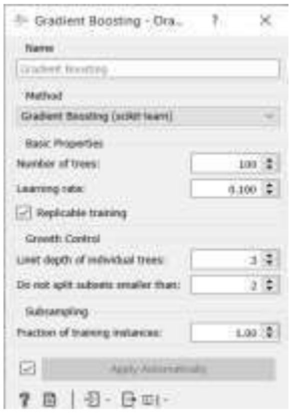


Model	AUC	CA	F1	Precision	Recall
kNN	0.804	0.760	0.748	0.747	0.760

Fig. kNN Evaluation Parameter

2.Gradient boosting

Gradient boosting is a boosting method used in classification tasks and regression and other problems, among others. It offers a prediction model in the form of a collection of weak prediction models that resemble decision trees.



Gradient Boosting - Ora...

Name: Gradient Boosting

Method: Gradient Boosting (scikit-learn)

Basic Properties

Number of trees: 100

Learning rate: 0.100

☑ Replicable training

Growth Control

Limit depth of individual trees: 3

Do not split subsets smaller than: 2

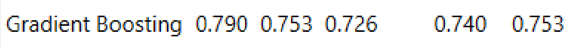
Subsampling

Fraction of training instances: 1.00

☑ Apply Automatically

Fig. Gradient Boosting Model Parameters

The model evaluation value

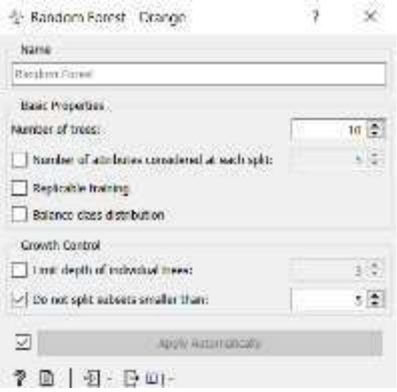


Gradient Boosting	0.790	0.753	0.726	0.740	0.753

Fig. Gradient Boosting Evaluation Parameter

3. Random forest

It is made up of many trees that work together as a whole. Each tree in the random forest makes a prediction, and the model prediction will be the one with the most votes. The random forest is a classifier that takes the average of a number of decision trees on different subsets of a given dataset to improve the accuracy of that dataset's predictions.The most trees in the forest give the most accurate results and stop the problem of overfitting.



Random Forest - Orange

Name: Random Forest

Basic Properties

Number of trees: 10

☐ Number of attributes considered at each split: 5

☐ Replicable training

☐ Balance class distribution

Growth Control

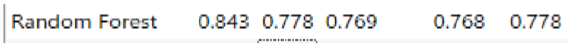
☐ Limit depth of individual trees: 3

☑ Do not split subsets smaller than: 5

☑ Apply Automatically

Fig. Random Forest Model Parameters

The model evaluation value

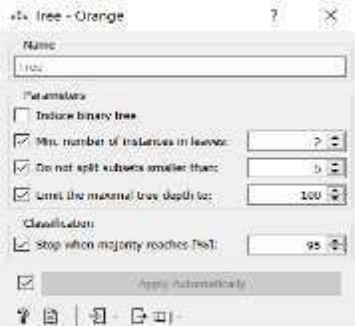


Random Forest	0.843	0.778	0.769	0.768	0.778

Fig. Random Forest Evaluation Parameter

4. Tree

Tree is an algorithm that divides data into nodes based on class purity. It comes before Random Forest. Both numerical and categorical datasets can be handled by Tree in Orange, which was created in-house.



Tree - Orange

Name: Tree

Parameters

☐ Induce binary tree

☑ Min. number of instances in leaves: 2

☑ Do not split subsets smaller than: 5

☑ Limit the maximal tree depth to: 100

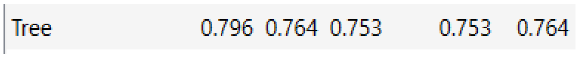
Classification

☑ Stop when majority reaches [%]: 95

☑ Apply Automatically

Fig. Tree Model Parameters

The model evaluation value

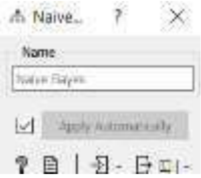


Tree	0.796	0.764	0.753	0.753	0.764

Fig. Tree Evaluation Parameter

5. Naïve Bayes

It is a type of probabilistic classifier called a Naive Bayes classifier, and it works by first estimating the conditional probabilities of the dependent variable based on the training data, and then using those estimates to classify fresh data instances. The algorithm processes discrete features relatively quickly but has a slower time when dealing with continuous features.

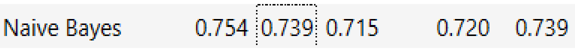


Naive...

Name: Naive Bayes

☑ Apply Automatically

The model evaluation value

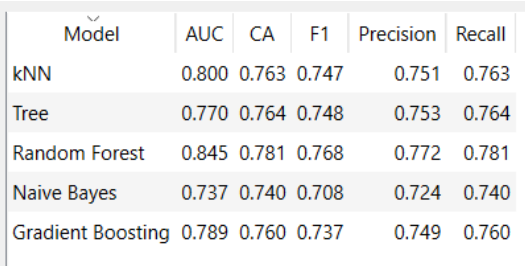


Naive Bayes	0.754	0.739	0.715	0.720	0.739

Fig. Naïve Bayes Evaluation Parameter

IV. RESULT AND DISCUSSION

This dataset consists of 1000 instances and 10 columns. Pass or Not is chosen as the target variable. Classification model applied in the datasets are KNN, Gradient boosting, Random Forest Tree and Naïve Bayes.The performance value obtained for each model is given below.

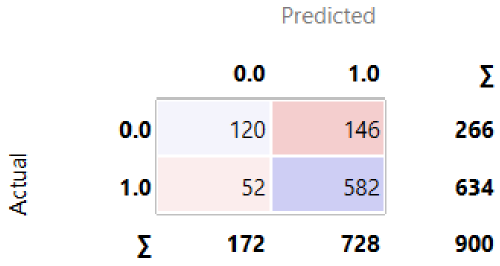


Model	AUC	CA	F1	Precision	Recall
kNN	0.800	0.763	0.747	0.751	0.763
Tree	0.770	0.764	0.748	0.753	0.764
Random Forest	0.845	0.781	0.768	0.772	0.781
Naive Bayes	0.737	0.740	0.708	0.724	0.740
Gradient Boosting	0.789	0.760	0.737	0.749	0.760

Fig. Evaluation Result

Among the applied classification models, Random Forest has highest accuracy which is 78.1%.

The confusion matrix of Random Forest is given below.



		Predicted		
		0.0	1.0	Σ
Actual	0.0	120	146	266
	1.0	52	582	634
	Σ	172	728	900

Here Pass or not ,Gender, Race/ethnicity , Parental levelof education, Lunch, ,Test preparation course are necessary for the classification for the dataset.

The table is given for the two-class classifier, which has two predictions "1" and "0." Here, 1 defines that the student is passed, and No defines that the student not passed. The classifier has made a total of 900 predictions. Out of 900 predictions, 702 are true predictions, and 198 are incorrect predictions.

Conclusion

The primary objective is to improve the academic performance of students who are considered to be at risk. , That the short-term goal of many students is to achieve good results on tests. There are a lot of students who put in a lot of work, time, and even money in order to earn good results on their exams. It is not enough to just put in a lot of effort in order to get good results on tests like this. You will be influenced by a variety of external elements such as the approach that you take to studying, the habits that you develop throughout the process of studying, as well as the context and settings of the study.

REFERENCES

[1] https://www.researchgate.net/publication/356848977_Educational_Data_Mining_Techniques_for_Student_Performance_Prediction_Method_Review_and_Comparison_Analysis

[2] https://www.researchgate.net/publication/358991518_Educational_data_mining_prediction_of_students%27_academic_performance_using_machine_learning_algorithms

[3] https://www.researchgate.net/publication/363045571_A_Novel_Hybrid_Ensemble_Clustering_Technique_for_Student_Performance_Prediction

[4]https://www.researchgate.net/publication/365514862_An_Efficient_2-Stages_Classification_Model_for_Students_Performance_Prediction

[5] https://www.researchgate.net/publication/358698343_Students_Performance_Prediction_Using_Educational_Data_Mining

[6] https://www.researchgate.net/publication/361647128_Student_Performance_Prediction_using_AI_and_ML

[7]https://www.researchgate.net/publication/366183430_STUDENT_PERFORMANCE_PREDICTION_IN_E-LEARNING_ENVIRONMENT_USING_MACHINE_LEARNING

[8]https://www.researchgate.net/publication/365013816_A_prediction_model_of_student_performance_based_on_self-attention_mechanism

[9]https://www.researchgate.net/publication/366232212_Analytical_Review_and_Study_on_Student_Performance_Prediction_A_Challenging_Overview

[10]https://www.researchgate.net/publication/368666444_Deep_neural_network_in_prediction_of_student_performance