A Project Report

on

# “*Student Performance Prediction Using Feedback*”

Submitted in partial fulfillment of the requirements for the Degree of B.Tech

in ***Information Technology***

By

**Shrey Agarwal (1806518)**
**Yashaswi Upmon (1806539)**
**Ganesh Bhandarkar (1806554)**
**Riyan Pahuja (1806566)**

under the guidance of

***Prof. Dr. Suresh Chandra Satapathy***

School of Computer Engineering
***Kalinga Institute of Industrial Technology Deemed to be University***
Bhubaneswar

12 May, 2021

KALINGA INSTITUTE
OF INDUSTRIAL TECHNOLOGY
Deemed to be University u/s 3 of the UGC Act, 1956

KIIT Deemed to be University
School of Computer Engineering
Bhubaneswar, ODISHA 751024

# CERTIFICATE

This is to certify that the project report entitled

“***Student Performance Prediction Using Feedback***”

submitted by

| | |
|---|---|
| Shrey Agarwal | (1806518) |
| Yashaswi Upmon | (1806539) |
| Ganesh Bhandarkar | (1806554) |
| Riyan Pahuja | (1806566) |

in partial fulfillment of the requirements for the award of the **Degree of Bachelor of Technology** in **Discipline of Engineering** is a bonafide record of the work carried out under my(our) guidance and supervision at School of Computer Science, Kalinga Institute of Industrial Technology, Deemed to be University.

Date: 12 May, 2021

(Prof. Suresh Chandra Satapathy)
Project Guide

# ACKNOWLEDGEMENTS

Our team would like to thank Prof. Suresh Chandra Satapathy for his guidance and supervision in the project. The project would not be possible without his expert guidance and encouragement. We would also like to thank Prof. Bhabani Shankar Mishra, Dean of school of computer engineering for this amazing opportunity which has enabled us to learn and grow. This Project won't be possible without the help of our Project supervisor and other professor who helped us in the selection of attributes of data. Lastly, we would also like to express our gratitude towards the scientific community which has enabled us to refer to world class research for our betterment.

| | |
|---|---|
| **Shrey Agrawal** | ShreyAgarwal |
| **Yashaswi Upmon** | Yashaswi upmon |
| **Riyan Pahuja** | Riyan Pahuja |
| **Ganesh Bhandarkar** | Ganesh |

# ABSTRACT

The education industry has gone through major changes amidst the recent COVID-19 pandemic. Facing unforeseen circumstances, educational institutions were forced to shift to an online learning model rather than an offline(classroom) based learning model. The sudden change in learning model impacted not only the students but also the teaching faculty. Even though many resources are available online, simulating a classroom-like study environment is not an easy task. Hence mapping student performance in the new learning model is an essential task. The main goal of our work is to predict the student performance in the online learning model implemented by many colleges and universities amidst the COVID-19 pandemic. An online survey was conducted to collect the data from the students who had undergone the aforementioned learning model for at least one semester. The data set for the research includes features that would have an impact on a student's performance. The model strives to predict a student's performance with good accuracy and help infer where the online learning model can be improved. Several classifiers such as KNN, Gradientboost, Adaboost, Decision tree, SVM ,Gaussian NB and many more algorithms were applied to the data set which explain later in this work and their performance was analysed and compared. The GradientBoost, Xgboost Classifier and SVM classifiers returned highest accuracies, in essence **97.46%, 97.45% and 97.45**% respectively. This indicates that the performance of the students is predictable with the given features**.**

*Keywords:* Machine Learning, Classification Model, Ensemble modeling, student performance prediction, Accuracy.

# TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

CHAPTER 2: BACKGROUND/BASIC CONCEPTS

CHAPTER3: PROJECT ANALYSIS/ PROJECT IMPLEMENTATION

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1 - INTRODUCTION

## 1.1 Purpose

Prior to 2020, classroom learning dominated the educational industry and was considered the norm. The increasing cases of COVID-19 in the country made a drastic change in the way educational institutions used to interact with the students. When a strict lockdown was imposed in the countries, education institutions shifted to teaching through online platforms which drastically increased the use of video conferencing applications such as Zoom, Google meet, skype, Microsoft teams and many more such softwares to interact with students online. Due to this change in the interaction between faculty and students there was bound to be some change in the performance of students. Forecasting performance of students is essential as there is a paradigm shift and will also help in making the system more flawless and robust.

## 1.2 Scope of Project/Model

This paper applies classification models on the gathered data, to predict student performance in the online learning system. The data has been collected from students who have participated in online semesters. Their performance in the classroom based learning model has been ignored to not introduce any bias in machine learning models. The objective of this project is to classify a student’s performance on the basis of their feedback in this mode as it can easily be predicted in the earlier mode as represented in the references. These features can be focused on by educational institutions to envisage the student performance.

## 1.3 Overview of Document

This paper is divided into 5 sections, the objective and rationale of the work is talked about in the introduction. A basic overview of the technology and concepts used is provided in the second, background section. The third section presents analysis and implementation of the project. The fourth section presents the achievements of the research along with final results. The paper is concluded in the fifth section with a short summary and a discussion about further work

# CHAPTER 2 - BACKGROUND

The main objective of the student performance prediction is to help teachers to know how the students will perform in the current semester using their feedback. Thus, this model may help in the improvement of student's performance in academia over a period of time. Further, teaching improvement can be made by creating a recommendation system for the universities and teaching faculty.

## 2.1 Machine Learning

Machine Learning is the subfield of artificial intelligence. By using a large number of data over a model/program we can train our model to give correct results over the similar data or any unseen data. We can train our models on multiple constraints and dimensions. It will ensure that there is not overfitting and any data leaks. To be Specific, machine learning models are the programs that make it possible to get future predictions easily with less errors. Machine learning algorithms are used in a wide variety of applications, such as in Medical, science, astronomy, energy, etc.

Machine Learning is divided into 3 Different parts :

1) Supervised Learning
2) Unsupervised Learning
3) Reinforcement Learning

### 2.1.1 Supervised Learning

Supervised learning is a method of learning in which our model is trained over a well structured and labeled data. i.e. there is data present with the correct answers provided to the model for training and some part of training data is used for testing the prediction . So the supervised learning analyses the testing data and returns the correct outcome.

Supervised learning have two different algorithms :

1) Classification
2) Regression

### 2.1.2 Classification

Machine learning classification algorithms deal with the classification problems when output is of categorical type.

There are two types of classification problems:

1) Binary classification : examples - spam detection
2) Multiclass classification: example - character recognition

### 2.1.3 Data Preprocessing

The data preprocessing is the method of processing the data to make it suitable for the model to understand. Here we convert the raw data into well structured data. This is the very first step in machine learning. A raw data contains noises, missing values, and in an unusable format which cannot be directly used for models. It is required for cleaning the data and making it suitable for a machine learning model which increases the accuracy and efficiency of a machine learning model.

### 2.1.4 Label Encoding

Label Encoding is basically the process of encoding the labels with the numeric form example : (tall, short, medium) -to- (0,1,2). It makes labels into machine readable form. This is the most important initial step of pre-processing the structured data in Supervised Learning .

# 2.2 System and Model Related

## 2.2.1 Algorithms Used

- KNN
- LOGISTIC REGRESSION
- ANN
- GRADIENT BOOSTING CLASSIFIER
- ADABOOST CLASSIFIER
- DECISION TREE CLASSIFIER
- SVM CLASSIFIER
- XGBOOST
- GAUSSIAN NB
- VOTING CLASSIFIER

## 2.2.2 Evaluation Metrics

The Evaluation Metrics give brief details about the performance Machine Learning model. This information helps us in improving over the accuracy of the model by making changes and tweaks in our Algorithm. This is very important to check the feedback frequently else the model will not make progress.

There are two important Evaluation Metrics to have feedback:

1) **Precision**: The percentage of your results which are relevant .
2) **Recall**: The percentage of total relevant results correctly classified by out algorithm.
3) **AUC - ROC curve**: This curve tells about the performance of model for classification model with different threshold tweaks.

* ( AUC - Area Under The Curve , ROC - Receiver Operating Characteristics )

The Evaluation metrics have a Score for the Model's Accuracy known as **F1-SCORE**.

**F1 Score** is the weighted average of Precision and Recall. It can be positive or negative as it is weighted . The Formula for calculating F1 Score is as follows :

**F1-Score = 2*((precision*recall)/(precision+recall))**

## 2.2.3 Software Requirements and Specifications

Operating System -

- WINDOWS 7/8/10
- LINUX
- MAC OS

Software Tools -

- ANACONDA
- JUPYTER NOTEBOOK
- JUPYTER LAB
- PYTHON  (__ver__ 3.8)

File Formats -

- CSV
- PICKLE
- NOTEBOOK (ipynb)
- TXT

Libraries used -

- Pandas
- Numpy
- Scipy
- Scikit - learn
- Keras
- Matplotlib

# CHAPTER 3 - PROJECT ANALYSIS / PROJECT IMPLEMENTATION

This section consists of a Model Architecture and explains the stepwise process of data collection, data cleaning, data preprocessing and other important steps. It also explains the different classification models that were considered in this work.

## 3.1 Model Architecture:

The comprehensive research architecture described in Figure 1. represents the flow of the constructed model.

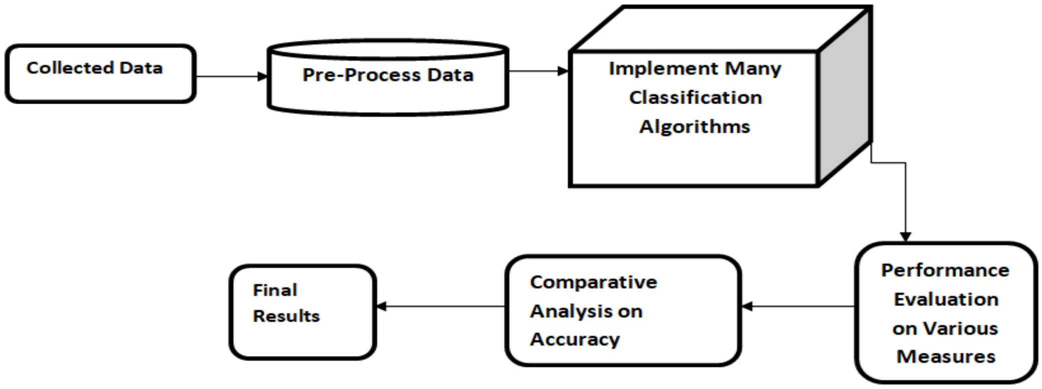

Collected Data
Pre-Process Data
Implement Many Classification Algorithms
Performance Evaluation on Various Measures
Comparative Analysis on Accuracy
Final Results

Figure 1. Layout of Research Architecture

The step are as follows:

- The data has been gathered from an online survey. Google form platform has been used for conducting the survey. Feedback is submitted by students from different universities pursuing different fields of studies. Feedback for the online semester was only considered while collecting the data.
- Preprocessing has been done after collecting the data by increasing the correlation between the different attributes, synthetically oversampled the data as the dataset was imbalanced.

- Classification Technique: The classification technique used are as follow:
    - KNN : This algorithm finds k nearest neighbors and uses distance metrics to find the distance between training sample and test sample for classification for numerical record.[7]
    - Gradient Boosting- This algorithm uses gradient descent and boosting to increase the efficiency of the prediction model. It combines weak learners into one strong learner.
    - Adaboost Classifier- This boosting algorithm introduces a weak learner in each stage to overcome the weakness of the existing weak learners. It uses high-weight data points to identify the weakness of the existing weak learners.
    - Decision Tree Classifier- It is a hierarchical model that makes use of decision rules to partition the feature space of a dataset into single class subspaces.[8]
    - SVM Classifier- It classifies the data using a hyperplane which maximizes the margin from the nearest data point of all the classes.
    - Gaussian Naive Bayes Classifier- It computes conditional class probabilities to find the most probable class for the training data. [9]
    - Logistic Regression Classifier- It computes the probability of the target class happening and then applies a natural log to output the logit of the target class.
    - Stochastic Gradient Descent Classifier- It selects random samples of the data in each iteration and applies gradient descent to it. It is useful when the dataset is large.
    - ANN Classifier- It is a classifier in which 3 layers of neurons are used for classifying the data which are input layer, hidden layer and output layer.
    - XGboost Classifier- It is a decision tree based machine learning classification algorithm. It uses gradient boosting for improving upon decision tree classifiers.
    - Voting Classifier- It is a classification model that employees multiple classification models. Due to the presence of multiple classification models it is most useful in a situation where there is confusion about which classification model to use. Voting classifiers would use the prediction which is the most frequent.

## 3.2 Data Information:

Using an online survey we collected data from the students of numerous colleges pursuing majors in different courses. Platform used for this collection of data was google form. Only online semester feedback data was collected during the survey.

Table 2. is resepenting the attributes present in the dataset.

| | | |
|---|---|---|
| Attributes | Q1. Did you face any internet connectivity issues during the class? | Q10. Was the Examination Portal well managed for this subject? |
| | Q2. How many times on average did the meeting disconnect? | Q11. The alternative is provided by the University if the uploading option in the exam portal is not working? |
| | Q3. Were you able to understand the concept? | Q12. Was the professor available for the doubts? |
| | Q4. Which platform was used for taking the class? | Q13. Lecture recording was provided? |
| | Q5. Attendance in the subject? | Q14. Interest in the subject? |
| | Q6. The interactive tools used for teaching in class | Q15.For how much time on average did you attentively attend the classes? |
| | Q7. How many questions did the teacher ask in the class? | Q16.Did you get the choice of your Teacher in the subject? |
| | Q8. How many questions did you ask the teacher? | Q17.How many questions in the exam you could have answered on your own? |
| | Q9. Which study material was provided by the university that you used for studying? | Q18.Your grade in the subject |

Table 3.2.1. Various Attributes In the Dataset

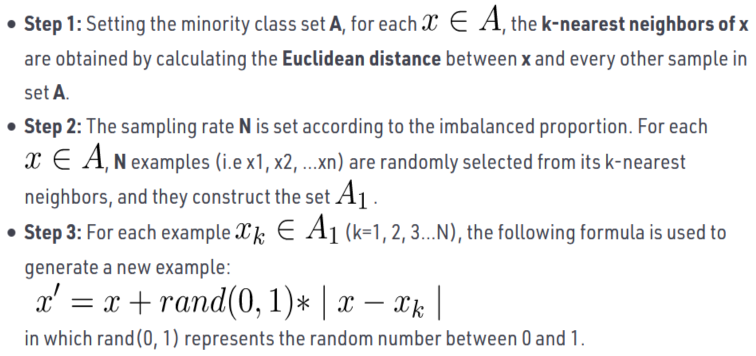
These attributes are used to analyze their performance in the particular subjects. During this current online learning method we found out that these attributes or features influence the performance of students in the significant manner.  This is finalized after consulting with our supervisor and the other professor in our university who helps us in this project.

## **3.3. Data preprocessing**:

After Data has been gathered the NaN and inconsistent rows were imputed with the mode in the dataset. We have split the continuous variable of our target value i.e. the grade of the student into three different categories i.e “Below Average”, “Average”, ”Above Average” that has been label encoded along with other columns of the dataset and the standardization of the feature are also made using standardscaler that scale the data to the unit variance by removing the mean. The dataset was imbalanced which might have given good accuracy but would not be able to anticipate the minority classes accurately, and  this would have caused a problem.

To resolve this issue we have also implemented SMOTE( synthetic minority oversampling technique).

SMOTE oversamples the minority class data by matching the number of records of the majority class and thereby it makes the dataset balanced. Steps followed by SMOTE Algorithm have been described in Figure 2.



- **Step 1:** Setting the minority class set **A**, for each $x \in A$, the **k-nearest neighbors of x** are obtained by calculating the **Euclidean distance** between **x** and every other sample in set **A**.
- **Step 2:** The sampling rate **N** is set according to the imbalanced proportion. For each $x \in A$, **N** examples (i.e x1, x2, ...xn) are randomly selected from its k-nearest neighbors, and they construct the set $A_1$ .
- **Step 3:** For each example $x_k \in A_1$ (k=1, 2, 3...N), the following formula is used to generate a new example:

$$x' = x + rand(0,1) * \mid x - x_k \mid$$

in which rand(0, 1) represents the random number between 0 and 1.

Figure 2. SMOTE Algorithm

Figure 3. represent the number of the record of each class before using the SMOTE Algorithm i.e. “Class 0” is “Average”, “Class 1” is “Below Average” and “Class 2” is “Above Average”. Figure 4. describe the number of records of each class after using the SMOTE Algorithm the description of the xlabel i.e. classes in Figure 4. is the same as that for Figure 3.

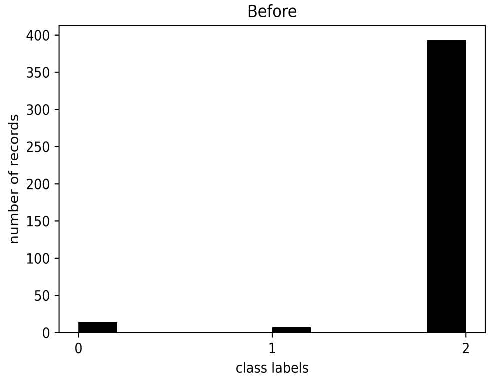


Figure 3. Number of records before using SMOTE

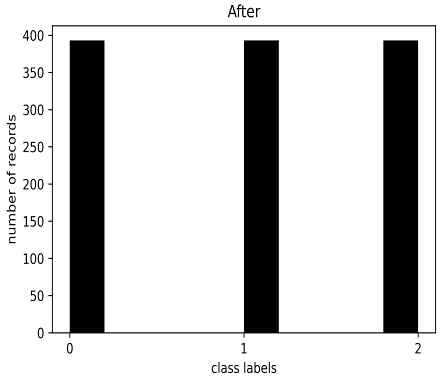


Figure 4. Number of records after using SMOTE

## **3.4 Implementation**:

Then in order to classify the data and predict the performance of the students, eleven classification techniques has been used i.e. KNN, gradient boosting,adaboost, decision tree, SVM, Gaussian Naive Bayes, Logistic regression, SGD, ANN, xgboost and Voting classifier. Voting classifier uses the model with the maximum votes with the individual models being KNN classifier, gradient boosting classifier and Xgboost classifier were used as mentioned. Out of which the SVM Classifier, Xgboost Classifier, and adaboost are performing best in the classification.

The hyperparameter tuning of each algorithm has been done using the method i.e. GridSearchCV which is a function that helps to find the hyperparameter that gives the best performance across the listed hyperparameters. By iterating throughout the predefined hyperparameters and fitting the model on the training set. Except for the KNN algorithm we have plotted the accuracy vs K-Value of KNN in Figure 5. From this we can find that the best accuracy we got is 92.5% at k=3.

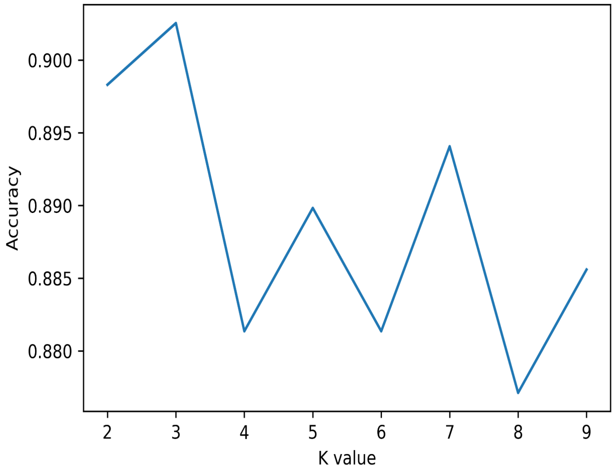
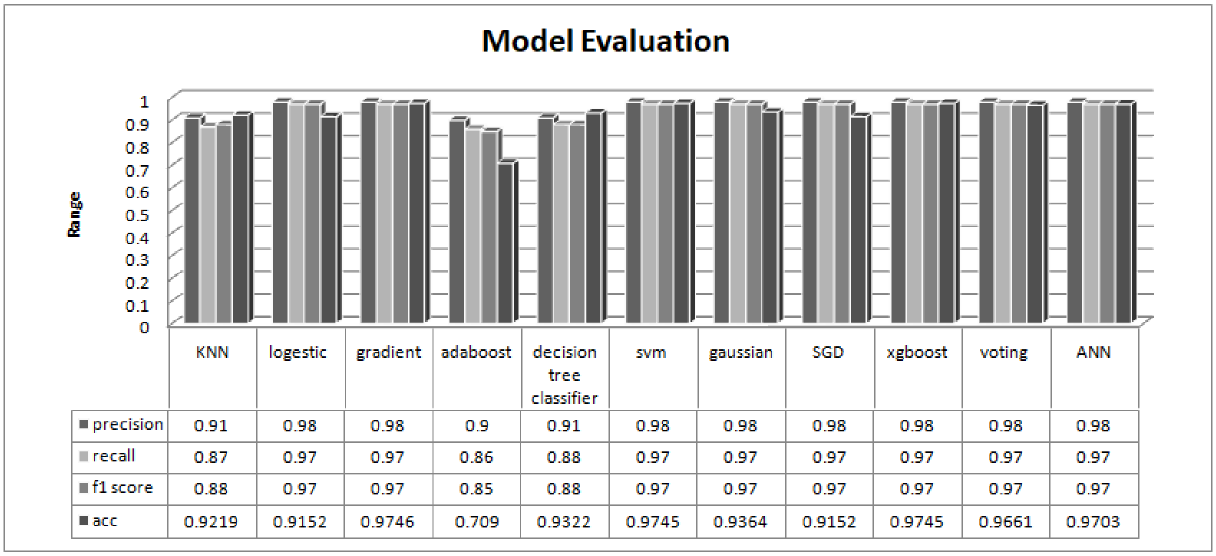


Figure 5. K-value vs Accuracy for KNN Algorithm

# CHAPTER 4 -
# RESULTS & DISCUSSIONS

From Figure 3. It has been found that SVM Classifier, Xgboost Classifier, and Gradient Classifier are best in  predicting the performance of students as compared to the other implemented models in this work. In Figure 3. Performance of all the classification models are plotted on the basis of various measures i.e. precision, recall, F-Measure, Accuracy. Precision quantifies the number Of true positives over “Below Average”,” Average” and “Above Average” classes of the student performance with sum of true positive and false positive over all the mentioned classes. Recall is tp calculated sensitivity of the classifiers. Harmonic means of precision and recall is defined as F-Measure. Accuracy of the algorithm in predicting the performance of students is defined as Accuracy.



| | KNN | logestic | gradient | adaboost | decision tree classifier | svm | gaussian | SGD | xgboost | voting | ANN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ■ precision | 0.91 | 0.98 | 0.98 | 0.9 | 0.91 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| ■ recall | 0.87 | 0.97 | 0.97 | 0.86 | 0.88 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |
| ■ f1 score | 0.88 | 0.97 | 0.97 | 0.85 | 0.88 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |
| ■ acc | 0.9219 | 0.9152 | 0.9746 | 0.709 | 0.9322 | 0.9745 | 0.9364 | 0.9152 | 0.9745 | 0.9661 | 0.9703 |

Figure 3. Performance of Various Classification Models

# CHAPTER 5 - CONCLUSION AND FUTURE WORK

## 5.1 Conclusion

During this pandemic online mode of learning is the new norm. Our work will be helpful in every university and benefit all the professors to take an instant action and keep the track of daily performance of the student on the basis of their feedback. In this work eleven classification models have been implemented as mentioned earlier and results have been analyzed on the basis of various measures. SVM Classifier, Xgboost Classifier, and Gradient Boosting classifier achieved the best accuracy of 97.45%, 97.45%, and 97.46% respectively on the data collected from an online survey. In Future, Recommendation system can be created for the professor to increase the efficiency and performance of students. More attributes of feedback can be taken into consideration to improve the accuracy.

## 5.2 Future Work

The Future work on this model is as follows:

We are planning to deploy the web application for the model based on a python library known as Streamlit.

Streamlit is the basic python library that allows the user to make the live implementation of any machine learning model with the help of a pickle file processed to generate the output. It also allows the user to make the 3D depiction of all the related visualization and graphs in real time.

The plan is to make this model public for all the schools and colleges where the organization submit their document with all queries. These queries will be processed by the model and the output will be generated.

*(Limitations : The Data uploaded by organizations should be clean)

## 5.3 Planning And Project Management

**Table 5.3.1 showing details about project planning and management**

| **Activity** | **Starting week** | **Number of weeks** |
|---|---|---|
| Ideation | 1st week of December | 3 |
| Literature review | 4th week of December | 3 |
| Finalizing problem | 3rd week of January | 1 |
| Designing Questions for Survey | 4th week of January | 1 |
| Finalizing the attributes of data after consults with professor | 1nd week of February | 1 |
| Time taken by online Survey | 2nd week of February | 4 |
| Data Preprocessing and Data analysis | 2nd week of March | 2 |
| Implementation of various classification models | 4th week of March | 2 |
| Hyperparameter tuning of the models | 2nd week of April | 2 |
| Finalizing the results | 4rd week of April | 2 |
| Preparation of project report | 1st week of May | 1 |
| Preparation of project presentation | 2nd week of May | 1 |

**The Gantt Chart is given below:**

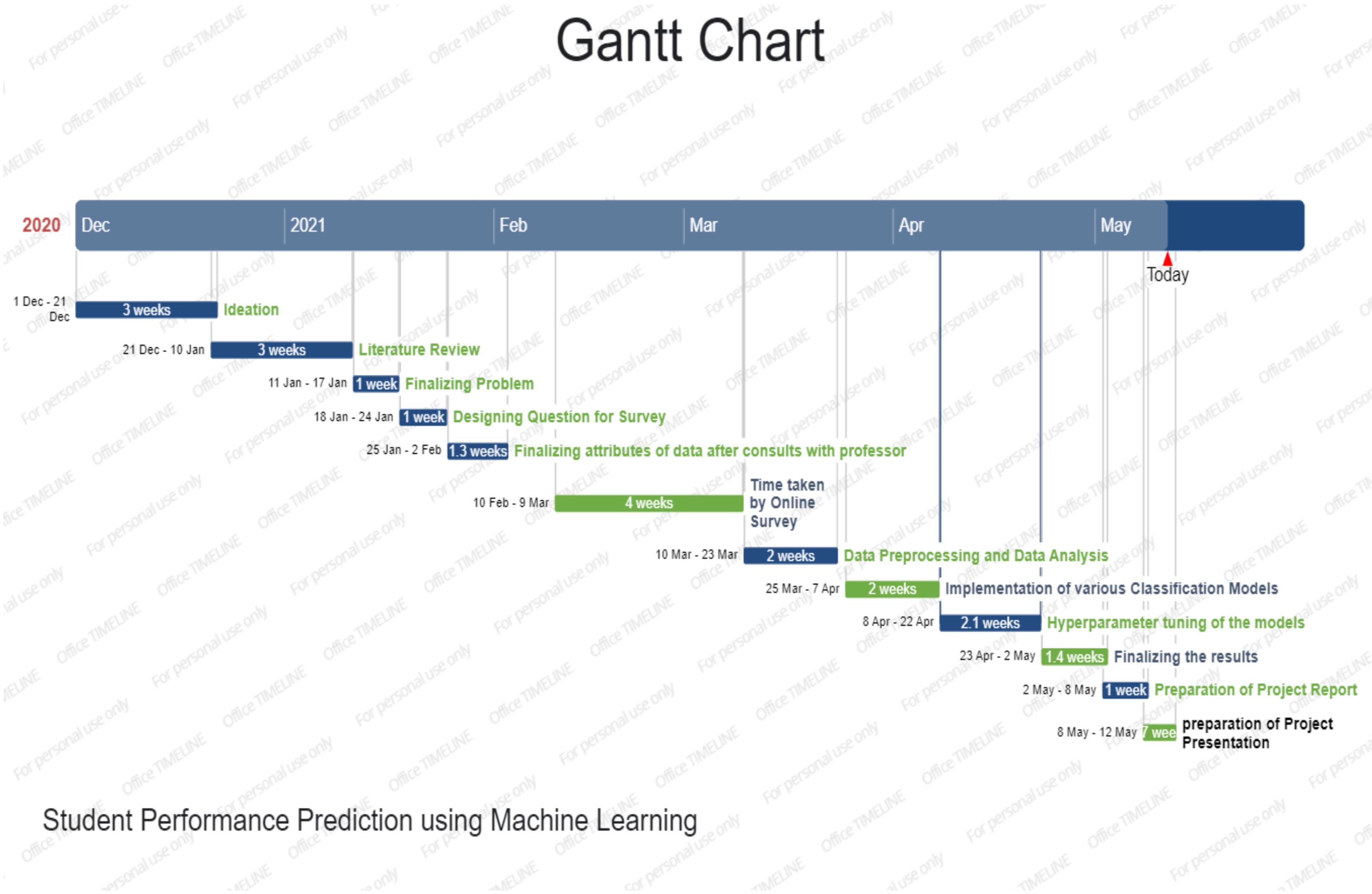


Fig 4:: Gantt chart

## REFERENCES

1. Sa, Chew Li, Emmy Dahliana Hossain, and Mohammad bin Hossin. "Student performance analysis system (SPAS)." In *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, pp. 1-6. IEEE, 2014.

2. Su, Yu, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris Ding, Si Wei, and Guoping Hu. "Exercise-enhanced sequential modeling for student performance prediction." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. 2018. (EERNN method)

3. Sekeroglu, Boran, Kamil Dimililer, and Kubra Tuncal. "Student performance prediction and classification using machine learning algorithms." In *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, pp. 7-11. 2019. (EERNN method)

4. Sorour, Shaymaa E., Tsunenori Mine, Kazumasa Goda, and Sachio Hirokawa. "A predictive model to evaluate student performance." *Journal of Information Processing* 23, no. 2 (2015): 192-201.

5.Yang, Fan, and Frederick WB Li. "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining." *Computers & Education* 123 (2018): 97-108. (BP-NN)

6.Alshabandar, Raghad, Abir Hussain, Robert Keight, and Wasiq Khan. "Students Performance Prediction in Online Courses Using Machine Learning Algorithms." In the 2020 *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7. IEEE, 2020.

7. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

8. Myles, Anthony J., Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. "An introduction to decision tree modeling." *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, no. 6 (2004): 275-285.

9. Moraes, Ronei M., and Liliane S. Machado. "Gaussian naive bayes for online training assessment in virtual reality-based simulators." *Mathware & Soft Computing* 16, no. 2 (2009): 123-132

# Student Performance Prediction Using Machine Learning

**YASHASWI UPMON**

**1806539**

## Abstract

The aim of the project is to create a bridge between the performance of the students in online exams and Learning made by them during these online classes to improve the grades with changes in Online teaching techniques. In this project, we have collected real-life data through an online survey regarding this student feedback and performance of the student in a particular subject. The data contain many attributes which have been mentioned in this project. The preprocessing of data is done using multiple libraries and techniques mentioned above. The project is using multiple classification models to classify between the grades and different attributes and is compared on the basis of many measures. The best performance was achieved by SVM, Xgboost, and Gradient boosting i.e. 97.45%, 97.45%, 97.46% respectively.

**Individual contribution and findings**:

Initially, the online survey has been conducted for the gathering of the data from the student regarding the feedback of this online learning model specifically about their respective subjects. A small set of questions or attributes has been decided by my team and me which being as students we find will affect any student performance in this learning mode during the pandemic. All those attributes have been mentioned above in this report. After collecting this real-life data during EDA(exploratory data analysis) I have found that the data is quite imbalanced. That would have created a huge problem for our work as when imbalanced data has fed to our model it might give good accuracy but will fail to accurately predict the minority class samples. To address this I have split the continuous variable into the categorical variable which helps in the prediction of the specific classes of the target value more precisely. The grades of the students are grouped together and the final target classes are labeled as “above average”, “average” and “below average". I have implemented many classification models like logistic regression classifier it is a supervised learning classification algorithm used to predict the probability of a target value, gradient boosting classifier, Adaboost classifier, SGD classifier, xgboost classifier, and voting classifier which have used the maximum votes from each model as KNN classifier, gradient boosting classifier and Xgboost classifier. Each of these algorithms is calculated by another team member on which he takes various performance measures like recall, precision, F-Measure, Accuracy. The accuracy of each measure after learning from my team member has been plotted by me. The best classifier models such as SVM, Xgboost, and Gradient boosting. I have hyperparameter tuned the models using gridsearchcv which helps in incrementing the accuracy of the model in significant numbers. In this project, I have learned many new technical skills like how to work on the real-life dataset? What are the key factors and how do you do the preprocessing before just feeding your data into the model? how to do hyperparameter tuning of the models?. Along with the technical skills I have learned, the main skills which help in life are team management and working with other co-members that i learned while working on this project.

**Individual contribution to project report preparation:** My work in the project report preparation was in Project analysis/ project implementation, Results & Discussion, Conclusion And Future Work. The Project analysis represents and contains the detailed brief of Model Architecture, Data Information, Data Preprocessing and Data Implementation. The results have been described in the Result and Discussion section with an detailed analysis of accuracy of each model used in this project. The work is concluded and the future scope of this project is described in Conclusion and Future Work scope.

**Individual contribution for project presentation and demonstration**: All the Co members of the team have contributed equally in the presentation preparation. My contribution in the ppt presentation is the subsection of Data preprocessing and Model evolution, Results Analysis and Conclusion my co member shrey agarwal has also contributed in this subsections.

Full Signature of Supervisor: Full signature of the
student:

Yashaswi upmon

(Prof. Dr. Suresh Chander Sathapaty ) YASHASWI
UPMON

# Student Performance Prediction Using Machine Learning

Ganesh Bhandarkar

1806554

**Abstract**： The aim of the project is to create a bridge between performance of the students in online exams and Learning made by them during these online classes to improve the grades with changes in Online teaching techniques. In this project, we have collected the real life data through an online survey regarding this student feedback and performance of the student in a particular subject. The data contain many attributes which have been mentioned in this project. The preprocessing of data is done using multiple libraries and techniques mentioned above. The project is using multiple classification models to classify between the grades and different attributes and are compared on the basis of many measures. The best performance achieved by SVM, Xgboost and Gradient Boosting.e. 97.45%, 97.45%, 97.46% respectively.

**Individual contribution and findings**: This project helped me learn multiple things to add in my technical stack. At the initial stage of the project, we started with reading multiple research papers and evaluating them with the categories. My teammates and I tried out multiple ideas and After discussion with our project guide we came up with the most crucial problem statement during the time of online examinations, everyone mutually agreed to proceed further with it .

We created a small scale survey of students that have attended the online classes and exams .This survey was circulated around multiple colleges around India including government colleges, private colleges and different universities. My project team and I created a curated list of questions that were involved in the survey. So these questions were depending on multiple factors, attributes and situations faced during online classes or exams. At the beginning of practical implementation of the project, we distributed work in a team . My role in the project was to test the algorithms and analyse the data at time of preprocessing. I also implemented several classification algorithms that can help in improving accuracy of the models. I also create a record for the progress our team made and algorithms implemented upto date. I tested out the ensemble learning Model to gather the decisions combined by multiple models. The findings were that if the higher grades and other grades are grouped then the accuracy obtained is better. My experience was exciting as I learned many things about multiple algorithms and Classification models. I learned about research patterns used in research papers while reading through multiple research papers .

My teammates helped me out whenever I needed help and motivated me to learn multiple concepts implemented during this project. This project helped me to implement the algorithms that were only known by me theoretically. Our project guide was very helpful and supportive during this period of project creation. He guided us through all the obstacles we encountered during implementing our project.

**Individual contribution to project report preparation:** My contribution in preparing project report was working around completing certificate forms, Chapter 2 (2.1 - Machine Learning, 2.2 - System Related), Gantt Chart, Table Content, List of Tables, List of Figures. Chapter 5 . Future works and Progress tables. I created a gantt chart with the help of a web application . Finally , I cured the indentations and size constraints in this document.

**Individual contribution for project presentation and demonstration**: Every team member has participated equally in the preparation of the project presentation . Mainly focus was around preparing the Overview of the project presentation, various algorithms used in preparing the project, and process of time management with scheduling i.e. Gantt chart. I also wrote the small brief about all the algorithms used in the project .

Full Signature of Supervisor:

…………………………….

(Prof. Dr. Suresh Chander Sathapaty )

Full signature of the student:

Ganesh

……………………………..

(Ganesh Bhandarkar)

# Student Performance Prediction Using Machine Learning

Shrey Agarwal

1806518

**Abstract**： The aim of the project is to create a bridge between performance of the students in online exams and Learning made by them during these online classes to improve the grades with changes in Online teaching techniques. In this project, we have collected the real life data through an online survey regarding this student feedback and performance of the student in a particular subject. The data contain many attributes which have been mentioned in this project. The preprocessing of data is done using multiple libraries and techniques mentioned above. The project is using multiple classification models to classify between the grades and different attributes and are compared on the basis of many measures. The best performance achieved by SVM, Xgboost and Gradient Boosting.e. 97.45%, 97.45%, 97.46% respectively.

**Individual contribution and findings**: After reading different papers related to student performance prediction, I came up with the idea of predicting student performance in the online teaching method by taking feedback from students. I shared this idea with my group members and my project guide, and everyone mutually agreed to proceed further with it. We all created a survey that we thought would most impact a student's performance and our supervisor finally reviewed it. After that, we all started taking surveys from the students at our university as well as in other universities. I gathered about 200 records. After this, I preprocessed the data by removing all the NaN rows and converting the required data type from object to int, float, and string. After reading different papers related to student performance prediction, I came up with the idea of predicting student performance in the online teaching method by taking feedback from students. I shared this idea with my group members and my project guide, and everyone mutually agreed to proceed further with it. We all created a survey that we thought would most affect a student's performance and our supervisor finally reviewed it. After that, we all started taking surveys from the students in our university as well as in other universities. I gathered about 200 records. After this, I preprocessed the data by removing all the NaN rows and converting the required data type from object to int, float, and string. After which I performed exploratory data analysis, in which I described the table and calculated value counts of each column to get a good outlook of the dataset. After which I applied different classification models like KNN, decision tree, SVM, and ANN. But we were not getting good accuracy and so Yashaswi came up with the idea of categorizing the class labels as well as other attributes which would improve the performance and I helped him do it. I classified the class labels 8-10 as above average, 6-7 as average, and below 5 as below average and this increased the accuracy of our models but the f1 score of our model was very low as the data of the last two labels was very less, so I thought of oversampling the data synthetically by using synthetic minority over-sampling technique and also plotted the number of records before vs after oversampling. After which I printed the classification report for all the models, which gave good results. Through this project, I implemented many things, which I knew only theoretically. I could analyze things myself better and this project overall boosted my confidence in the skills I have and it also taught me how to look at various angles to solve a problem. Working in a group taught me how to coordinate, the importance of the push and motivation by other team members and how people perceive a problem, and the different ways they react to it made me more open-minded and the experience gained from working on this project made me more intuitive.

**Individual contribution to project report preparation:** My work in the project report preparation was in chapter-3 and chapter-4. Chapter-3 describes the model architecture, data information, data preprocessing and implementation. Chapter-4 gives a good overview of the results found after implementing our project.

**Individual contribution for project presentation and demonstration**: All the Co members of the team have contributed equally in the presentation preparation. My contribution in the ppt presentation is the subsection of Data preprocessing and Model evolution, Full Signature of Supervisor: Full signature of the student:

Shrey Agarwal

……………………………. ……………………………..

(Prof. Dr. Suresh Chander Sathapaty ) (Shrey Agarwal)

## Student Performance Prediction Using Machine Learning

RIYAN PAHUJA

1806566

### Abstract

The aim of the project is to create a bridge between performance of the students in online exams and Learning made by them during these online classes to improve the grades with changes in Online teaching techniques. In this project, we have collected the real life data through an online survey regarding this student feedback and performance of the student in a particular subject. The data contain many attributes which have been mentioned in this project. The preprocessing of data is done using multiple libraries and techniques mentioned above. The project is using multiple classification models to classify between the grades and different attributes and are compared on the basis of many measures. The best performance achieved by SVM, Xgboost and Gradient i.e. 97.45%, 97.45%, 97.46% respectively.

**Individual contribution and findings**:

The world has been experiencing a major COVID-19 outbreak for over a year now. The education system faced drastic changes in teaching and learning patterns during this period. The educational institutions have shifted to teaching the students through online models to ensure everyone’s safety during such trying times. This situation called upon the need to analyse the student’s performance in such a new system. To predict their performance to identify weak and strong performing students and to ensure the viability of this system along with improving it at the same time.

As we decided to do an online survey for collection of student performance data for analysis. I helped organise the form and analyse the questionnaire. I formulated questions that would have the highest impact on a student’s performance in an online learning model based on previous research combined with some general sense of the educational system that I have developed. After the questionnaire had been developed it was upon us to properly advertise it thoroughly not in our college but across multiple colleges. I was responsible for getting it advertised in multiple colleges in my vicinity. There were some challenges that I faced in this task. The most concerning was that not all colleges had given out their results. Which was understandable as they were overwhelmed during implementing a new system which has now developed fantastically. This challenge delayed the data collection process but we got around it by keeping in touch with all the respondents and asking them to complete the survey as soon as their results came out. We completed the data collection problem after this without a hitch.

When the data collection process had finished, we started with the data mining task. I did the preliminary data mining tasks of grouping data, handling the missing values, and identifying and handling outliers. Since the data was limited, we couldn’t possibly delete any records that contained missing data and outliers. Hence missing data handling techniques such as highest frequency value imputation, median imputation were used depending on the variable.

Further, I managed the data visualization task such as plotting scattergrams and correlation matrix for feature selection. I was also called upon by my teammates for some initial model implementations which could be then improved upon by them. I would like to mention that my teammates were a great help in all my tasks and would aid me if I ever needed them.

**Individual contribution to project report preparation:**

My work in the project report preparation was in abstract, introduction and related work sections. The abstract section gives a brief overview of the domain of the research itself. The introduction gently lets the reader know about the research and what they can expect from the project and about its main objective. The references section required me to read and understand the works that were in the same domain and then mention the most important and correlated ones in the report.

**Individual contribution for project presentation and demonstration**:

In the powerpoint presentation I have worked on understanding the problem, feature information, features collected, and the data analysis sections.

Full Signature of Supervisor:

…………………………..

(Prof. Dr. Suresh Chander Sathapaty )

Full signature of the student:

Riyan Pahuja

(Riyan Pahuja)