

ACKNOWLEDGEMENT

My sincere thanks to our internal project guide **Dr. Amit Thakkar, Prof. Purvi Prajapati, Prof. Chadni Shah, Prof. Jay Patel** for giving us valuable inputs and ideas right from the selection of topic for project till its successful completion. Without his constant encouragement we could not have been able to achieve what we have.

I thank **Dr. Parth Shah(HOD IT department)** for his ongoing support and encouragement in every aspect. Last but not the least, entire staff of Department of IT Engineering for guiding me thoughts and vision. The successful completion of my project would not have been possible without the dedicated support from all our mentors, family and friends.

It's my good fortune that we had support and well wishes of many. We are thankful to all and those names which have been forgotten to acknowledge here but contributions have not gone unnoticed.

With Sincere Regards,
Priya Nasit(16IT057)
Dhvani Raval(16IT106)

ABSTRACT

Student performance prediction is an area of concern for educational institutions. Machine Learning plays an important role in the business world and it helps to the educational institution to predict and make decisions related to the students' academic status. The existing system is a system which maintains the student information in the form of numerical values and it just stores and retrieve the information what it contains. So the system has no intelligence to analyze the data. The prediction with high accuracy in student's performance is beneficial as it helps in identifying the student with low academic achievements at the early stage of academics.

In this work, classifier techniques based on six representative learning algorithms, namely Naive Bayes, Logistic Regression, k-Nearest Neighbour, Support Vector Machine, Decision Tree classification, Random Forest classification and ANN. These six learning methods have been compared separately with respect to the training and test sets. Random Forest Classification is found to be the best classifier for predicting the student's result based on the marks obtained in the semester, Permanent City, Gender, Actual Cast Category, 12th Percentage, Board. We would also discuss how these machine learning models can help to improve an education system by considering the different factors in terms of accuracy, Specificity, Precision, Prevalence, Recall, F-Measure in results.

Keywords and terms: student performance, machine learning, naïve Bayes classification, Logistic Regression, k-Nearest Neighbour, Support Vector Machine, Decision Tree classification, Random Forest classification and ANN(Artificial Neural Network).

TABLE OF CONTENTS

Acknowledgement.....	I
Abstract.....	II
Chapter 1 Introduction.....	1
1.1 Project Overview.....	2
1.1.1 Machine Learning.....	3
1.2 Scope	3
1.3 Objective.....	3
1.4 Problem Statement and solution.....	4
Chapter 2 System Analysis	4
2.2 Tools & Technology.....	4
2.2.1 Software Requirements	4
2.2.2 Programming Language.....	4
Chapter 3 System Design.....	5
3.1 Work Flow	6
Chapter 4 Implementation.....	7
4.1 Methodology.....	7
4.2 Algorithms.....	7
4.2.1 Logistic Regression.....	7
4.2.2 Naïve Bayes Classifier.....	7
4.2.3 K Nearest Neighbor.....	7
4.2.4 Support Vector Machine.....	7
4.2.5 Decision Tree Classification	7
4.2.6 Random Forest Classification.....	8
4.2.7 ANN	8
4.3 Feature Engineering.....	9
4.3.1 Features Correlation.....	9
4.4 Evaluation Methods.....	10
4.5 Dataset Features.....	12
4.6 Coding Standards.....	12
4.7 Result.....	15
4.2.1 Random Forest Classification.....	15
4.2.2 Decision Tree Classification.....	16
4.2.3 K Nearest Neighbor.....	16

4.2.4 Support Vector Machine.....	17
4.2.5 Naïve Bayes	18
4.2.6 Logistic Regression.....	19
4.2.6 ANN.....	19
4.8 Analysis.....	20
Chapter 5 Future Enhancement.....	21
Chapter 6 Conclusion.....	22
References.....	23

LIST OF FIGURES

3.1 System Design	5
4.1.1 Work Flow	6
4.4.1 Two Class Confusion Matrix	10
4.7.1.1 Confusion Matrix Of Random Forest	15
4.7.2.1 Confusion Matrix Of Decision Tree	16
4.7.3.1 Confusion Matrix Of KNN.....	16
4.7.4.1 Confusion Matrix Of SVM(poly).....	17
4.7.4.3 Confusion Matrix Of SVM(RBF).....	17
4.7.5.1 Confusion Matrix Of Naïve Bayes	18
4.7.6.1 Confusion Matrix Of Logistic Regression.....	19
4.7.6.1 Result Of ANN.....	19

LIST OF TABLES

4.4.2 1 Calculation Table	11
4.4.3 1 Three Class Confusion Matrix.....	11
4.5.1 Data Features.....	12
4.7.1.2 Result Of Random Forest.....	15
4.7.2.2 Result Of Decision Tree	16
4.7.3.2 Result Of KNN.....	16
4.7.4.2 Result Of SVM(poly).....	17
4.7.4.4 Result Of SVM(RBF).....	17
4.7.5.2 Result Of Naïve Bayes	18
4.7.6.2 Result Of Logistic Regression	19
4.8.1 Analysis Table	20
4.8.2 Comparison Of All Classification Algorithms	20