# STUDENT PERFORMANCE EVALUATION USING PREDICTIVE ANALYSIS

**Manju M N[1], Mayur G Kodekal[1], Shachi D Sharma[1], Chandan J[2]**

[1]*Dept. of ISE, The National Institute of Engineering, Mysore*
[2]*Asst. Professor, Dept. Of ISE, The National Institute of Engineering, Mysore*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *In this paper, we are going to do class result prediction using Machine Learning. The focus of this is to predict the student's result based on collecting data of each student in the university. Which is the part of machine learning and scientific computing theory and which is subfields of artificial intelligence[7]. Machine learning is one of the booming technology in current industry , where we can give solutions to the business problems that are under the requirements of the client. By collecting and analyze large amount of data and find out who are eligible for writing exams. Every time we must analyze correctly and find out the result in an accurate manner using best fit algorithms. The ability to predict a student's performance is based upon diverse factors like personal, social, psychological and other environmental variables and by consisting these factors a lot of institutions use this kind of prototype where we can judge a success rate of student.*

*Predicting student's performance can help identify the students who are at risk of failure and thus management can provide timely help and take essential steps to coach the students to improve his performance. To overcome this problem student result prediction model in the form of web service will be proposed here.*

***Key Words*: SQL server *2005/2008, Classification Prediction, Accuracy Score, Visualizations, Supervised learning, Machine Learning.***

## 1. INTRODUCTION

Machine learning is a sub-domain of computer science which evolved from the study of pattern recognition in data, and also from the computational learning theory in artificial intelligence[3]. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions. Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, etc.

Supervised learning or classification is the machine learning task of inferring a function from a label data.

In Supervised learning, we have a training set and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, which requires the model to generalize from the training set in a reasonable way.

In this paper we have made an attempt to predict results of a batch of students based on the previous performance using supervised machine learning algorithms such as "Naive Bayes", "KNN algorithm" and decision tree[1].

## 2. DATA COLLECTION

We have used data collected from a sql server2005. In sql server where we can get each and every characteristics of the student(collecting the data) ex: University serial number, how many hours a student has been studied, and a student should be regular to class or not and what about his interaction in the class, how student maintained his time management, and what about his extracurricular activities. And the main thing is collecting previous year results ex: 10th, 12th and previous sem marks data.

By applying these relational attributes of every student to the predict whether a student is excellent or good or bad based on applying classification algorithms like Naive Bayes and KNN algorithms and Decision Tree, which one is suited for us we can take that algorithm[4].

Based on accuracy score of each algorithm we used, which one we got highest accuracy i.e. of ex: 90% we can take as a best one. Our project is mainly based on data that is collected based on behavior of student. The goal of behavior is to collect data while exhibiting good behavior and then train a model correctly and test the result with predicting results if they result are matches then is best fit algorithm.

## 3. MODEL ARCHITECTURE:

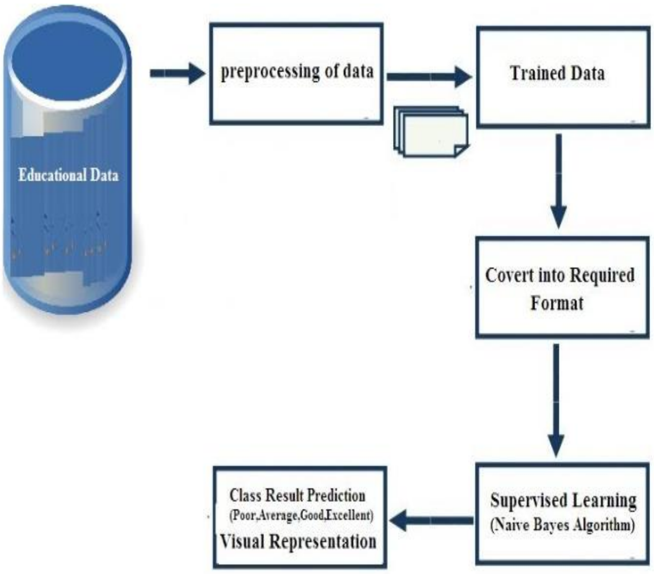

Educational Data
preprocessing of data
Trained Data
Covert into Required Format
Supervised Learning (Naive Bayes Algorithm)
Class Result Prediction (Poor,Average,Good,Excellent) Visual Representation

**Fig-1:** Architecture Diagram For How Result Prediction Is Done .

Preprocessing of data: check correct datasets to be loaded or not whether it is dbms connected sql file or .csv file, it needs to be checked.

If relational data is categorical variable, then we should apply LabelEncoder() class to decode into integer, why should we do this? In machine it only knows binary values(0 or 1) to execute every line of data[2]. It would be converted into machine language to execute the data to get the results.

Check whether we should apply feature scaling or not based on requirement. While coming to train and test the data in the relations based on how much data we can get. Suppose we get 1000 data entries; we need to filter 80% of the data will be training data i.e 800 data points for training and 20% of the data i.e 200 data points will be testing data. Normally training data will be more. Once testing data is got then we will fit the model and then predict with the testing set data.

## 4. METHODOLOGICAL APPROACH

Following approach can be considered here:

### 4.1Naive Bayes classifier

A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. It is used by administrator to predict pass percentage and fail percentage of the overall students.

Algorithm is a simple probabilistic classifier that calculates a set of probability by counting the frequency and combination of values in a given dataset with independent assumptions between predictors[6].

Algorithm assumes that all attributes are independent given the value of the class variable. This property holds good for problem considered, since the score of the student in each subject is independent, though it could be related with similar subjects. Therefore, this classifier is very effective for this problem.

**Reasons for Selecting Naive Bayes:**

1. works fine for n number of parameters.
2. works fine for small datasets as well as huge datasets.
3. more efficient than other algorithms.
4. more accurate results compared to other algorithms

**Steps involved:**

Step 1: Scan the dataset (storage servers)
Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]
Step 3: Apply the formulae

P (attribute value(ai)/subjectvaluevj) = (n_c + mp)/(n+m)

Where:

- n = the number of training examples for which v = vj
- nc = number of examples for which v = vj and a = ai
- p = 1/number of subject values
- m = the equivalent sample size [number of attributes]

Step 4: Multiply the probabilities by p
Step 5: Compare the values and classify the attribute values to one of the predefined sets of class.

### 4.2 KNN classifier

Basis of K nearest neighbour is categorization of unknown record or data point in which its class is already known.
Nearest neighbour is calculated according to k-value that determines the number of nearest neighbours to be considered.
Euclidean distance is considered as a measure to calculate the distance of the test label with the centres and assign the class label to the test sample by majority nearest neighbour. The test data sample will be assigned the class label by determining which centre is the nearest one.

**KNN Algorithm can be expressed as:**

1. Determine value of k, which is positive integer typically small with best choice value depending on dataset used in study.
2. Calculate Euclidean distance.
3. Determine k-min distance neighbours of testing data from trained data on a distance neighbours metrics.
4. Gather category Y values of nearest neighbours.
5. Use simple majority of nearest neighbours to predict value of query value given.

KNN is simplest of all machine learning algorithms. It is easy to implement and understand and also works fine in a sample of many class labels. It is memory inefficient in some cases and can be improved according to their distance from new sample record. This limitation of memory can be overcome by reducing the size of dataset which can be done by eliminating repeated patterns and removing some records which doesn't affect the result.

### 4.3 Decision Tree

Decision tree depicts the classification of a dataset into groups. It trains a model based on a sample of known observation as input and known responses as output. It is a non-parametric method, assuming no predefined data probability distribution occurs. Decision tree is robust and also efficient in classifying both categorical and numerical variable hence output is easily interpreted. This algorithm can provide information about attribute which are of predictive important i.e. closer the attribute is to the root, the more important it tends to be.

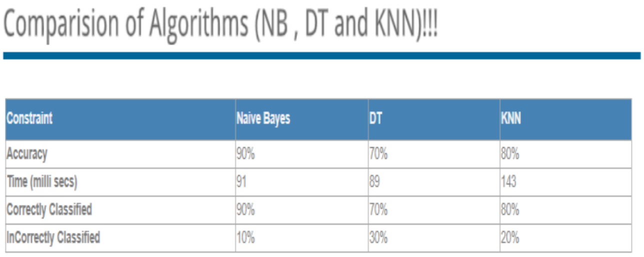
**Steps involved:**

1. Scan the data set.
2. For each attribute, calculate conccurrences(gain).
3. Let a_best be the attribute of highest gain (highest count).
4. Create a decision node on a best retrieval of records(nodes) where the attribute values match with a_best.
5. Recur on the sub-lists and calculate the count of results(outcomes) termed as sub nodes. Based on highest count we classify the new node or record.

## 5. RESULTS

The models created by each of the algorithms are used to create an output for each of the student. This output is then compared with the actual results of that semester, and the accuracy is determined.

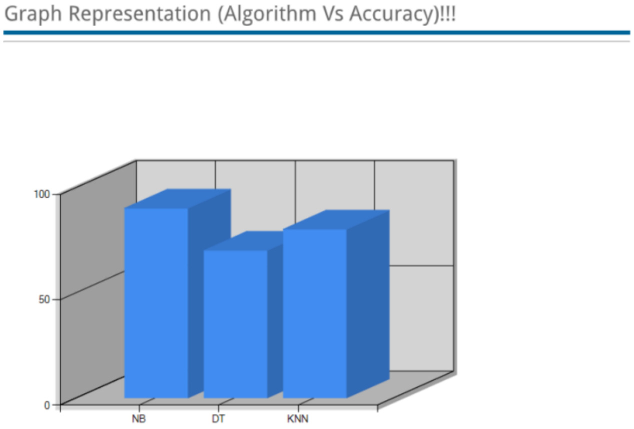
The output each of the algorithm with accuracy results is tabulated below.



Comparision of Algorithms (NB , DT and KNN)!!!

| Constraint | Naive Bayes | DT | KNN |
|---|---|---|---|
| Accuracy | 90% | 70% | 80% |
| Time (milli secs) | 91 | 89 | 143 |
| Correctly Classified | 90% | 70% | 80% |
| InCorrectly Classified | 10% | 30% | 20% |

**Fig-2**: Accuracy Results For All Three Algorithms.



Graph Representation (Algorithm Vs Accuracy)!!!

**Fig-3**: Visualization of Each Algorithm with Accuracy.

## 6. CONCLUSIONS

The academic achievement of every student is very important in India and it is a turning point of the life. But there are determinants like demographic, academic and socio-economic factors of students that restrict the student's performance.

This necessitates the need for some forecasting systems to Predict the academic performance of students. Student academic performance pre-dominatly by means of different feature selection of data mining. Proposition of suitable performance prediction model with reasonable predictive accuracy have been framed as the main objectives for the present investigation. As the present investigation pertains to the model prediction of the performance of students in higher education, it will be useful to parents, educators, academics, and policy makers for taking appropriate decisions for the development of student community in the highly competitive world to-day. But also, on the prediction of the academic performance of the students using different classification algorithm in data mining[5].

From the review of literature, it was found that a number of studies identified different factors that could influence the academic achievement of students. It was also found that various prediction models were proposed in different contexts.

This led to allowing learning algorithms to operate faster and more effectively to produce better classification results with high predictive accuracy.

## 7. REFERENCES

[1] Sunil Ray, Common Machine Learning Algorithms. Analytics vidya. https://www.analyticsvidhya.com/blog/2015/08/commonmachine- learning-algorithms/

[2] Micheal Bowles, Machine Learning in Python: Essential Techniques for Predictive Analysis. John Wiley & Sons, Inc. 2015

[3] Jack Clark. Google turning its lucrative web search over to AI machines, 2015. www.bloomberg.com/news/articles/2015-10-26/googleturning- its-lucrative-web-search-over to-a machine. 1212

[4] Aggelos Konstantinos Katsaggelos, Jeremy watt, and Reza Borhani, Machine Learning Refined: Foundations, Algorithms.

[5] S. Ren, K. He, R.Girishick and J.son, "Faster R- CNN: with real world examples in Machine Learning.

[6] *Naive Bayes classifier from Wikipedia the free encyclopedia*, [online] Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[7] *An executive's guide to machine learning*.