RESEARCH ARTICLE OPEN ACCESS

Result Prediction Using Different Machine Learning Algorithms

Ashna KJ, Aswathy Vs, Sruthi Rajeev, Abhishek A A
Department of Computer Science and Engineering
Toc H Institute of Science and Technology, Arakkunam
Kerala - India

ABSTRACT

Machine learning (ML) is one of the intelligent methodologies that have shown promising results in the domains of classification and prediction. One of the expanding areas necessitating good predictive accuracy is result prediction, our project is to implement software that will predict the result of students with maximum efficiency. For this we are employing machine learning technology with ensemble method. Machine Learning algorithms are employed in order to implement several features. Along with result prediction we are also planning to do result analysis with help of Data Visualization technique. Also, we are planning to compare efficiency of various ML algorithms and choosing the best algorithm that yields to better result and maximum efficiency. The parameters that are considered for result prediction are attendance, internal marks and mark of previous semester. The decision tree, Naive Baye's and Support Vector Machine algorithms are used for result prediction with maximum efficiency. Ensemble methods are machine learning technique that combines several base models in order to produce one optimal predictive model. Data Visualization helps to analyze performance like did he/she improved her performance compared to previous semester, rise and fall in performance, percentage increase or decrease in CGPA and so on. We can see graphically the pass percentage so it will be easier to analyze performance of each and every student and also, we can continuously monitor the performance of student.

Keywords: Machine Learning, Result Prediction, Result Analysis, Naïve Baye's algorithm, Support Vector Machine Algorithm, Data Visualization

I. INTRODUCTION

Computers do not learn as well as people do, but many machine- learning algorithms have been found that are effective for some types of learning tasks. They are especially useful in poorly understood domains where humans might not have the knowledge needed to develop effective knowledge engineering algorithms. Generally, Machine Learning (ML) explores algorithms that reason from externally supplied instances (input set) to produce general hypotheses, which will make predictions about future instances. The externally supplied instances are usually referred to as training set. To induce a hypothesis from a given training set, a learning system needs to make assumptions about the hypothesis to be learned. A learning system without any assumption cannot generate a useful hypothesis since the number of hypotheses that are consistent with the training set is usually huge.

The aim is to predict student performance. Data about students is used to create a model that can predict whether the student is successful or not, based on other properties. First, the training data set is taken as input. There are two different data sets, containing different types of information. These data sets are in tabular format, where each row represents a student and each column, or variable, contains certain information about a student, such as previous semester marks, attendance, assignment marks and current semester internal marks. In addition, a column representing the success of the student is used as the variable that the algorithms trying to predict. The algorithm creates a model, which is a function that outputs success or failure of the student, using other variables as input. The parameters that are considered for result prediction are attendance, internal marks and mark of previous semester. The decision tree, Naive Baye's and Support Vector Machine algorithms are used for result prediction with maximum efficiency. Ensemble methods are machine learning technique that combines several base models in order to produce one optimal predictive model. Data Visualization helps to analyze performance like did he/she improved her performance compared to previous semester, rise and fall in performance, percentage increase or decrease in CGPA and so on. We can see graphically the pass percentage so it will be easier to analyze performance of each and every student and also, we can continuously monitor the performance.

II. LITERATURE SURVEY

Analysis Of Student Performance Based On Classification And Mapreduce Approach In Bigdata

To identify students with educational risks and develop predictive models to predict student's performance, and to help identify student's final outcome. Student's academic performance will be related to many features. The scope of this study is limited to the study of progress in learning. The analysis of student performance consists of two functions:

i) Students who are at the academic risk
ii) Predict the student performance. iii)Students who are at the academic risk

Data is collected from different educational sectors and the lowest six subject marks are selected, it requires proper method of extracting knowledge from large repositories for better decision making. This presents an important challenge to organizations that use data management mechanisms to analyze, process and store large data sets. Therefore, a new model called "big data analytics" needs to be defined to reevaluate the existing system and manage and process large amounts of student data. Here we implement a new component of big data analytics called "learning analytics". It refers to performing various data produced by students in order to evaluate learning progress like, predict future performance, and identify probable problems.

The first step of learning analytics is to collect data from different educational institutions, this step is very difficult, because of the real information about students' details are fetched from the respective institutions. Data collected from various institutes and since it contains a large amount of data those data sets are considered as the big data. Here MongoDB framework can be used. It's very cost-effective and it generates

uicker data processing. MongoDB is free and open-source cross-platform, document-oriented database program. Classified as a NoSQL database program. MongoDB can be used as a file system with load balancing and data replication features over multiple machines for storing files. JavaScript can be used in queries, aggregation functions (such as Map Reduce), and sent directly to the database to be executed. B. Predict Student Performance For educational institutes, the student's result analysis and prediction are very important because of their quality of education.

The main process is the ability to meet the needs of students. Analyzing the past performance of these students will give a better understanding. It should change their future results. This can be done well using the concept of predictive analytic and performance analytics. Predictive analytics includes various statistical techniques from modeling, machine learning, and data mining to analyze current and past data to predict the future. In the prediction methods, the main process is collecting data. Here the data collected from different educational institutes will include students' information and split the data into sample data and test data. The predictive model is developed with sample data with the help of statistical methods and apply the model to test data to predict the result.

Advantages of proposed system

i) The time complexity of the system is reduced.
ii) It supports for big data. So, it will be applicable in several real applications.
iii) It also supports the online data.
iv) The classification accuracy is increased even for big data. v)Support redundant data.
vi) Low computational complexity comparing with the existing work
vii) The system efficiency is improved.

III. RESULT PREDICTION USING DATA MINING TECHNIQUE

In Indian education system checking student's performance is a very essential in higher education. But we don't have any fixed criteria to evaluate the student performance. Some institutions student performance can be observed by using internal assessment and co-curriculum. In the Indian context, an institution with the higher degree of reputation using the good academic record as its basic criteria for their admissions. There are lots of definitions of student academic performance prediction should be given in the literature. Different authors are using different student factors/attributes for analyzing student performance. Most of the author used CGPA, Internal assessment, External assessment, Examination final score and extra co-circular activities of the student as prediction criteria. Most of the Indian institution and universities using final examination grade of the student as the student academic performance criteria.

The final grades of any student depend on different attributes like internal assessment, external assessment, laboratory file work and viva-voce, sessional test. The performance of the student depends upon how many grades a student score in the final examination. In Educational data mining field, making a prediction about student academic performance is usually done. To build a predictive modelling we need to take different data mining techniques into consideration like classification, clustering association rule mining and regression analysis. In almost every research paper, the only classification algorithm is taken into consideration for predicting student academic performance.

There are so many classification techniques available for prediction but we are taking into consideration only decision tree, Naive Bayes, Support Vector Machine (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbor, SMO, Linear Regression, Random Forest, Random Tree, REPTree, LAD Tree, J48 etc gave a brief finding of different research papers with their author's name, main attributes helpful for prediction accuracy with different data mining algorithm used. Mostly used data mining algorithm for SAP is Decision Tree (DT), Naive Bayes (NB), Artificial Neural Networks (ANN), Rule-based (RB) and K-Nearest Neighbor (KNN). In Decision tree algorithm the maximum and minimum accuracy for predicting student's academic performance are 99.9% and 66.8% respectively. To find the maximum prediction accuracy Maria Goga, Shade Kuyoro and Nicolae Goga used the combination of student's attribute like family, PEP, EES, end of first session result. In Naive Bayes algorithm, the maximum and minimum accuracy for predicting student's academic performance are 100% and 63.3% respectively. Maria Koutina et. al. used the different combination of student's attribute like Gender, Age, Marital Status, Number of children, Occupation, Job associated with the computer.

Advantages of the proposed system

i) It is helpful to predict future trends

Most of the working nature of the data mining systems carries on all the informational factors of the elements and their structure. One of the common benefits that can be derived with these data mining systems is that they can be helpful while predicting future trends. And that is quite possible with the help of technology and behavioral changes adopted by the people.

ii)Helps in decision making

There are some people who make use of these data mining techniques to help them with some kind of decision making. Nowadays, all the information about anything can be determined easily with the help of technology and similarly, with the help of such technology one can make a precise decision about something unknown and unexpected.

iii) Quick fraud detection

Most parts of the data mining process is basically from information gathered with the help of marketing analysis. With the help of such marketing analysis one can also find out those fraudulent acts and products available in the market. Moreover, with the help of it one can understand the importance of accurate information.

Limitations of proposed system

i) It violates user privacy:

It is a known fact that data mining collects information about people using some market base techniques and information technology. And these data mining process involves several numbers of factors. But while involving those factors, data mining system violates the privacy of its user and that is why it lacks in the matters of safety and security of its users. Eventually, it creates Mis-communication between people.

ii)Additional irrelevant information

The main functions of the data mining systems create a relevant space for beneficial information. But the main problem with the information collection is that there is a possibility that the collection of information process can be little overwhelming for all. Therefore, it is very much essential to maintain a minimum level of limit for all the data mining techniques.

iii)Misuse of information

As it has been explained earlier that in the data mining system the possibility of safety and security measure are really minimal. And that is why some can misuse this information to harm others in their own way. Therefore, the data mining system needs to change its course of working so that it can reduce the ratio of misuse of information through the mining process.

iv)An accuracy of data

Most of the time while collecting information about certain elements one used to seek the help from their clients, but nowadays everything has changed. And now the process of information collection made things easy with the mining technology and their methods. One of the most possible limitation of this data mining system is that it can provide accuracy of data with its own limits. Finally the bottom line is that all the techniques, methods and data mining systems help in discovery of new creative things. And at the end of this discussion about the data mining methodology, one can clearly understand the feature, elements, purpose, characteristics and benefits with its own limitations.

Predicting students' performance in distance learning using machine learning techniques

The ability of prediction of a student's performance could be useful in a great number of different ways associated with university-level distance learning. Students' key demographic characteristics and their marks in a few written assignments can constitute the training set for a supervised machine learning algorithm. The learning algorithm could then be able to predict the performance of new students thus becoming a useful tool for identifying predicted poor performers. The scope of this work is to compare some of the state of art learning algorithms. Two experiments have been conducted with six algorithms, which were trained using data sets provided by the Hellenic Open University. Among other significant conclusions it was found that the Naive Bayes algorithm is the most appropriate to be used for the construction of a software support tool, has more than satisfactory accuracy, its overall sensitivity is extremely satisfactory and is the easiest algorithm to implement this method uses existing ML techniques in order to predict the students' performance in a distance learning system. It compares some of the state of art learning algorithms to find out which algorithm is more appropriate not only to predict student's performance accurately but also to be used as an educational supporting tool for tutors. To the best of our knowledge, there is no similar publication in the literature. Whittington (1995) studied only the factors that impact on the success of distance education students of the University of the West Indies. For the purpose of our study the 'informatics' course of the Hellenic Open University (HOU) provided the training set for the ML algorithms.

The basic educational unit at HOU is the module and a student may register with up to three modules per year. The 'informatics' course is composed of 12 module and that leads to Bachelor Degree. The total number of registered students with the course of informatics in the academic year 2000-1 was 510. Of those students 498 (97.7%) selected the module Introduction to Informatics (INF10). This fact enabled the authors to focus on INF10 and collect data only from the tutors involved in this module. The tutor in a distance-learning course has a specific role. Despite the distance between him/her and his/her students, he/she has to teach, evaluate and continuously support them. The communication between them by post, telephone, e-mail, through the written assignments or at optional consulting meetings helps the tutor to respond to this complex role. In all circumstances, the tutor should promptly solve students' educational problems, discuss in a friendly way the issues that distract them, instruct their study, but most of all encourage them to continue their studies, understanding their difficulties and effectively supporting them.

Furthermore, tutors have to give them marks, comments and advice on the written assignments and they have to organize and carry out the face-to-face consulting meetings. For all above mentioned reasons, it is important for the tutors to be able to recognize and locate students with high probability of poor performance (students at risk) in order to take precautions and be better prepared to face such cases. Regarding the INF10 module of HOU during an academic year the students have to hand in 4 written assignments, participate in 4 optional face to face meetings with their tutor and sit for final examinations after an 11-month period. The students' marking system in Hellenic Universities is the 10-grade system. A student with a mark >=5 'passes' a lesson or a module while a student with a mark <5 'fails' to complete a lesson or a module. Key demographic characteristics of students (such as age, sex, residence etc) and their marks in written assignments constituted the initial training set for a supervised learning algorithm in order to predict if a certain student will eventually pass or not a specific module. A total of 354 instances (student's records) have been collected out of the 498 who had registered with INF10. Two separate experiments were conducted. The first experiment used the entire set of 354 instances for all algorithms while the second experiment used only a small set of 28 instances corresponding to the number of students in a tutor's class.

The application of Machine Learning Techniques in predicting students' performance proved to be useful for identifying poor performers and it can enable tutors to take remedial measures at an earlier stage, even from the beginning of an academic year using only students' demographic data, in order to provide additional help to the groups at risk. The probability of more accurate diagnosis of students' performance is increased as new curriculum data has entered during the academic year, offering the tutors more effective results. Some very basic Machine Learning definitions are given in discusses the data collection process, the attribute selection, the algorithm selection and the research design. Section 4 presents the experiment results for both experiments and all six algorithms and at the same time compares these results using as criteria the overall sensitivity, accuracy and specificity of the algorithms. Finally, section 5 discusses the conclusions and some future research directions.

Description of the used machine learning techniques

Murthy (1998) provides a recent overview of existing work in decision trees. Decision trees are trees that classify instances by sorting them based on attribute values. Each node in a decision tree represents an attribute in an instance to be classified, and each branch represents a value that the node can take. Instances are classified starting at the root node and sorting them based on their attribute values. The main advantage of decision trees in particular and hierarchical methods in general, is that they divide the classification problem into a sequence of sub problems which are, in principle, simpler to solve than the original problem. The attribute that best divides the training data would be the root node of the tree. The algorithm is then repeated on each partition of the divided data, creating sub trees until the training data are divided into subsets of the same class. Artificial Neural Networks (ANNs) are another method of inductive learning based on computational models of biological neurons and networks of neurons as found in the central nervous system of humans (Mitchell, 1997). A multi-layer neural network consists of large number of units (neurons) joined together in a pattern of connections. Units in a net are usually segregated into three classes: input units, which receive information to be processed, output units where the results of the processing are found, and units in between called hidden units. Classification with a neural network takes place in two distinct phases. First, the network is trained on a set of paired data to determine the input-output mapping. The weights of the connections between neurons are then fixed and the network is used to determine the classifications of a new set of data.

Naive Bayes classifier is the simplest form of Bayesian network (Domingos and Pazzani, 1997). This algorithm captures the assumption that every attribute is independent from the rest of the attributes, given the state of the class tribute. Naive Bayes classifiers operate on data sets where each example x consists of attribute values <a1, a2 ... ai> and the target function f(x) can take on any value from a pre-defined finite set V= (v1, v2 ... vj). Instance-based learning algorithms belong in the category of lazy-learning algorithms (Mitchell, 1997), as they defer in the induction or generalization process until classification is performed. One of the most straightforward instance based learning algorithms is the nearest neighbour algorithm (Aha, 1997).

K-Nearest Neighbour (KNN) is based on the principal that the instances within a data set will generally exist in close proximity with other instances that have similar properties. If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbors. Logistic regression analysis (Long, 1997) extends the techniques of multiple regression analysis to research situations in which the outcome variable (class) is categorical. The relationship between the classifier and attributes. The dependent variable (class) in logistic regression is binary, that is, the dependent variable can take the value 1 with a probability of success pi, or the value 0 with probability of failure 1- pi. Comparing these two probabilities, the larger probability indicates the class label value that is more likely to be the actual label. The SVM technique revolves around the notion of a ‘margin’, either side of a hyperplane that separates two data classes. Maximizing the margin, and thereby creating the largest possible distance between the separating hyperplane and the instances on either side of it, is proven to reduce an upper bound on the expected generalization error (Burges, 1998). Nevertheless, most real- world problems involve non-separable data for which no hyperplane exists that successfully separates the positive from negative instances in the training set. One solution to the inseparability problem is to map the data into a higher-dimensional space and define a separating hyperplane. This higher-dimensional space is called the feature space, as opposed to the input space occupied by the training instances. Generally, with an appropriately chosen feature space of sufficient dimensionality, any consistent training set can be made separable.

Analysis

This project aims to fill the gap between empirical prediction of student performance and the existing ML techniques. To this end, six ML algorithms have been trained and found to be useful tools for identifying predicted poor performers in an open and distance learning environment. With the help of machine learning methods the tutors are in a position to know which of their students will complete a module or a course with sufficiently accurate precision. This precision reaches the 62% in the initial forecasts, which are based on demographic data of the students and exceeds the 82% before the final examinations. Our data set is from the module ‘Introduction in informatics but most of the conclusions are wide-ranging and present interest for the majority of programs of study of the Hellenic Open University. It would be interesting to compare our results with those from other open and distance learning programs offered by other open Universities. So far, however, we have not been able to locate such results.

Two experiments were conducted using data sets of 354 and 28 instances respectively. The above accuracy was the result of the first experiment with the large data set; however, the overall accuracy of the second experiment for all algorithms was less satisfactory than the accuracy in the 1st experiment.

The 28 instances are probably few if we want more accurate precision. After a number of experiments with different number of instances as training set, it seems that at least 70 instances are needed for a better predictive accuracy (70.51% average prediction accuracy for the Naive Bayes algorithm).Besides the overall accuracy of the algorithms, the differences between sensitivity and specificity are quite reasonable since ‘pass’ represents students who completed the INF10 module getting a mark of 5 or more in the final test, while ‘fail’ represents students that suspended their studies during the academic year (due to personal or professional reasons or due to inability to hand in 2 of the written assignments), as well as students who did not show up in the final examination. Furthermore, ‘fail’ also represents students who sit for the final examination and get a mark less than 5. Furthermore, the analysis of the experiments and the comparison of the six algorithms has demonstrated sufficient evidence that the Naive Bayes algorithm is the most appropriate to be used for the construction of a software support tool. The overall accuracy of the Naive Bayes algorithm was more than satisfactory (72.48% for the first experiment) and the overall sensitivity was extremely satisfactory (78.00% for the first experiment). Moreover, the Naive Bayes algorithm is the easiest to implement among the tested algorithms.

Future work we intend to study if the use of more sophisticated approaches for discretization of the marks of WRIs such as the one suggested by Fayyad and Irani (1993) could increase the classification accuracy. In addition, because of FTOFs did not add accuracy and the run time of inductive algorithms grows with the number of attributes, we will examine if the selection of a subset of attributes could be useful. Finally, since with the present work we can only predict if a student passes the module or not we intend to try to use regression methods in order to predict the student’s marks instead.

IV. METHODOLOGY

The main aim of our project is to produce a model with better efficiency and to ensure correctness. We try to implement the system in such a way that error rate is very less and also to make the model flexible to make necessary corrections in case of any error identified. Also we use ensembling method to produce improved machine learning results .Thus ensuring the correctness of the model.

System Study

System study of a project includes system analysis and system design. System analysis and design includes all activities, which help the transformation of requirement specification into implementation. Requirement specifications specify all functional and non-functional expectations from the software. These requirement specifications come in the shape of human readable and understandable documents, to which a computer has nothing to do. System analysis and design is the intermediate stage, which helps human-readable requirements to be transformed into actual code. System design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. The system includes Architectural design, logical design and physical design.

The various modules in our system is user, training, testing, ensembling and data visualization. The hardware requirements of our stem is a processor with 4 GB RAM and the software requirements are Python is the preferred language for coding and for data visualization we use math plot lib. The python libraries are available in module named flask. PHP is used for server-side scripting. Software named pandas is used for developing the artificial neural networks (ANN). JSON modules are also linked with python libraries. For database creation My SQL is used. Prototyping model is followed.

V.DESIGN AND IMPLEMENTATION

Training

Here we provide the data set which contains internal marks, assignment marks, results of previous semester and attendance. We train our model according to our input data by supervised learning technique. Here the output/result is already provided and is known. And we have to check whether the model is giving correct result based on the training data. For example we have to predict the semester 5 result which is already known. So we have results from semester 1 to 5 and the results from semester 1 to 4 is used for training and 5 is used for validation and hence we can check the correctness.

Testing

Testing is done based on various machine learning algorithms. This module is intented to check and ensure the correctness of the model that is how correct the prediction was and the error rate is also calculated. Modifications are done based on the error rates. Error comparisons with using various algorithms is done in this module. For example we have to check the semester 6 result which is already known by us but not provided in training set. We provide results till 5th semester only. So to do this we’ll give results from semester 1 to 4 as training and 5th semester result will be used for validation. Here system will predict a whether a student passed or failed. So classification task is carried out and classification algorithm is employed.

Data Visualization

As the name suggests function of this module is to represent the data graphically. This module actually performs the result analysis. The data is represented using bar graphs, histograms, pie chart and so on. It represents large number of data in a small space and make large data set coherent. Thus we can compare different pieces of data irrespective of dimension. In this project we use the tool math plot lib for data visualization purpose.

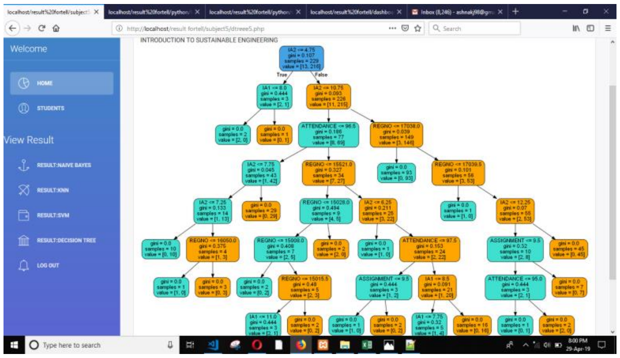


Fig 4.1 A Sample Visualization

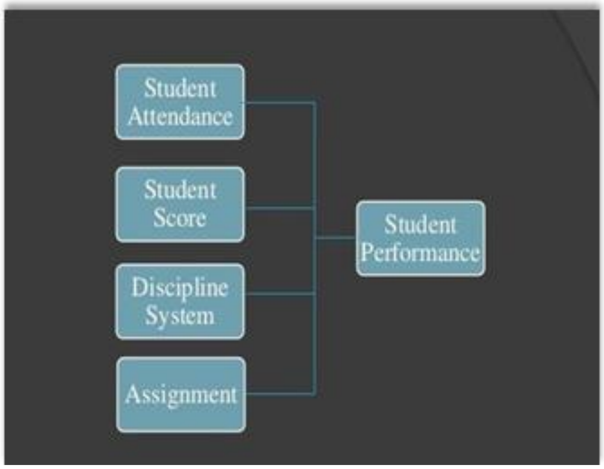


Fig 4.2 block diagram

Algorithms used

Naïve Baye's Algorithm

Naive Bayes is the classification machine learning algorithm that relies on the Bayes Theorem. It can be used for both binary and multi-class classification problems. The main point relies on the idea of treating each feature independently. Naive Bayes method evaluates the probability of each feature independently, regardless of any correlations, and makes the prediction based on the Bayes Theorem. That is why this method is called "naive", in real-world problems features often have some level of correlation between each other.

To understand the algorithm of Naive Bayes, the concepts of class probabilities and conditional probabilities should be introduced first.

i) Class Probability is a probability of a class in the dataset. In other words, if we select a random item from the dataset, this is the probability of it belonging to a certain class.

$$P(C) = \frac{count(instances\ in\ C)}{count(instances\ in\ Ntotal)}$$

ii) Conditional Probability is the probability of the feature value given the class.

$$P(V|C) = \frac{count(instances\ with\ V\ and\ C)}{count(instances\ with\ V)}$$

iii) Given the probabilities, we can calculate the probability of the instance belonging to a class and therefore make decisions using the Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

iv) Probabilities of the item belonging to all classes are compared and the class with the highest probability if selected as a result.

The advantages of using this method include its simplicity and easiness of understanding. In addition to that, it performs well on the data sets with irrelevant features, since the probabilities of them contributing to the output are low. Therefore they are not taken into account when making predictions. Moreover, this algorithm usually results in a good performance in terms of consumed resources, since it only needs to calculate the probabilities of the features and classes, there is no need to find any coefficients like in other algorithms. As already mentioned, its main drawback is that each feature is treated independently, although in most cases this cannot be true.

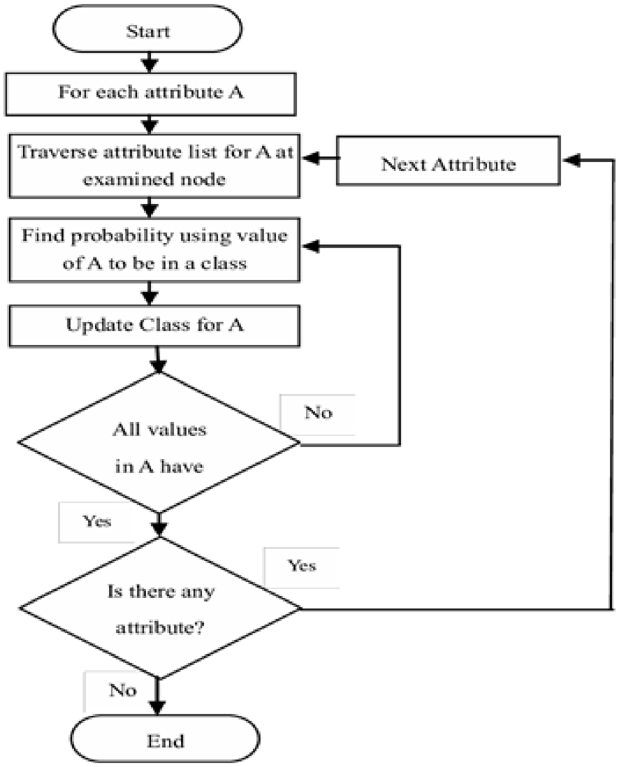


Fig 4.3 Flow Chart of Naïve Baye's

Support Vector Machine(SVM)

Support Vector Machines (SVM) is another machine learning algorithm that is generally used for classification problems. The main idea relies on finding such a hyperplane that would separate the classes in the best way. The term 'support vectors' refers to the points lying closest to the hyperplane, that would change the hyperplane position if removed. The distance between the support vector and the hyperplane is referred to as margin.
Intuitively, understand that the further from the hyperplane our classes lie, the more accurate predictions we can make. That is why, although multiple hyperplanes can be found per problem, the goal of the SVM algorithm is to find such a hyperplane that would result in the maximum margins. On Figure, there is a dataset of two classes. Therefore, the problem lies in a two- dimensional space, and a hyperplane is represented as a line. In general, hyperplane can take as many dimensions as we want.

The common algorithm for decision trees is ID3 (Iterative Dichotomiser 3). It relies on the concepts of the Entropy and Information Gain. Entropy here refers to the level of uncertainty in the data content. For example, the entropy of the coin toss would be indefinite, since there is no way to be sure in the result. Contrarily, a coin toss of the coin with two heads on both sides would result in zero entropy, since we can predict the outcome with 100% probability before each toss. (Mitchell 1997).
In simple words, the ID3 algorithm can be described as follows: starting from the root node, at each stage we want to partition the data into homogenous (similar in their structure) dataset. More specifically, we want to find the attribute that would result in the highest information gain, i.e. return the most homogenous branches.
Algorithm of Decision Tree
1) Calculate the entropy of the target.

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

margin
Algorithm

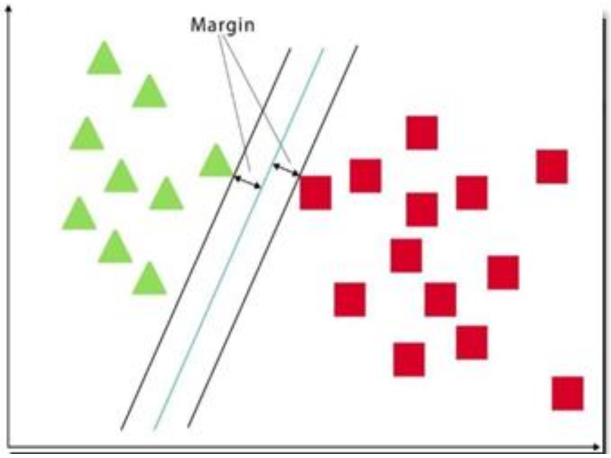


Fig 4.4 Graph of SVM
1.) We define X and Y as the input and output sets respectively. (□1,1),…,(□□,□□) is the training set.
2) Given x, we want to be able to predict y. We can refer to this problem as to learning the classifier y=f(x, a), where a is the parameter of the classification function.
3) F(x, a) can be learned by minimizing the training error of the function that learns on training data. Here, L is the loss function, and □□□□ is referred to as empirical risk.

$$R_{emp}(a) = \frac{1}{m}\sum_{i=1}^{m} l(f(x_i,a),y_i) = Training\ Error$$

4) We are aiming at minimizing the overall risk, too. Here, P(x,
y) is the joint distribution function of x and y.

5) We want to minimize the Training Error +

Complexity term. So, we choose the set of hyperplanes, so f(x) = (w · x)+b:

$$\frac{1}{m}\sum_{i=1}^{m} l(w \cdot x_i + b, y_i) + ||w||^2 \text{ subject to } min_i|w \cdot x_i| = 1$$

SVMs are generally able to result in good accuracy, especially on "clean" datasets. Moreover, it is good with working with the high-dimensional datasets, also when the number of dimensions is higher than the number of the samples. However, for large datasets with a lot of noise or overlapping classes, it can be more effective. Also, with larger datasets training time can be high.

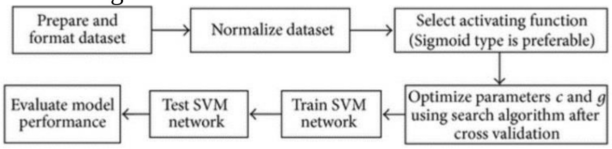

Prepare and format dataset
Normalize dataset
Select activating function (Sigmoid type is preferable)
Optimize parameters c and g using search algorithm after cross validation
Train SVM network
Test SVM network
Evaluate model performance

Fig 4.5 Flow chart of SVM

Decision Trees

2) Split the dataset and calculate the entropy of each branch. Then calculate the information gain of the split that is the differences in the initial entropy and the proportional sum of the entropies of the branches.

3) The attribute with the highest Gain value is selected as the decision node.

4) If one of the branches of the selected decision node has an entropy of 0, it becomes the leaf node. Other branches require further splitting.

5) The algorithm is run recursively until there is nothing to split anymore.

Decision tree method achieved its popularity because of its simplicity. It can deal well with large datasets and can handle the noise in the datasets very well. Another advantage is that unlike other algorithms, such as SVM or KNN, decision trees operate in a "white box", meaning that we can clearly see how the outcome is obtained and which decisions led to it.

These facts made it a popular solution for medical diagnosis, spam filtering, security screening and other fields. A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. ... Programmatically, they can be used to assign monetary/time or other values to possible outcomes so that decisions can be automated.Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. A decision tree is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. They can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically. A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a treelike shape. There are three different types of nodes: chance nodes, decision nodes, and end nodes. A chance node, represented by a circle, shows the probabilities of certain results. A decision node, represented by a square, shows a decision to be made, and an end node shows the final outcome of a decision path.

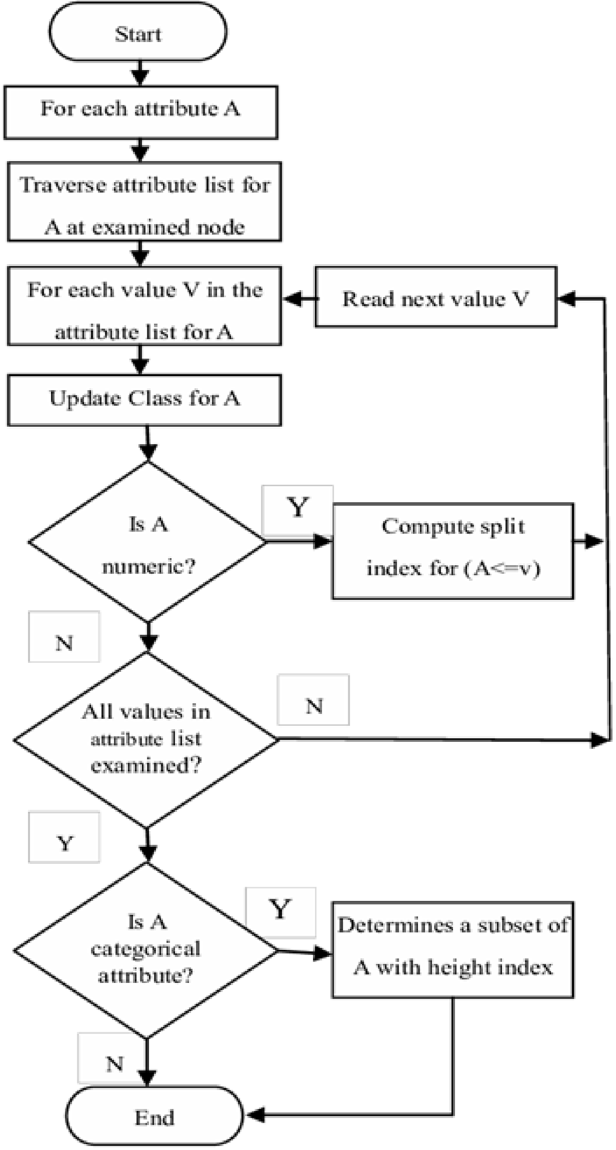
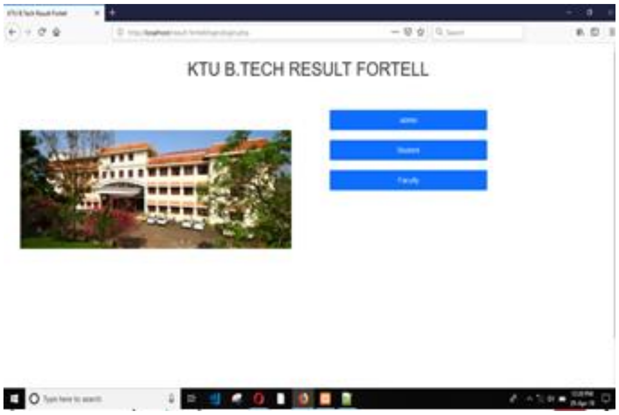

Start
For each attribute A
Traverse attribute list for A at examined node
For each value V in the attribute list for A
Read next value V
Update Class for A
Is A numeric?
Y
Compute split index for (A<=v)
N
All values in attribute list examined?
N
Y
Is A categorical attribute?
Y
Determines a subset of A with height index
N
End

Fig 4.5 Flow chart of decision tree



KTU B.TECH RESULT FORTELL

Fig 4.6 User Interface

VI. CONCLUSION

The success of machine learning in predicting student performance relies on the good use of the data and machine learning algorithms. Selecting the right machine learning method for the right problem is necessary to achieve the best results.However, the algorithm alone cannot provide the best prediction results. Results of both data sets show similarities and differences with their use in the original studies. In the first data set, similarity is that recall values were consistently higher than precision values. Difference was in the accuracy values. The accuracy reached in this thesis was higher than in the original research. This can be attributed to the difference in dependent variables. In this project the upcoming semester exam result of student is predicted based on certain features like attendance, internal marks and previous semester results. The proposed system will give the real time implementation of result prediction system we have used real time data in training and testing stages.Data Visualization was used to do the analysis graphically also it made comparison easier. Correlation was also done by using Data Visualization. This project can also be extended for similar applications in future.

Algorithm	Accuracy values Obtained(%)
KNN	80,94,91,98,96
Naïve Bayes	77,91,85,57,98,96
SVM	98,100
Decision Tree	98,100

Table 5.1 Accuracy obtained for each algorithms

The maximum efficiency obtained is for SVM and Decision Tree algorithm about 98% and least efficiency was obtained for Naïve Baye's algorithm

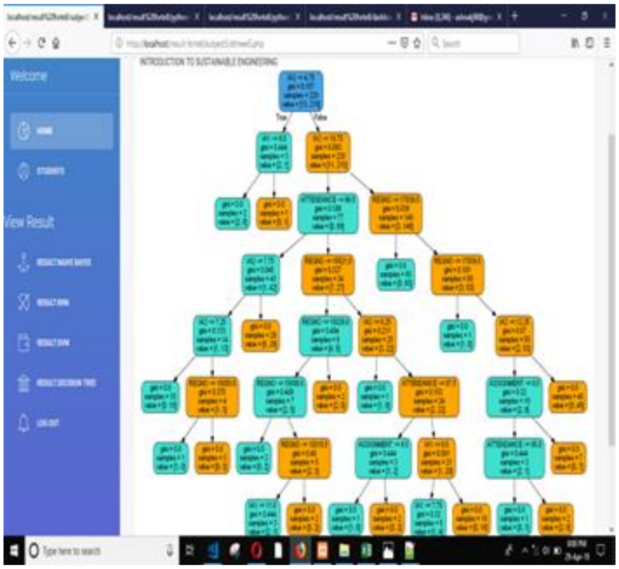
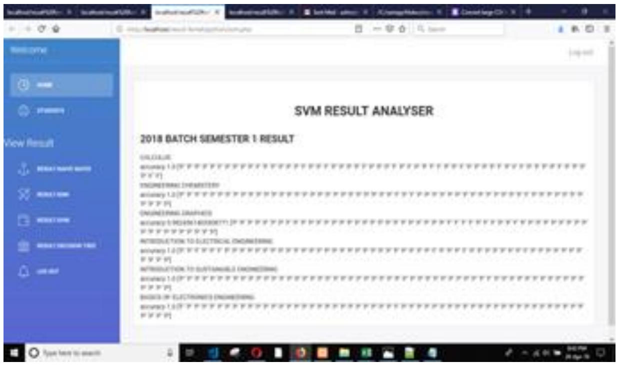


Fig 5.1 Visualization of Decision tree

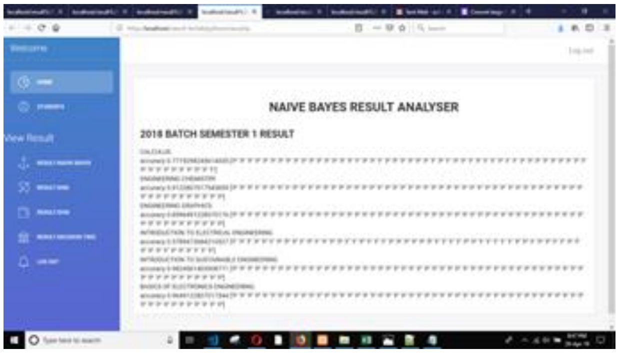


SVM RESULT ANALYSER

2018 BATCH SEMESTER 1 RESULT

CALCULUS
ENGINEERING CHEMISTRY
ENGINEERING GRAPHICS
INTRODUCTION TO ELECTRICAL ENGINEERING
INTRODUCTION TO SUSTAINABLE ENGINEERING
BASICS OF ELECTRONICS ENGINEERING

Fig 5.2 Result of SVM Algorithm

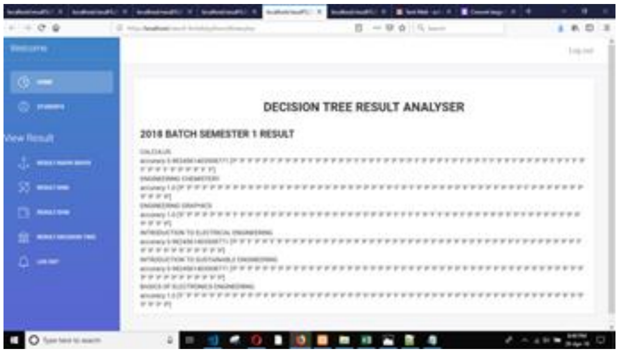


NAIVE BAYES RESULT ANALYSER

2018 BATCH SEMESTER 1 RESULT

CALCULUS
ENGINEERING CHEMISTRY
ENGINEERING GRAPHICS
INTRODUCTION TO ELECTRICAL ENGINEERING
INTRODUCTION TO SUSTAINABLE ENGINEERING
BASICS OF ELECTRONICS ENGINEERING

Fig 5.3 Result of Naïve Baye's Algorithm



DECISION TREE RESULT ANALYSER

2018 BATCH SEMESTER 1 RESULT

CALCULUS
accuracy [illegible]
ENGINEERING CHEMISTRY
accuracy [illegible]
ENGINEERING GRAPHICS
accuracy [illegible]
INTRODUCTION TO ELECTRICAL ENGINEERING
accuracy [illegible]
INTRODUCTION TO SUSTAINABLE ENGINEERING
accuracy [illegible]
BASICS OF ELECTRONICS ENGINEERING
accuracy [illegible]

Fig 5.4 Result of Decision Tree Algorithm

VII. FUTURE WORKS

One limitation of this project is that we took only a small data set. We constrained our project in a way to predict result of students of a particular branch only. So our project can be extended to predict result of students in entire college or university. In this project we have used Decision tree, Naïve Bayes and Support Vector Machine(SVM) algorithms. Other methods like clustering, regression and Artificial Neural Networks(ANN) can be used to have a better understanding of importance of method selection and also to improve efficiency of the model. Also to improve the accuracy of prediction we can combine 2 or more ML algorithms that yields a better result.

REFERENCES

[1] Using Machine Learning to Predict Student Performance, University of Tampere Faculty of Natural Sciences Software Development M. Sc. Thesis Supervisor: Jorma Laurikkala June 2017

[2] Student Performance Prediction via Online Learning Behavior Analytics Wei Zhang, Xujun Huang*, Shengming Wang, Jiangbo Shu, Hai Liu, Hao Chen National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China,2017

[3] Analysis of student performance based on classification and mapreduce approach in bigdata, Dr.R Senthil Kumar1, Jithin Kumar.K.P2 1,2Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, MysuruCampus Mysuru, India sen07mca@gmail.com1, jithinkumarkp18
[4]http://www.academia.edu/24311300/Predicting_Students_Academic_Perfomace_using_Naive_Bayes_Algorithm

[5] Student's Performance Prediction in Education using Data Mining Techniques Mukesh Kumar, Prof. A.J. Singh, Dr. Disha Handa,Himachal Pradesh University, Summer-Hill, Shimla (H.P) Pin Code: 171005, India. IT Consultant, DesktekTeam, 2017