# CREATING DATA BACKBONES FOR STUDENT BEHAVIOUR ANALYSIS USING DECISION SUPPORT SYSTEM

**Mr. V. Vivekanandan** **Ms. N. Karpagavalli** **Mr. R. Manoj Kumar** **Ms. A. Devipriya**

*Computer Science and Engineering,*
*SriGuru Institute of Technology,*
*Coimbatore, Tamilnadu, India*

***Abstract—Data mining techniques are applied to predict school failure and idler of the Students. That use real data on school students for prediction of failure and dropout. It implements white-box classification methods, like induction rule and decision tree. DT is a decision support tool that represented as like graph or a model of decision. It consists of nodes, in which the internal nodes are denoted as test on attributes. Attribute is nothing but real data of students that collected from school in middle or secondary education. A path from root to leaf is represents classification rule and it consists of three types of nodes which includes chance node, decision node, and end node. It is mostly used in decision analysis. Using this technique to attempt to improve their correctness for predicting which students might dropout or fail by first, using all the available attributes next, and selecting the best attribute. Attribute selection done by using WEKA tool. WEKA is a Data Mining tool widely used in classification and prediction of data. WEKA tool supports several standard data mining tasks like clustering, classification, data pre-processing and feature selection. Data is rebalanced using cost responsive classification that is Naive Bayes Algorithm. The naive Bayes classifier is works based on Bayes rule of conditional probability and it accepts all attributes are contained in dataset, it takes some sample for making classification. The outcome was compared and the models with the results are exposed.***

***Keywords— KDD, Classification, Prediction, NavieBaye, J48.***

## I. INTRODUCTION

Here DSS proposed that when students were found in danger, and appointed to a coach so as to produce them with each educational support and steerage for motivating and attempting to stop student failure.It have shown that classification algorithms can be used with success so as to predict a student's educational performance and, particularly, to model the distinction between Fail and Pass students [1][2].

Data mining could be a broad method that consists of many stages and includes several techniques, among the data. This knowledge discovery method contains the steps of pre-processing, the appliance of DM techniques and also the analysis and reading of the results.DM is aimed toward operating with terribly massive amounts of information (millions and billions). The statistics doesn't typically work well in massive databases with high spatiality.

## II. EASE OF USE

### *2.1 Classification*

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. The individual observations are analyzed into a set of quantifiable properties, known as various explanatory variables, features, etc.

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category[3][4].

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering (or cluster analysis), and involves grouping data into categories based on some measure of inherent similarity (e.g. the distance between instances, considered as vectors in a multi-dimensional vector space).

### *2.2 Naive Bayes Classifier*

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". An overview of statistical classifiers is given in the article on Pattern recognition.

In simple terms, a naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features[5].

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of Naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix [6][7].

## III. PROBABILISTIC MODEL

Probability model for a classifier is a conditional model.

$$p(C|F_1,\dots,F_n)$$

Over a dependent class variable $C$ with a small number of outcomes or *classes*, conditional on several feature variables $F_1$ through$F_n$. The problem is that if the number of features $n$ is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. Therefore reformulate the model to make it more tractable.

In plain English the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on $C$ and the values of the features $F_i$ are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

$$p(C, F_1, \dots, F_n)$$

Which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

### *3.1 Constructing a classifier from the probability model*

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier, a Bayes classifier, is the function $\mathrm{classify}$ defined.

### *3.2 Predictive analytics*

Predictive analytics encompasses a variety of techniques from statistics, modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events.In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.Predictive analytics is used in actuarial science, marketing, financial services, insurance, telecommunications, retail, travel and other fields.

One of the most well-known applications is credit scoring, which is used throughout financial services. Scoring models process a customer's credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time. A well-known example is the FICO score.

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future.

For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

## IV. PREDICTION

A prediction or forecast is a statement about the way things will happen in the future, often but not always based on experience or knowledge. While there is much overlap between prediction and forecast, a prediction may be a statement that some outcome is expected, while a forecast is more specific, and may cover a range of possible outcomes. Although guaranteed information about the future is in many cases impossible, prediction is necessary to allow plans to be made about possible developments; Howard H. Stevenson writes that prediction in business "... is at least two things: Important and hard.

## V. PROPOSED SYSTEM

DSSs include knowledge-based systems. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from a combination of raw data, documents, and personal knowledge, or business models to identify and solve problems and make decisions. Here data mining techniques are applied to predict school failure and idler of the Student. That use real data on school students for prediction of failure and dropped out.

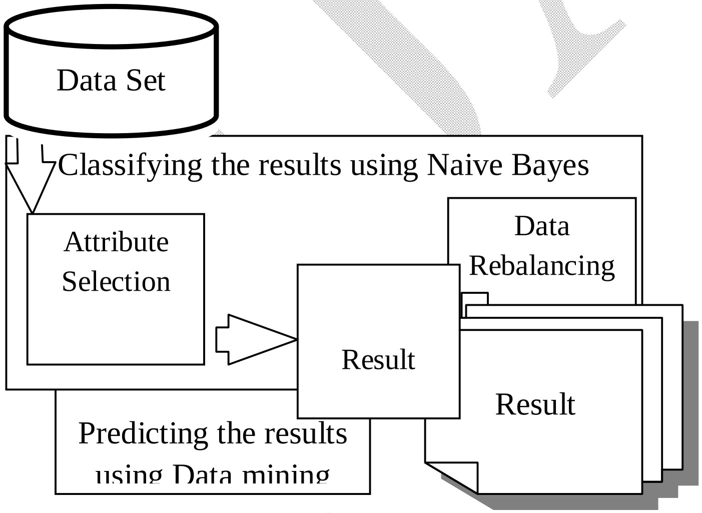
Data Set
Classifying the results using Naive Bayes
Attribute Selection
Data Rebalancing
Result
Result
Predicting the results using Data mining

*Fig 1: Architecture Diagram*

### *5.1 Creating Dataset*

The process of data gathering that involves in collecting all information about students. The set of attributes should be identified that can affect student's performance and collected from available data source. The collected characteristics that can influence to students failure or dropout. Characteristics that contain the information about students educational, cultural, social, background, economic status, academic progress and psychological profile. In which most of the students are aged between 15 and 16 and this is the years with the highest rate of failure. Finally the survey is to obtain family and personal information to identify important factors of all students and school services provides the score obtained by the students in all subjects. All those information's are integrated into a single dataset.

### *5.2 Data Partitioning - Pre Processing*

Dataset is prepared for applying the data mining techniques. Before applying data mining technique, pre-processing methods like data partitioning and variable transformation and other technique for attribute selection must be applied. For example New attribute of age is created using date of birth of each student. The continues variables are transformed into discreet variable that is scores obtained by each student is changed into categorical values that is Excellent score between 9.5 and 10,Very good the score between 8.5 and 9.4.all information's are integrated in single dataset that is stored in attribute relation file format of Weka tool. Finally entire dataset is divided randomly into 10 pairs of test and training data files. After pre-processing it will have variables or attributes for each student. Each training and test file will contain best attributes and rebalanced.

### *5.3 Data Mining*

Data mining technique is going to be applied. Here the data mining technique is mainly used for classifications. The classification is based on best attribute selection from data set. In which the naive bays algorithm is implemented for classification of data. Traditionally the Weka Software tool is used for data mining. It contains verity of data mining algorithms. Weka implements decision tree, it is a set of condition organized in hierarchical structure. Decision tree algorithms are like J48, AD Tree, C4.5, Random Tree etc. Here the classification algorithms were executed using cross- validation and all available information. Finally the result with the test file of classification is shown.

## VI. ANALYZING DATA

The obtained results are analyzed to predict student failure or dropout. To achieve this previous results are taken for the comparison. In this stage rules for classification techniques are applied to predict relevant factors and relationships that lead to student fail or pass. There are variables that indicate that student who failed are older than 15 year and some of the attribute are not presented, shows marks of poor and regular students. Finally the student's characteristics are analyzed with previous results of classification algorithms.

## VII. CONCLUSION

As seen, predicting student failure at school can be a difficult task not only because it is a multifactor problem (in which there are a lot of personal, family, social, and economic factors that can be influential) but also because the available data are normally imbalanced. To resolve these problems, it has shown the use of different DM algorithms and approaches for predicting student failure. It has carried out several experiments using real data from high school students in Mexico. Here applied different classification approaches for predicting the academic status or final student performance at the end of the course. Furthermore here shown that some approaches such as selecting the best attributes, cost-sensitive classification, and data balancing can also be very useful for improving accuracy.

It is important to notice that gathering information and pre-processing data were two very important tasks in this work. In fact, the quality and the reliability of the used information directly affect the results obtained. However, this is an arduous task that involves a lot of time to do. Specifically, had to do the pick out of data from a paper and pencil survey and had to integrate data from three different sources to form the final dataset.

Starting from the previous models (rules and decision trees) generated by the DM algorithms, a system to alert the teacher and their parents about students who are potentially at risk of failing or drop out can be implemented. As an example of possible action, propose that once students were found at risk, they would be assigned to a tutor in order to provide them with both academic support and guidance for motivating and trying to prevent student failure.

## References

[1] Rahul SharanRenu, Gregory MockoandAbhiramKoneru"Use of Big Data and Knowledge Discovery to Create Data Backbones for Decision support Systems" Complex Adaptive Systems, Publication 3,Elsevier 2013.

[2] L. A. Alvares Aldaco, "Comportamiento de la deserción y reprobación en el colegio de bachilleres del estado de baja california: Caso plantel ensenada," in *Proc. 10th Congr. Nat. Invest. Educ.*, 2009, pp. 1–12.

[3] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university dropout rates," *Comput. Educ.*, vol. 53, no. 3, pp. 563–574, 2009.

[4] M. N. Quadril and N. V. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global J.Comput. Sci. Technol.*, vol. 10, pp. 2–5, Feb. 2010.

[5] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.

[6] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[7] S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education," *Based Syst.*, vol. 23, no. 6, pp. 529–535, Aug. 2010.