**DATA MINING PROPOSAL**

**Unish Shah (**[ushah2@student.gsu.edu](mailto:ushah2@student.gsu.edu)**)**

**Ashish Tiwari (**atiwari5[@student.gsu.edu](mailto:ushah2@student.gsu.edu))

**Abstract**: **The aim of this project is to improve the current trends in the higher education systems and to find out which factors might help in creating successful students. It is really necessary to find successful students as it motivates higher education systems to know them well and one way to know this is by using valid management and processing of the student’s database. Data mining methods represent a valid approach for the extraction of precious information from existing students to manage operations with future students. It will help to indicate which type of students will potentially be successful and what areas to concentrate upon in higher education systems for creating more successful students. For this purpose, the data mining framework is used for mining related to academic data from enrolled students**.

1. Introduction

The ability to predict a student’s performance is very important in educational environments. Students’ academic performance is based upon diverse factors like personal, social, psychological and other environmental variables. A very promising tool to attain this objective is the use of Data Mining. Data mining techniques are used to operate on large amount of data to discover hidden patterns and relationships helpful in decision making. Modern educational organizations start developing and enhancing the educational system increasing their capability to help the decision makers obtain the right knowledge, and to make the best decisions by using the new techniques such as data mining methods [1].

**2. Problem Statement**

Student success rate in any higher education system is one of the greatest concerns to higher education managements. Several factors may affect the student success rate in Institute. One of the biggest challenges that higher education faces today is predicting which major factors play a role in generating successful students.

**3. Data Description and Pre-processing**

Data source link: <http://archive.ics.uci.edu/ml/datasets/student+performance>[2]

Data format: Integer

Size: 396 rows X 33 columns

Number of Instances: 396

Number of Attributes: 33

This data is of student’s achievement in secondary education of Portuguese school. The data attributes include student grades, demographic, social and school related features) and it was collected by using questionnaires and school reports. Dataset are provided regarding the performance in subject: Mathematics. The target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade, while G1 and G2 correspond to the 1st and 2nd period grades.

During the data pre-processing set we found out that data present in our dataset was clean, as a result we did not had to perform the data cleaning methods.

In our dataset we had 33 attributes and as result we had to reduce some of the attributes which were not so important, to get better accuracy and low-cost tree. In organizations these kind of strategies is performed to reduce the data, so we also decided to do the same.

4. **Description of Algorithms**

**We have used four algorithms in our project. The algorithms we have used and implemented are classification algorithms.**

They are listed and explained below:

1. Decision tree: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences
2. Naive Bayesian: Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
3. SVM: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane.

4 K nearest neighbour: In pattern recognition, the k-nearest neighbour’s algorithm (k-NN) is a non-parametric method used for classification and regression.

We are going to compare the Accuracy, Classification error, Sensitivity, Specificity using example models of various classification algorithms stated above as our evaluation metrics.

5. Experiments

We have implemented our algorithms with the help of Python. We have made use of in-built python libraries and packages to implement our classification algorithms. We have made use of the following libraries and packages:

1) Numpy

2) Pandas

3) Scikit-learn

4) Matplotlib

The First algorithm which we have implemented is Decision Tree.

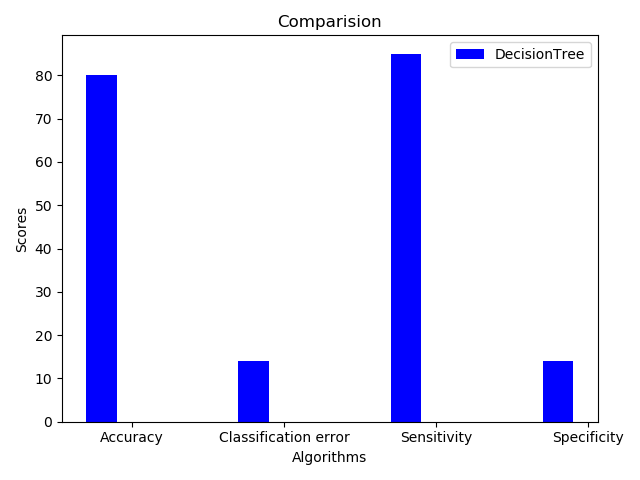
The results of implementing are as below:

Accuracy is 0.80

Error rate is 0.141414141414

Sensitivity is 0.858585858585

Specificity is 0.141414141414



The Second algorithm which we have implemented is Naïve Bayesian.

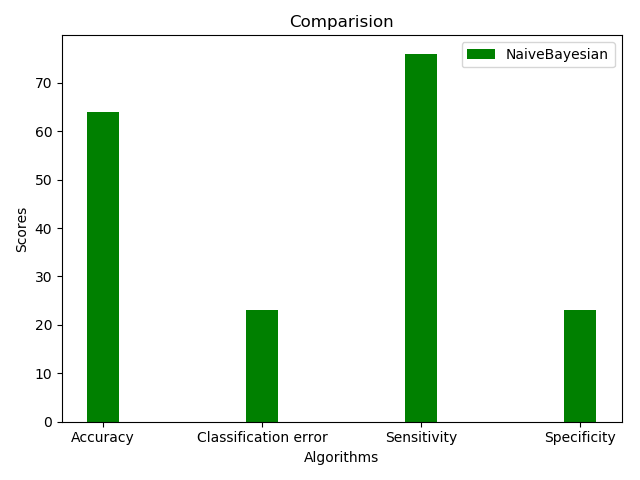
The results of implementing are as below:

Accuracy is 0.64135021097

Classification error is 0.35864978903

Sensitivity is 0.64135021097

Specificity is 0.35864978903



The Third algorithm which we have implemented is K-Nearest Neighbor.

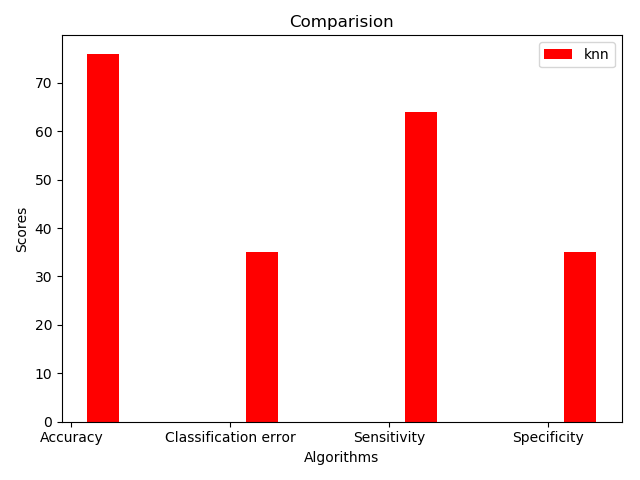
The results of implementing are as below:

Accuracy is 0.737373737374

Classification error is 0.262626262626

Sensitivity is 0.737373737374

Specificity is 0.262626262626



The fourth algorithm which we have implemented is SVM.

The results of implementing are as below:

Accuracy is 0.670498084291

Classification error is 0.329501915709

sensitivity is 0.670498084291

specificity is 0.329501915709

A screenshot of a cell phone

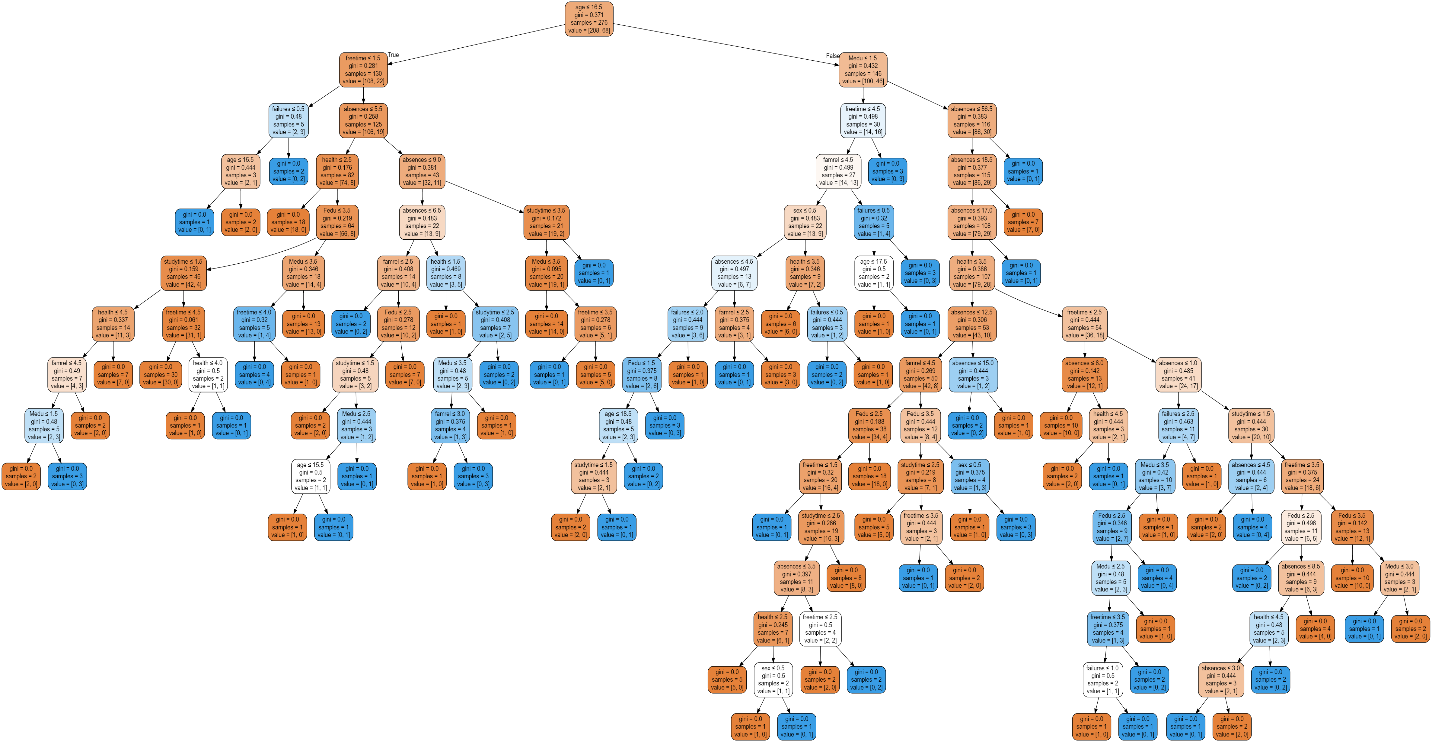
Description generated with very high confidence

Comparisons of various algorithms are displayed below:

A screenshot of a cell phone

Description generated with high confidence

Here is the result of our decision tree:



Conclusions

With the help of classification methods performed above we can now extract the precious information from existing students to manage operations with the future students. The classification algorithms can help us identify and indicate that which type of students will potentially be successful and it will also help the schools to get more information regarding which areas to concentrate upon in higher education system so that create more successful students.

**References**

1. Higher Education Enhancement Project (HEEP), 2007, <http://www.heep.edu.eg>.
2. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Division of Work and Contributions

|  |  |
| --- | --- |
| **Task** | **People** |
| 1. Collecting and preprocessing data | Unish Shah, Ashish Tiwari |
| 2. Implementing Decision Tree | Unish Shah |
| 3. Implementing Naïve Bayesian | Unish Shah |
| 4. Implementing K- Nearest means | Unish Shah |
| 5. Implementing SVM | Ashish Tiwari |
| 6. Evaluating and comparing algorithms | Unish Shah, Ashish Tiwari |
| 7. Slides | Ashish Tiwari |
| 8. Demo and Presentation | Unish Shah, Ashish Tiwari |
| 9. Writing report | Unish Shah, Ashish Tiwari |