3 Research Methodology

3.1 Research Question

Can machine learning algorithms improve the accuracy in predicting the application status of students aspiring to enrol for Masters in Computer Science course at universities in the USA?

3.2 Methodology

CRoss-Industry Standard Process (CRISP) methodology (Azevedo, 2008) was followed in this research.

Business Understanding: Initially good amount of time was spent on understanding the problem statement by understanding the concerns of students regarding the current application process, the objectives of the research were defined in this process.

Data Understanding: Data required for the research was collected from multiple data sources. Different features of the data were analyzed based on their importance and rel- evance. Data-set would be explained in more detail further.

Data Preparation: In this phase, the data from multiple data sources were integ- rated into a final data-set. Further the data was cleaned by removing unwanted columns, performing transformation and cleaning activities on the data.

Modelling: Multiple machine learning models were developed to predict the likelihood of success of the student's application in a particular university. The user interface was developed to allow the users to access these models.

Evaluation: Models developed were evaluated based on their performance and accuracy. More information will be presented in the evaluation section of the paper.

Deployment: Once the models were evaluated they were integrated with code developed for user interface using the Shiny package in R.

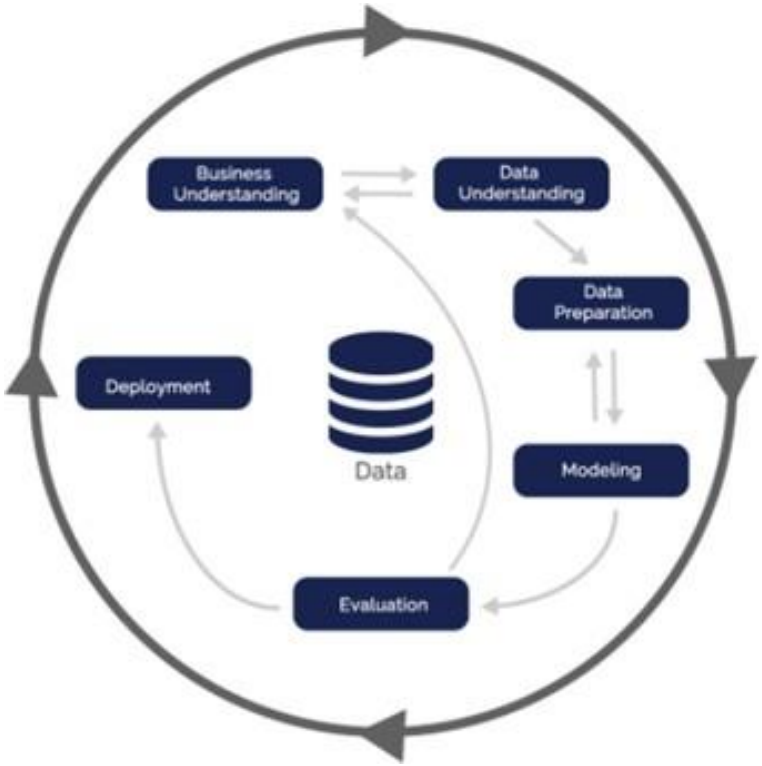

Figure 1: CRISPDM

3.3 Research Significance

The principal objective of the research is to help the students who are aspiring to pursue their education in the USA. The SAP system will help them to evaluate the chances of success in a particular university without being dependent on any education consultancy firm. It will help them in saving a huge amount of time and money spent in the applica- tion process.
Also, it will help them to limit the number of applications made by the students by sug- gesting them the best universities where they have high chances of securing admission thereby by saving the amount of money spent by the students by applying in universities where they have less chance to secure admit based on their profile

3.4 Research Limitations

Student Admission Predictor system will only take into consideration the data related to the Indian students pursuing Masters in Computer Science from universities in the USA.

4 Implementation

4.1 Data-set

This section describes, in brief, the data that has been used for the research. Data from multiple sources was used in this project, the major amount of data was extracted from public website Yocket(Yocket), data regarding the rankings, fees and enrolment in col- leges was obtained from a leading educational consultancy firm The Mentors Circle in India. Data from both the sources was integrated together to form a staging data-set. For predicting the chance of a student getting shortlisted in universities the final data-set was divided into multiple data-sets each representing a particular university. For predicting the list of universities suitable for students based on their profile data of all the students the staging data-set was updated only to have records of students who had successfully secured admission in the universities. Below table shows the different features of the data-sets.

Field	Description
GRE	Marks scored by the student in GRE
Language	Marks scored by the student in TOELF/IELFTS exam
UGPA	Result of the student in their undergraduate course
SOP	Quality of students Statement Of Purpose document
LOR	Quality of students Letter Of Recommendation document
Work_Exp	Students work experience in months
Intake	Type of intake Fall/Spring
Status	Status of application Accept/Reject (To be predicted using KNN)
Rank	Rank of Universities 1/2/3(To be predicted using Decision Tree)

Figure 2: Data-set

4.2 Data-set Extraction and Transformation

Data related to the college ranking was collected in .csv format, the data related to students profile was extracted from (Yocket (2017)) using data extraction tool provided by (Mozenda (n.d.))in .csv files. Data being from public portal had multiple records with missing and irrelevant values; data cleaning was performed in Microsoft Excel by deleting the records having

unwanted and missing values. Unwanted columns were removed from the data-set. Once the data-set was cleaned data was transformed to be suitable for the model. The original data-set had TOEFL or IELTS score as a representation of language, to have a consistent metrics for the language score of all the records were converted to IELTS scale using the conversion table.(Prescholar (2017)). Similarly, the UGPA score of the students was represented in terms of percentage and CGPA; all the records of percentage were converted to CGPA by multiplying percentage score by 9.5. The values of Intake, Status and Rank fields were changed as shown below to have numerical values for the data to achieve better results for KNN.

Rank	Value
Top 10-30	1
Top 40-60	2
Top 100	3

Status	Value
Accept	1
Reject	0

Intake	Value
Fall	0
Spring	1

Figure 3: Data

4.3 Algorithms

Multiple machine learning algorithms were used for this research, K- Nearest Neighbour and Multivariate Logistic Regression algorithms were used to predict the likelihood of the students getting admission into university based on their profile. Decision Tree algorithm was used to predict the rank of the college that would be suitable for the students based on their profile and suggest the list of universities accordingly.

K-Nearest Neighbours: It is an algorithm which is used widely for classification and regression problems. Due to its simplicity and effectiveness, it is easy to implement and understand. It is a supervised machine learning algorithm that uses available data to create the model and further that model can be applied to classify the new data. The class of new data is determined by the class of its neighbours. Distance is calculated between the unseen data sample and the all other data samples already present in the data-set. Depending on the value of K, that many nearest neighbours are selected and their class is identified. The class of neighbours which has majority is assigned to the class of the new data sample. Generally, Euclidean distance is used to calculate the distance between the records. Multiple values of K should be tried and tested, and the value of K at which best performance is observed must be selected for the model.

Logistic Regression: Logistic regression algorithm is used to identify the probab- ility of occurrence of an event based on single predictor variable. Multivariate Logistic regression can be used to determine the probability of the occurrence of an event based on multiple predictor variables. The class variable that has to be predicted has to be bin- ary or dichotomous. Logistic Regression is also a supervised machine learning algorithm which used data with

predetermined classes to create a model and perform predictive analysis on unseen data.

Decision Tree: It is a supervised machine learning algorithm. Due to its simple logic, effectiveness and interpretability it the most widely used classification algorithm. The model works by creating a tree-like structure by dividing the data-set into several smaller subsets based on different conditional logic. The main components of the decision tree are the decision nodes, leaf nodes and the branches. Nodes with multiple branches are the decision nodes, nodes with no branches are called the leaf nodes, and the top node is called the root node of the decision tree. The nodes are connected to each other via branches based which are different conditions. The root and decision nodes are created by computing the entropy and information gain for the data-set.

Shiny Library: The Shiny package in R is used to create interactive standalone

and web-based applications. It allows creating a user interface for the R programs by providing a platform to integrate the presentation code and program code in a single.

4.4 Architecture

In this section, we will describe the architecture of the Student Admission Predictor system. The figure below explains the flow of the system:

- The student will enter his/her profile details using the user interface developed in shiny.
- The user interface code will interact with the KNN and Decision Tree models to provide the users with the required result.
- The KNN algorithm will be used to determine the chance of the student of securing admission in a particular university based on his/her profile.
- The Decision Tree algorithm will be used to determine the rank of college to which is most suitable for the student based on his/her profile and provide the student with the list of universities which fall in that rank.
- Once the models have been executed the result will be provided to the student as the output on the user interface.

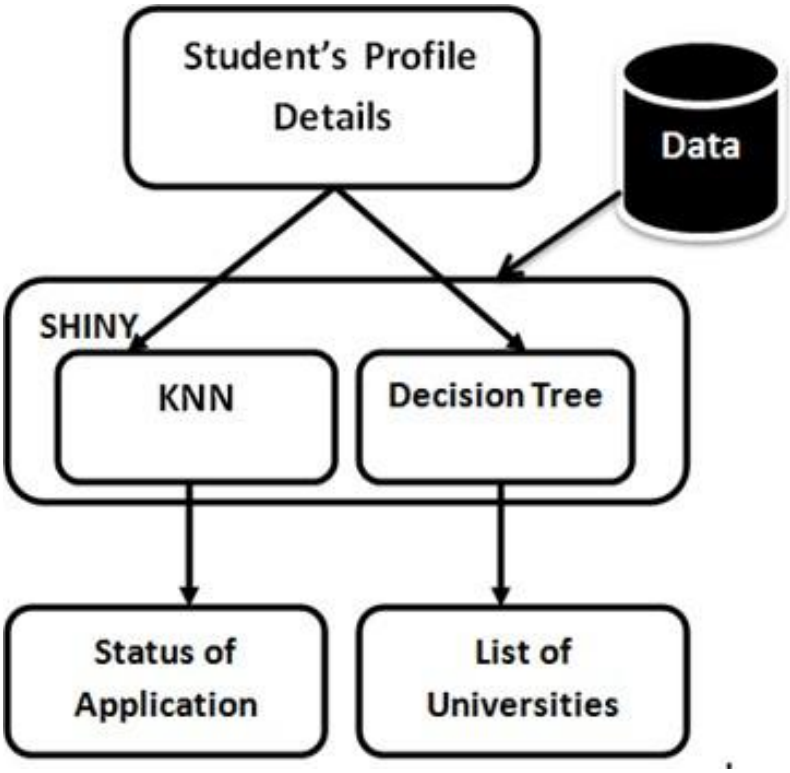

Figure 4: SAP Architecture

5 Evaluation

K-nearest neighbour and Multivariate logistic regression algorithms were used to create a model that can be used to predict the likelihood of success of a students application to the university based on his/her profile. Both algorithms were tested and their performance was evaluated based on different factors like Accuracy, Sensitivity, Specificity and Kappa value. As can be seen the figure given below model created using K-Nearest Neighbour outperformed the model created using Logistic Regression on all the performance meas- ures. Also by looking at the variance in the values of the data KNN seemed to be the best-fit algorithm to create the Student Admission Predictor System.

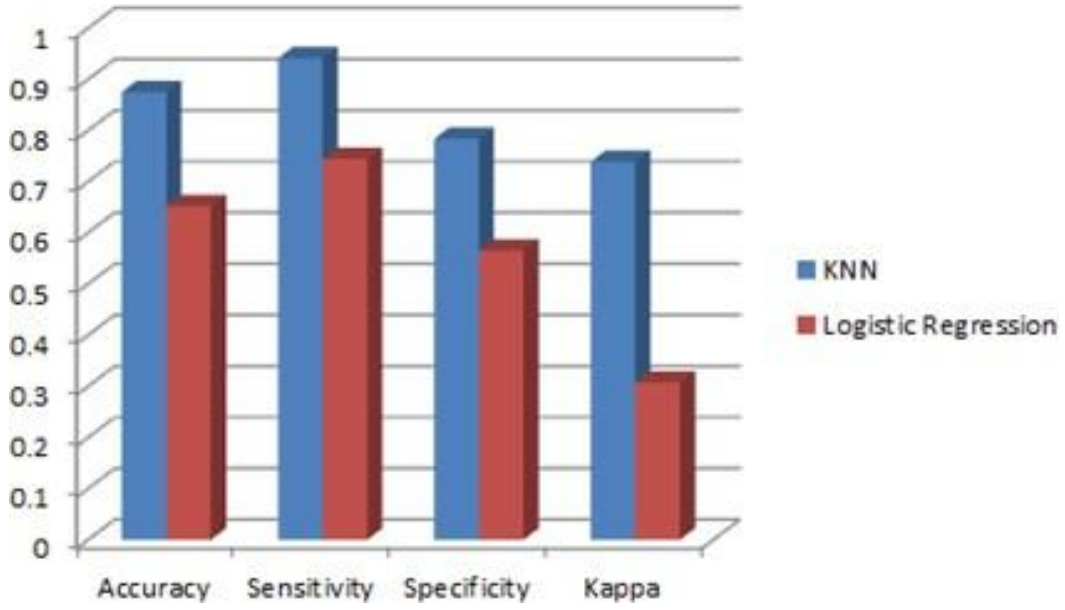

Figure 5: KNN vs Logistic Regression

Accuracy was considered to be main metric to evaluate the performance of the models, as the data used for creating the models was balanced. Also, prediction of the true positive and true negative scenarios was equally equivalent. The KNN model performed well with an overall average accuracy of 76%. The decision tree model which was created to predict the rank of the universities suitable for the student provided the result with an accuracy of 80%.

```
Confusion Matrix and Statistics

pred   1   2   3
   1 173  37   2
   2  35  97  23
   3   0  18 175

Overall Statistics

               Accuracy : 0.7946
                 95% CI : (0.7588, 0.8274)
    No Information Rate : 0.3714
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6894
 Mcnemar's Test P-Value : 0.4462

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.8317   0.6382   0.8750
Specificity            0.8892   0.8578   0.9500
Pos Pred Value         0.8160   0.6258   0.9067
Neg Pred Value         0.8994   0.8642   0.9319
Prevalence             0.3714   0.2714   0.3571
Detection Rate         0.3089   0.1732   0.3125
Detection Prevalence   0.3786   0.2768   0.3446
Balanced Accuracy      0.8605   0.7480   0.9125
```

Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 48  8
         1  3 29

               Accuracy : 0.875
                 95% CI : (0.7873, 0.9359)
    No Information Rate : 0.5795
    P-Value [Acc > NIR] : 1.642e-09

                  Kappa : 0.7387
 Mcnemar's Test P-Value : 0.2278

            Sensitivity : 0.9412
            Specificity : 0.7838
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.9062
             Prevalence : 0.5795
         Detection Rate : 0.5455
   Detection Prevalence : 0.6364
      Balanced Accuracy : 0.8625
```

KNN

Figure 6: Confusion Matrix and Statistics

The above figure provides the details of the confusion metrics and overall statistics of the KNN and Decision Tree algorithms.

Below is the plot of the decision tree developed to predict the rank of the university best suitable for a student based on his/her profile.