Abstract

In todays era we see a lot of students pursuing their education away from their home countries. The main country targeted by these international students is The United States of America. Majority of the international students in the United States of America are from India and China. In the past decade the number of Indian students pursuing post graduate education from the USA has rapidly increased. With the increase in the number of international students studying in the USA, each applicant has to face a tough competition to get admission in their dream university. Generally as the students dont have much idea about the procedures, requirements and details of the universities in the USA they seek help from the education consultancy firms to help them successfully secure admission in the universities which are best suitable for their profile, for this they have to invest huge amount of money as consultancy fees. Apart from these the education consultancy firms there are few websites and blogs that guide the students on the admission procedures. The drawback of the currently available resources is that they are very limited and also they are not truly dependable taking into consideration of their accuracy and reliability. The aim of this research is to develop a system using machine learning algorithms, we will name it as Student Admission Predictor (SAP). It will help the students to identify the chances of their application to an university being accepted. Also it will help them in identifying the universities which are best suitable for their profile and also provide them with the details of those universities. A simple user interface for the users to access the SAP system.

Keywords: Student Admission Predictor; Machine Learning; Data Mining

1   Introduction

A persons education plays a vital role in their life. While planning for education students often have several questions regarding the courses, universities, job opportunities, expenses involved, etc. Securing admission in their dream university is one of their main concerns. It is seen that often students prefer to pursue their education from universities which have global recognition. And when it comes to international students the first choice of the majority of them is the United States of America. With the majority of worlds highly reputed universities, wide range of courses offered in every sector, highly accredited education system and teaching, scholarships provided to students, best job market and many more advantages make it the dream destination for the international

students. According to research, there are above 8 Million international students studying in more than 1700 public and 2500 private universities and colleges across the USA. (MasterPortal (2017))

The majority of international students studying in the USA are from India and China. In the past decade, India has seen a huge increase in the number of students opting to pursue their education from foreign universities in countries like The USA, Ireland, Australia, Germany, etc. Although there are significant universities and colleges in India, students are finding it difficult to get admission in the highly ranked colleges and also getting a job is a challenge as the ratio of number students to the number work

opportunities available is quite high. India is one of the leading counties in the number of software engineers produced each year; it becomes tough for the students to find jobs in elite companies due to high competition. This motivates a good number of students to pursue post-graduation in their field. It is seen that the number of students pursuing Masters in Computer Science field from universities in the USA is quite high; the focus of this research will be on these students.
(Times (2015))

Majority of universities in the USA follow similar guidelines for providing admission to students. Universities take into consideration different factors like score on aptitude based examination like the General Record Examination (GRE), command over the English language is judged based on their score in English competency test like Test Of English as a Foreign Language (TOEFL) OR International English Language Testing System (IELTS), their work experience in same or other fields, the quality of the Letters Of Recommendation (LOR) and the Statement Of Purpose documents provided by the student etc. Based on the overall profile of the student decision is taken by the universities admission team to admit or reject a particular candidate.

Every candidate has to take all the required examination and build a strong profile to secure admission in their dream universities in the USA. Once the candidates have made their profile ready, they apply to the universities where they aim to secure admission. The students have to shortlist the universities which are best known for the courses they are looking for and also they should have an idea about their chances of securing admission in those universities based on their profile. This task of shortlisting the universities where the student has high chances of admission is difficult for mainly for the international students, so they end up with applying to many universities in hopes of getting admission in few of them thus investing an extra amount of money in the applications. There are several portals and websites which provide information and help to students in shortlisting the universities, but they are not reliable. Most of the students dont take the risk of evaluating the colleges by themselves, and they seek the help of the education consultancy firms to do it for them. Again for this students have to pay a huge amount of fee to the education consultant.

The primary objective of this research is to develop a system to solve the problems the international students are facing while applying for universities in the USA. We will be developing a Student Admission Predictor (SAP) system which will help the students to predict the chances of their application being selected for a particular university for which they wish to apply based on their profile. Also, the system will provide a recommendation of universities to the student to which the student has a high possibility of getting admission. Multiple machine learning classification algorithms were evaluated to develop the system. Finally, K Nearest Neighbours and Decision Tree algorithms were used as they were found to be the best fit for the system developed. Also, we will be creating a simple user interface which will help the users to input the data related to student profile and get the predicted result for the application based on the profile as output.

This research will thus eventually help students saving the extra amount of time and money they have to spend at the education consultancy firms. And also it will help them to limit their number of application to a small number by proving them the suggestion of the universities where they have the best chance of securing admission thus saving more money on the application fees.

2   Related Work

This section provides the literature review of the work that has previously done on predicting the chances of students enrolment in universities. There have been several project and studies performed on topics related to students admission into universities. (Bibodi et al. (n.d.)) used multiple machine learning models to create a system that would help the students to shortlist the universities suitable for them also a second model was created to help the colleges to decide on enrolment of the student. Nave Bayes algorithm was used to predict the likelihood of success of an application, and multiple classification algorithms like Decision Tree, Random Forest, Nave Bayes and SVM were compared and evaluated based on their accuracy to select the best candidates for the college. Limitation of this research as that it did only relied on the GRE, TOEFL and Undergraduate Score of the student and missed on taking into consideration other important factors like SOP and LOR documents quality, past work experience, technical papers of the students etc.

Bayesian Networks were used by (Thi et al. (2007)) to create a decision support system for evaluating the application submitted by international students in the university. This model was designed to predict the performance of the aspiring students by comparing them with the performance of students currently studying in the university and had similar profile during their application. In this way based on the current students profile the model predicted whether the aspiring student should be granted admission to the university. Since the comparisons were made only with the students who were already admitted in the university and the data of the students who were denied admission were not included in the research this model proved to be less efficient due to the problem of class imbalance.

(Abdul Fatah S; M (2012)) developed a model that can provide the list of universities/colleges where the which best suitable for a student based on their academic records and college admission criteria. The model was developed by applying data mining techniques and knowledge discovery rules to the already existing in-house admission prediction system of the university. (Mane (2016)) conducted a similar research that predicted the chance of a student getting admission in college based on their Senior Secondary School, Higher Secondary School and Common Entrance Examination scores using the pattern growth approach to association rule mining. The performance of both the models was good the only drawback was the problem statement was single university-centric. (Mishra and Sahoo (2016))conducted a research from a university point of view to predict the likelihood of a student enrolling in the university after the have enquired about of courses in the university. They used K-Means algorithm for clustering the students based on different factors like feedback, family income, family occupation, parents qualification, motivation etc. to predict if the student will enroll at the university or not. Depending upon the similarity of the attributes among the students they were grouped into clusters and decisions were made. The objective of the model was to increase the enrolment of the students in the university.

(Eberle et al. (n.d.)) used machine learning and predictive modelling to develop a model that to evaluate the admission policies and standards in the Tennessee Tech University. A well know version of the C4.5 algorithm, J48 was used to create the model. Like the models mentioned above they used the different factors of the student profile to evaluate the chances of their admission in the university. The model worked well in predicting the true positive scenarios where the student was had good profile to secure

the admission, but it failed in efficiently identifying the true negatives because of which student who does not satisfy the defined criteria.

In research conducted by (Jamison (2017)) the yield of college admission was predicted using machine learning techniques. Yield rate can be defined as the rate at which the students who have been granted admission by the university actually enrol for the course. Multiple machine learning algorithms like Random Forest, Logistic Regression and SVM were used to create the model; the models were compared based on their performance and accuracy, Random Forest outperformed the other models with 86% accuracy and was thus used to create the system. The factors that proved to be significant in predicting successful application were also highlighted.

GRADE system was developed by (Waters and Miikkulainen (2013)) to support the admission process for the graduate students in the University of Texas Austin Department of Computer Science. The main objective of the project was to develop a system that can help the admission committee of the university to take better and faster decisions. Logistic regression and SVM were used to create the model, both models performed equally well and the final system was developed using Logistic regression due to its simplicity. The time required by the admission committee to review the applications was reduced by 74% but human intervention was required to make the final decision on status if the application.(Nandeshwar et al. (2014)) created a similar model to predict the enrolment of the student in the university based on the factors like SAT score, GPA score, residency race etc. The Model was created using the Multiple Logistic regression algorithm, it was able to achieve accuracy rate of 67% only.

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 relevance. Data-set would be explained in more detail further.

Data Preparation: In this phase, the data from multiple data sources were integrated 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.
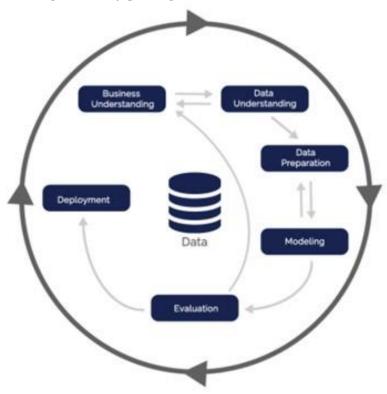
Figure 1: CRISPDM

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 application process.

Also, it will help them to limit the number of applications made by the students by suggesting 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 colleges 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 probability 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 binary 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.
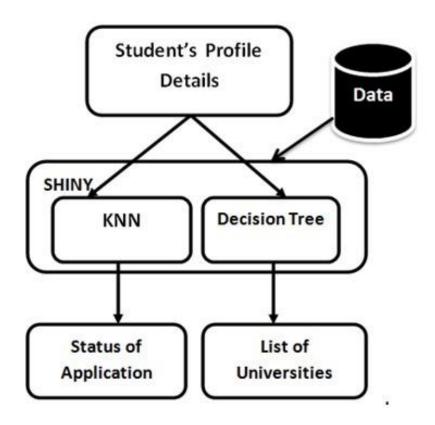

Figure 4: SAP Architecture

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 measures. 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.
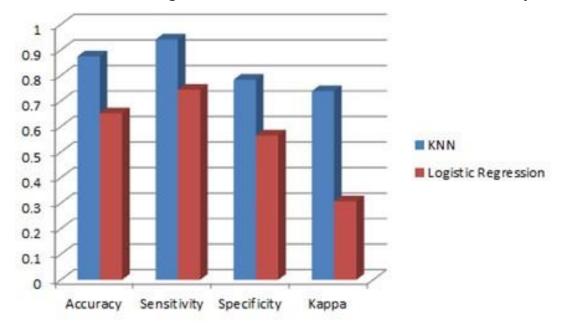

Figure 5: KNN vs Logistic Regression

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%.
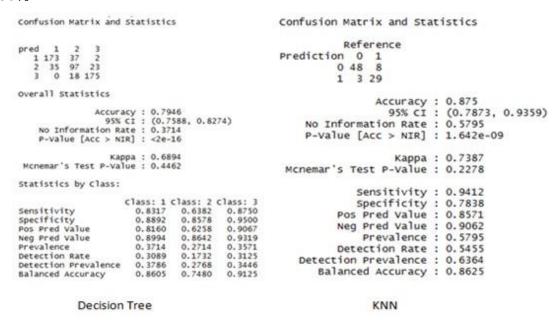

Figure 6: Confusion Matrix and Statistics

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.
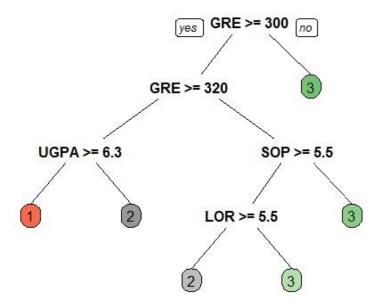

Figure 7: Decision Tree

5.1 Case Study 1

When a student with top grades evaluates his/her profile for admission in colleges Ranked in Top 10 to Top 40 in the USA.

Below is the screen-shot of the user interface before submitting the student profile.
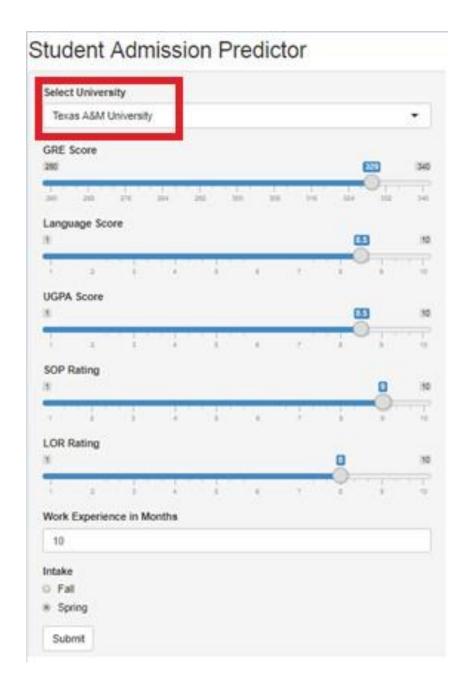
Figure 8: Case Study 1 Input

Below is the output, showing the student has 75% chance of getting admission in the Texas A&M University. It also suggests the student to try applying for the other universities listed. As the grades of the students are very good, universities ranked top 10 to top 40 were recommended. Also other important details regarding the recommended universities like State, Type, Rank, Acceptance Rate, Tuition Fees, Living Expense and Yearly Enrolment are provided to the student.

You can successfully secure admission. Your chances of getting admission is 75%

Below is the list of universities suitable for your profile:

University_Name	State	Type	Rank	Accpetance_Rate	Tution_Fees	Living_expense	Enrollment
University of Southern California	California	Private	Top 10	23	51500	8500	5142
University of California - Los Angeles	California	Public	Top 20	29	32000	8500	1928
Texas A&M University	Texas	Public	Top 20	24	12000	6700	3379
University of Illinois at Chicago	Illinois	Public	Top 30	20	56000	7200	1164

Figure 9: Case Study 1 Output

5.2   Case Study 2

When a student with bad grades evaluates his/her profile for admission in universities ranked Top 10 to Top 30 in the USA.

Below is the screen-shot of the user interface before submitting the student profile details.

Figure 10: Case Study 2 Input

Below is the output, showing the student has no chance of getting admission in the Texas A & M University and it advises the student to try in the list of universities which are suitable for their profile. Also as the student profile is below average it is recommended that the student should try applying in the universities which are ranked in Top 100 instead of applying in universities ranked Top 10 to Top 30.

Sorry you have low chances of securing admission. Please try in below listed universities.

Below is the list of universities suitable for your profile:

University_Name	State	Type	Rank	Accpetance_Rate	Tution_Fees	Living_expense	Enrollment
Syracuse University	New York	Private	Top 100	40	21000	4900	1174
State University of New York at Buffalo	New York	Public	Top 100	24	22300	5000	1923
University of Texas - Arlington	Texas	Public	Top 100	55	15000	5600	2646

Figure 11: Case Study 2 Output

5.3   Case Study 3

The same student who got rejected in Texas A & M University wants to evaluate his chance of getting admission in Syracuse University which was recommended to him by the our system.

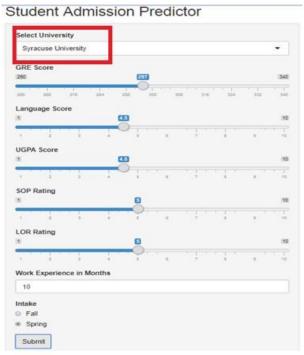Below is the screen-shot of the user interface before submitting the student profile details.

Figure 12: Case Study 3 Input

As expected there is 83% chance that the student will be able to secure admission in the in Syracuse University.

Figure 13: Case Study 3 Output

5.4   Case Study 4

A student with average profile evaluates his/her profile to get admission in universities ranked Top 10 to Top 30.

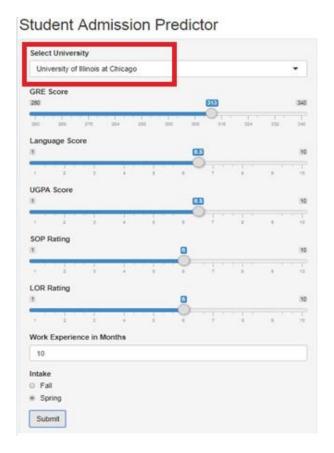Below is the screen-shot of the user interface before submitting the student profile.

Figure 14: Case Study 4 Input

Below is the output, showing the student has no chance of getting admission in the University of Illinois at Chicago and it advises the student to try in the list of universities which are suitable for their profile. Also as the student profile is average so it is recommended that the student should try applying in the universities which are ranked Top 40 to Top 60 colleges instead of applying in universities ranked Top 10 to Top 30.
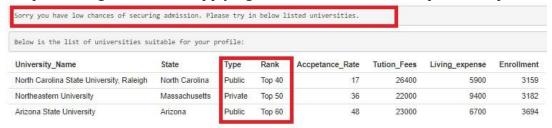

University_Name	State	Type	Rank	Accpetance_Rate	Tution_Fees	Living_expense	Enrollment
North Carolina State University, Raleigh	North Carolina	Public	Top 40	17	26400	5900	3159
Northeastern University	Massachusetts	Private	Top 50	36	22000	9400	3182
Arizona State University	Arizona	Public	Top 60	48	23000	6700	3694

Figure 15: Case Study 4 Output

5.5   Case Study 5

The same student who got rejected in University of Illinois Chicago wants to evaluate his chance of getting admission in Arizona State University which was recommended to him by the system.

Below is the screen-shot of the user interface before submitting the student profile details.
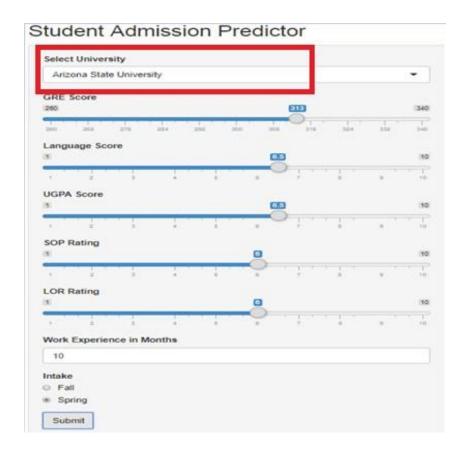
Figure 16: Case Study 5 Input

As expected there is 78% chance that the student will be able to secure admission in the in Arizona State University.

Figure 17: Case Study 5 Output

6   Conclusion and Future Work

The main objective of this research was to develop a prototype of the system that can be used by the students aspiring to pursue their education in the USA. Multiple machine learning algorithms were developed and used for this research. KNN proved to best-fit for development of the system when compared with the Logistic regression model. The model can be used by the students for evaluating their chances of getting shortlisted in a particular university with an average accuracy of 75%. Decision Tree algorithm was used to predict the universities which were best suitable for a student based on their profile. The decision tree algorithm proved to be 80% accurate. A simple user interface was developed to make the application interactive and easy to use for the users from the non-technical background. Shiny library from R was used to create the user interface. The overall objective of the research was achieved successfully as the system allow the

students to save the extra amount of time and money that they would spend on education consultants and application fees for the universities where they have fewer chances of securing admission. Also, it will help the students to make better and faster decision regarding application to the universities.

As discussed earlier in the limitation of the research we have created the models based only on the data of Indian Students studying Masters in Computer Science in the USA, we have considered only ten universities with different rankings. In future, more data related to additional universities and courses can be added to the system. Also, the system can be enhanced to a web-based application by making changes to the Shiny code. Other classification algorithms can be evaluated to resolve the problem if they perform better than the current algorithm the system can be easily updated to support the new algorithm by changing the server code in the Shiny app.

Acknowledgment

References

Abdul Fatah S; M, A. H. (2012). Hybrid Recommender System for Predicting College Admission, pp. 107–113.

Bibodi, J., Vadodaria, A., Rawat, A. and Patel, J. (n.d.). Admission Prediction System Using Machine Learning.

Eberle, W., Simpson, E., Talbert, D., Roberts, L. and Pope, A. (n.d.). Using Machine Learning and Predictive Modeling to Assess Admission Policies and Standards.

Jamison, J. (2017). Applying Machine Learning to Predict Davidson College ' s Admissions Yield, pp. 765–766.

Mane, R. V. (2016). Predicting Student Admission decisions byAssociation Rule Mining with Pattern Growth Approach, pp. 202–207.

MasterPortal (2017). MasterPortal.
   URL: *http://www.mastersportal.eu/countries/82/united-states.html*

Mishra, S. and Sahoo, S. (2016). A Quality Based AutomatedAdmission System for Educational Domain, pp. 221–223.

Mozenda (n.d.). Mozenda.
   URL: *https://www.mozenda.com/*

Nandeshwar, A., Chaudhari, S., Sampath, V., Flagel, A., Figueroa, C., Sugrue, P., Ahlburg, D. and Mcpherson, M. (2014). Predicting Higher Education Enrollment in the United States : An Evaluation of Different Modeling Approaches, *International Journal of Operational Research* **19**(26): 60–67.
   URL: *http://ezproxy.bethel.edu/login?url=http://inderscience.metapress.com/content/fr20314q18626 content/uploads/2008/11/DMWVU Project.pdf*

Prescholar (2017). Yocket.
   URL: *https://www.prepscholar.com/toefl/blog/toefl-to-ielts-conversion*

Thi, N., Hien, N. and Haddawy, P. (2007). A Decision Support System for Evaluating International Student Applications, pp. 1–6.

Times, T. E. (2015). The Economic Times.
   URL: *https://economictimes.indiatimes.com/wealth/personal-financenews/around-3-lakh-indian-students-go-abroad-annually-33-take-property-backedloans/articleshow/48703656.cms?intenttarget=no*

Waters, A. and Miikkulainen, R. (2013). G RADE : Machine Learning Support for Graduate Admissions.

Yocket (2017). Yocket.
   URL: *https://yocket.in/*