UNIVERSITY ADMIT ELIGIBILITY

PREDICTION SYSTEM

Introduction:

The world markets are developing rapidly and continuously looking for the best knowledge and experience among people. Young workers who want to stand out in their jobs are always looking for higher degrees that can help them in improving their skills and knowledge. As a result, the number of students applying for graduate studies has increased in the last decade. This fact has motivated us to study the grades of students and the possibility of admission for master's programs that can help universities in predicting the possibility of accepting master's students submitting each year and provide the needed resources.

Literature Review:

The dataset is related to educational domain. Admission is a dataset with 500 rows that contains 7 different independent variables which are:

• Graduate Record Exam1 (GRE) score. The score will be out of 340 points.

• Test of English as a Foreigner Language2 (TOEFL) score, which will be out of 120 points.

• University Rating (Uni.Rating) that indicates the Bachelor University ranking among the other universities. The score will be out of 5

• Statement of purpose (SOP) which is a document written to show the candidate's life, ambitious and the motivations for the chosen degree/ university. The score will be out of 5 points.

• Letter of Recommendation Strength (LOR) which verifies the candidate professional experience, builds credibility, boosts confidence and ensures your competency. The score is out of 5 points

• Undergraduate GPA (CGPA) out of 10

• Research Experience that can support the application, such as publishing research papers in conferences, working as research assistant with university professor (either 0 or 1). One dependent variable can be predicted which is chance of admission, that is according to the input given will be ranging from 0 to 1. It is worth to mention that all tests will be done using R language. Models will be created using Weka, and statistical test will be performed using PHStat.

Technical background:

A. Shapiro-Wilk Normality Test

The Shapiro-Wilks test is a test performed to detect whether a variable is normally distributed or not depending on the pvalue. In case the p-value was less than or equal 0.05, the test will reject the null hypothesis. Otherwise, the variable is normally distributed. It is good to mention that Shapiro test does has limitations. Moreover, it is biased toward large samples. The larger the sample, the more possibility to get a statistically significant results.

B. **Multiple Linear Regression** Multiple linear regression is a statistical technique used to predict a dependent variable according to two or more independent variables. As well as, present a linear relationship between them and fit them in a linear equation.

The format of the linear equation is as following:

$y_i=\beta_0+\beta_1 x_{i1}+...+\beta_n x_n+\epsilon$

where, for i=n observations:

y_i=dependent variable

x_i= independent variables

β_0=y-intercept

β_n=slope coefficients for each independent variable

ϵ=the model's error term or residuals

C. **K-Nearest Neighbor** K-nearest neighbor (KNN) is a supervised machine learning algorithm used for classification and regression problems. It is based on the theory of similarity measuring. Therefore, to predict a new value, neighbors should be put into consideration. KNN uses some mathematical equations to calculate the distance between points to find neighbors. In a regression problem, KNN is used to find the mean of the k labels. While in classification problems, the mode of k labels will be returned.

D. **Random Forest** The random forest algorithm is one of the most popular and powerful machine learning algorithms that is capable of performing both regression and classification tasks. This algorithm creates forests within number of decision reads. Therefore, the more data is available the more accurate and robust results will be provided. Random Forest method can handle large datasets with higher dimensionality without overfitting the model. In addition, it can handle the missing values and maintains accuracy of missing data.

References:

1. J. Han, and M. Kamber, "Data Mining: Concepts and Techniques, 2nd edition", Morgan Kaufmann Publishers, 2006

2. J. W. Seifert, Data Mining: An Overview, C-RS Report for Congress, Dec. 16, 2004, www.fas.org/irp/crs/RL31798.pdf.

3. G. Ganapathy, and K. Arunesh, "Models for Recommender Systems in Web Usage Mining Based on User Ratings" Proceedings of the World Congress on Engineering, Vol. I WCE 2011.

4. D. Mican, and N. Tomai, "Association-Rules-Based Recommender System for Personalization in Adaptive Web-Based Applications" http://gplsi.dlsi.ua.es/congresos/qwe10/fitxers/QWE10-Mican.pdf.

5. S. Liao, T. Zou, and H. Chang,"An Association Rules and Sequential Rules Based Recommendation System", Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference, 12-14 Oct. 2008.

6. Q. Li, and B. M. Kim, "Clustering Approach for Hybrid Recommender System" Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03), 2003.

7. S. Nadi, M.H. Saraee, and A. Bagheri, "Hybrid Recommender System for Dynamic Web Users", International Journal Multimedia and Image Processing (IJMIP), Vol. 1, Issue 1, March 2011.

8. S. Tiwari, "A Web Usage Mining Framework for Business Intelligence", International Journal of Electronics Communication and Computer Technology (IJECCT) Vol. 1 Issue 1 Sep.2011.

9. X. Su, and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques" Advances in Artificial Intelligence Volume 2009, Article ID 421425.

10. J. A. Freeman, and D. M. Skapura, "Neural Networks: Algorithms. Applications. And Programming", Addison-Wesley Pub (Sd), June 1991.

11. E. Gottlieb, "Using integer programming to guide college admissions decisions: a preliminary report", Journal of Computing Sciences in Colleges, Volume 17, Issue 2, Pages: 271-279, 2001.

12. I. Hatzilygeroudis, A. Karatrantou, and C. Pierrakeas, "PASS: An Expert System with Certainty Factors for Predicting Student Success" Knowledge-Based Intelligent Information & Engineering Systems 2004. www.informatik.uni-trier.de/~leydb/conf/kes/kes2004-1.html.