Research Article

STATISTICAL ANALYSIS USING DATA MINING TECHNIQUES

Jay Mehta., Avadhut Mainkar and Seema Yadav

K. J. Somaiya Institute of Engineering and Information tech. Mumbai, India

ARTICLE INFO	ABSTRACT
	This paper is based on web based application which is used to predict and suggest different institutes of engineering for admission just by providing basic information of the aspiring students. Engineering is evergreen, advanced and one of the fastest growing fields of technology and education background. Students facing many problems to decide which branch to choose, which institute is better for particular branch and to search latest updated information of colleges. This system is providing all the important information to students by performing statistical analysis on the updated information of many colleges from different regions of Maharashtra. Clustering and classification algorithms are used to make process more reliable in flexible. By using this system students get to know about pass year cutoffs of colleges, placement information, important details about scholarships etc. Recommendation system is also very helpful to find out best fit college according to student academic details.

INTRODUCTION

In the growing fields of data mining, its applications are also increasing widely.

Wikipedia definition of data mining is the practice of examining large pre-existing databases in order to generate new information.

Different data mining techniques are:

- Classification.
- Clustering.
- Regression.
- Anomaly detection.
- Association rules.
- Reinforcement learning.

There are more than 100 institutes which are providing engineering degree of around 40 different branches. Engineering aspiring students may be not familiar with different rare branches. We are developing a web portal consisting of all the information of all the institutes in many regions of Maharashtra.

Large volume and high speed may bring heterogeneity, noise accumulation, counterfeit relations and random endogeneity, creating issues in computational feasibility and algorithmic stability. Either with newly developed statistical methodologies and/or computational methodologies can be approached to these obstacles.

Different colleges are ranked on basis of different parameters like college grade, location, placement policies, performance of students etc. System includes information like fee structure, placement history, students review, past year cutoffs, location etc.

This system will also recommend a list of shortlisted colleges according to their interest, location, preferences and performance in their curriculum. Step by step online process of admission and scholarship finder for deserving students will also be there in the system.

The flow of the system will go like this

All the data of different institutes of various branches will store in the system. Along with this scholarship finder is also set off for the students.

Clusters of the colleges will be created by their location, branch, grade etc.

User needs to give their information and their preferences to choose colleges and accordingly system will recommend the shortlisted colleges along with all the information so that students understand it in better way.

College reports will be created on different parameters to help students to choose institute accordingly.

Corresponding author: **Jay Mehta**

K. J. Somaiya Institute of Engineering and Information tech. Mumbai, India

Survey

Based on the reports and investigation of the previous system, this system has some additional changes than the previous one.

Data mining methods are used for the college recommendation system. Random forest have been considered as most effective model as it yields the highest accuracy of 54.12%. Previously system only using the rank scored by the student to predict the list of institutes that student can apply for, but in this system different parameters are used like placements policies, locations, grades of the institute etc. to predict the more preferable institute for the students.

Students only have to provide some basic information to the system to make the system more reliable and more interactive. The new feature can be add to the system which is scholarship finder which is used to help the students to make aware of different kind of scholarship for which they can apply for according to various eligibility criteria. Every information of all the institutes, students can get at one place very easily using this system, which is not present in the previous system. Different statistical formals and hypothesis can be used to make the results more accurate and precise.

Front end, Interface and database these three modules are very much important in the system.

Technology used in existing system

In the previous existing system, [3]analysis was done only on ranks of the students. In our system statistical analysis is done by considering different parameters like cutoff marks, location, grades, placement policies etc. various kinds of hypotheses can be created to apply statistical formulas and get accurate results.

Initially student has to give their academic information and then system will do analysis on that data. The list of institutes can be classifying using Naïve Bayes classification algorithm.

Data can be clustered on basis of location of institutes by using Random Forest algorithm. By doing statistical analysis system predict the number of institutes on the basis of entered preferences by the student.

By considering above all the parameters rating of the institute can also be done which will out of 5. List of institute will be show according to this rating only.

In the system, there is another feature of scholarship finder is present which is used to provide the information of all the scholarship details where particular students can apply on basis of different criteria.

Report generation of all the institutes can be done by using Business Intelligence tool which will help the student to understand the details of institutes very clearly.
Prediction done by the system is more accurate than existing system because of statistical analysis and advanced algorithms.

System architecture

System is based on client-server architecture. User first need to log in to the system, then user will provide require information to the system to get the list appropriate list of institutes. System will analyze the data and process it with the data which is already stored in the database and by doing statistical analysis most accurate and reliable results can be given by the system.

Data mining techniques are used for classification and clustering of the data to process the data faster and to avoid redundancies. Business Intelligence tool is used for the report generation of every institute by considering different parameters as placement, cutoff marks, category etc.

The system architecture is consisting of three phases which are client based, server based and database.
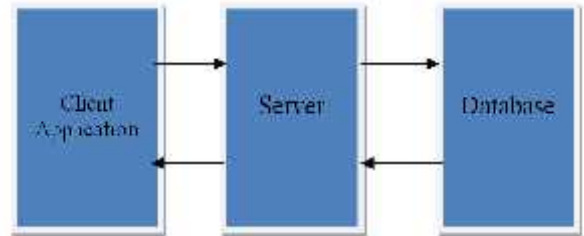
System Archiecture

In the very first stage, at client application the user will login to the account and then user will provide some basic academic information to the system and then system will do analysis on that data.
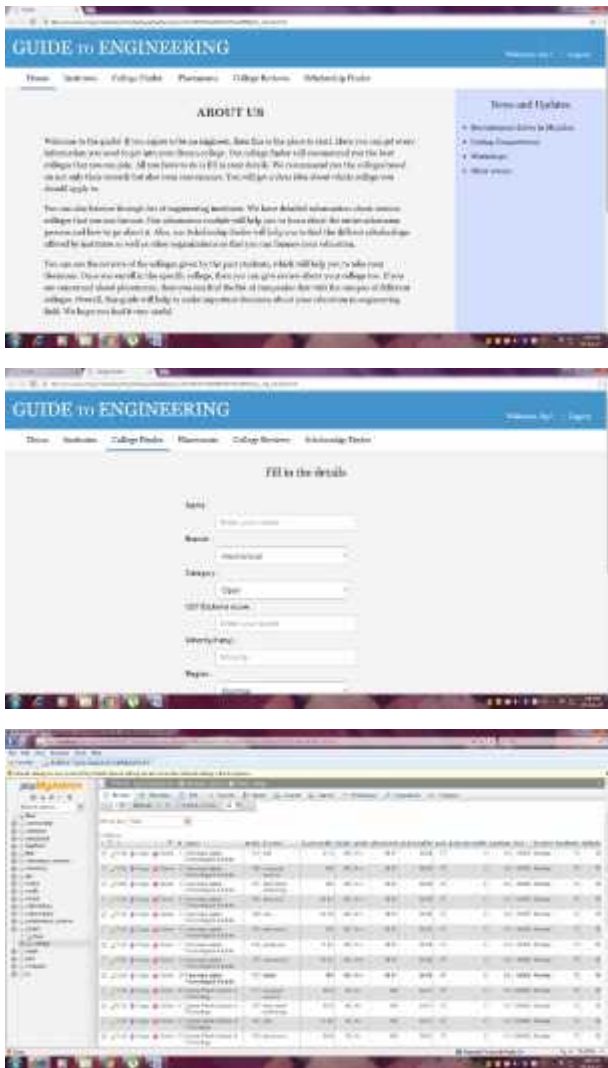
The best analytic techniques are run on server side which are data mining techniques including exploratory data analysis and preprocessing, frequent patterns mining and associations, classification /characterization, clustering and outlier deduction. The analysis on data that is, data mining algorithms and statistical analysis in done at the server side and then results will be display to the user. Server will request for the stored data of institutes to the database and given data are processed together.

Database section has already store all the information of all the institutes so that it can be given to server as per requirement. This information will consist of cutoff marks, intakes, grades, performances and many more different parameters for the reliable analysis.

Comparison between existing system and proposed system

	Existing system	Proposed System
1	Algorithm used is not much reliable.	Algorithms are very reliable and accurate.
2	No hypotheses are created for statistical analysis.	No hypotheses are created for statistical analysis.
3	In the existing system, only rank of the student is considered as parameter for analysis.	In this system, cutoff marks, location, grades, placements like this all parameters are considered for better results.
4	Data mining techniques are not used.	Data mining techniques are used to process the data faster.
5	Predictions which are obtained are less reliable; operations are more complex and time consuming.	Predictions which are obtained are more reliable and accurate; operations are quite simple because of data mining and statistical analysis.
6	Feature of scholarship finder is not present in existing system.	Scholarship is present which gives information of all scholarship based on different criteria.

Screenshots of User Interface and Stored Database

Institutes getting rating out of 5 by considering parameters like placements, grades, intakes and many more. Reports can be generated by BI tool.

Scholarship finder is another feature which is helpful to find out different scholarships for the students who are eligible by various criteria.

Statistical analysis will give more reliable and accurate prediction of the system using data mining techniques.

Acknowledgement

References

1. M.S.Hajirahimova, A.S.Aliyeva "Reviewof statistical analysis of high dimensional data" Republic of Azerbaijan, April 2015.
2. "Intelligent data prediction using predctive data mining techniques" By Pankaj Kumar, Anju Bala, CSED Department, Thapar University, Punjab,India. ICICT International Conference on 26th August 2016.
3. "MapReduce: Simplified Data Processing on Large Clusters" By Jefferey Dean & Sanjay GEmawat, Communication of the ACM, January 2008/vol 51, No. 1.
4. "Electricity consumption analysis by data mining techniques" By Usman Ali, Concettina Buccella, Carlo Cecati, Italy. IECON 2016on 22nd december 2016.

CONCLUSION

This is web based application which is very useful for engineering aspiring students to find out appropriate list of institutes from all the institutes. The predicted institutes can be obtained by using various algorithms and other techniques.

****

How to cite this article:

Jay Mehta., Avadhut Mainkar and Seema Yadav.2017, Statistical Analysis Using Data Mining Techniques. *Int J Recent Sci Res.* 8(1), pp. 15358-15360.