CHAPTER 7
CONCLUSION

7.1. Conclusion

The main objective of any higher educational organization is to provide quality education with improved quality of managerial decisions. In many places existing education system has just not been capable to fulfill the industrialized demands. Accurate prediction of student’s success in higher learning institution may become a great approach to reach the highest level of quality in higher education system [83]. In this work to enhance the quality of education select data mining. In educational area data mining applications are very well known. There are many prediction models available with different approaches. There is no certainty if there are any predictors that accurately determine whether a student will be an academic genius, a drop out, or an average performer.

This work presents steps involved in a data mining process, providing both theoretical and practical contributions. The context of work is classification of large volume of noisy data, which may contain irrelevant and/or redundant pieces of information. However, the correct identification of underrepresented classes is generally of increased importance, which results in different errors possessing different degrees of gravity.

Create data set on the basis of collected data is first objective of work. The present thesis provides extensive systematic different categories of data set. Desired data collection was a big problem in data analysis. To achieve this objective data has been collected from college, and some Govt. schools so that work examines on different parameters. This work analysis on different data set, different parameters to performance prediction. From the data set of college students received parameters of present result for prediction. School data provided parent qualification, living area, caste, etc. Almost methods need different parameters to present result. Collect 150 data from a private educational organization through newly enrolled students. This
thesis presents that parent qualification, living location, previous grade (Chapter 3), class 12th result, caste (Chapter 4), attendance in present class (Chapter 5), sport interest, parents income, class test performance (Chapter 6) can be affect the students’ performance. To analyze data some times it is necessary to transformed qualitative data into categorical form. Data transformations have been done by Ms-excel. It proved facility to covert data range into required format. Work formed a fuzzy set with linguistic modifier (good, vgood, excellent, outstanding), into numerical value transformed data (Figure 4.3), (Figure 5.3), (Figure 6.2) through excel.

Attributes extraction which effect performance is extremely backbreaking and domain dependent task, resulting in situations in which not all attributes recorded are relevant for the classification task, while some attributes convey redundant information. Irrelevant or redundant attributes have been shown to harm the classification process. Apply validation rule on data set to eliminate repeated data and missing data is next objective so that get the validate data for work. Present work has established the fact that the method achieves better stability than individual feature selection performed via different methods, while reducing the number of attributes significantly. In chapter 3 excel used for data filtration to eliminate empty cell or missing data. Chapter 4 and chapter 5 used filter operator of rapid miner. This operator takes data set as input and returns a new example set including only those examples that satisfy the specified condition. Rapid miner provides several predefined conditions as compare to excel. Users can select any of them and can also define their own conditions to filter examples. Filter operator may reduce the number of examples in an example set but it has no effect on the number of attributes. The select Attributes operator is used to select required attributes. The Filter Examples operator is frequently used to filter examples that has or do not have missing values.

The existing records are commonly incomplete. The popular classification techniques have not been designed to deal with missing values, and often employ basic and inefficient approaches to deal with such issues. Thesis use validates and handles missing data set by cross validation using rapidminer. Validated data provide more accurate result compare to non-validate data. With 10 cross-validation, divide it just once, but in 10 folds divide into, 10 pieces. For training uses 9 of the pieces, and the last piece use for testing. Perform the whole thing for 10 times and every time use
different segment for testing (section 3.5 and 5.2). In cross validation technique some part of the data set is recurrently and analytically used to test the suitability for the model. Cross-validation conducted at multiple levels (section 6.1.2). Thesis used rapid miner tool to apply data validation. The validation operator of expects an example set for training a model i.e. training data set. The training process must return a model, which is accomplished on the input example set. The example set that was given as input is without changing passed for the output to the output through validation tool. This is usually used to reuse the same example set in further operators or to view in the results workspace. The testing sub process returns a performance vector. The methods presented in this work have been successfully validated on a number of standard real-world datasets that is stridently employed by real-world scenarios.

Build model for early prediction of study outcome using the student enrolment data and construct a prediction model using classification data mining techniques on the basis of data set is main objective. To accomplish main objective in chapter 3 after data preprocessing and transformation form a decision tree on the basis of information gain to select root node (section 3.2) and select best split point on the basis of split information (Table 3.4). Model evaluated by confusion matrix (section 3.3). Model implemented through weka tool. It provides 79% accuracy that is acceptable and proved that student’s performance is affected by qualitative factors. Thesis presents others qualitative data also affects performance.

Fuzzyfication and defuzzyfication can improve model performance (Chapter 4). Convert data into rough fuzzy set through triangular membership (section 4.4). Output calculated through membership value. Through this model work shows that any subjective variable can be added in fuzzy logic approach. Subjective variable also affects the performance of students. Work presents that performance of any student affect by previous grade and caste also (Figure 4.10 and 4.11) through matlab.

Combination of fuzzy logic and value of information gain can generate more accurate model. Convert data into fuzzy data and calculate entropy of attributes to generate a fuzzy decision tree. Generated fuzzy decision rule (section 5.3) shows how much students attendance and previous result is play role in student’s performance. Performance vector (section 5.4) shows model accuracy is more than 86%, which is
higher then non-fuzzy set. An educational field can provide huge data and numerous attributes also. Combination of different attributes and large data may affect the prediction result even possibilities to received more accurate result. After transforming data into fuzzy values to find out more significant attributes calculate weight of attribute using chi square test (section 6.2.2). After Calculating weight (Figure 6.3) apply validation and CHAID to construct Decision tree. CHAID will "build" non-binary trees (Figure 6.4). Performance vector, which shows accuracy of model, is 99% and classification error is 0.89% (Figure 6.7). Chapter 6 also presents correlation of different attributes. It shows how much an attribute is associated to each other. Implemented this model using rapid miner. To identify factors which affect the students’ performance is aim of model. Task represents through decision tree that previous grade and living location are the important factor for performance of student (decision tree Figure 3.3).

The tree produces a graphic representation of all possible outcomes, returns and follow-up decisions in one document. Each subsequent decision resulting from the root is also depicted on the tree, so you can see the overall effect of any one decision. As go through the tree and make choices, you will see a specific path from one node to another. It seems from Figure 4.10 and 4.11 that include a subjective variable in fuzzy logic approach. Subjective variable also affects the performance of students. It can see from Figure 5.6 somewhere students living location; previous result and even caste status may impact the performance. Decision tree generated by fuzzy set is more accurate than non-fuzzy set shown in chapter 5. Figure 6.2 and Figure 6.3 shows that if attendance of a candidate average then most probable performance will be poor specially when candidate from rural or semi urban area. If attendance is good then performance value will be increases. Figure 6.5 shows fuzzy performance rule or decision rule, extract from decision tree. All parameters are not plays vital role in performance prediction it can be easily determine by decision tree.

Result presentation is essential task to beneficiate user from the result. Thesis presents result in form of decision tree (figure 3.3, figure 5.5, Figure 6.5). Main motivation behind used decision tree is it visualizes the solution. Users can follow the path through tree. Decision Trees (DT) are like those used in decision analysis where each non-terminal node represents a test or decision on the data item considered.
Depending on the outcome of the test, one chooses a certain branch. To implement model different tools has been used in thesis; Tools provided graphical user interface so produced result easily understandable by any user. The decision tree generated by weka and rapidminer tool is shown through diagram (Figure 3.3, Figure 5.5). Another tool matlab produced out put shown in the form of graph (Figure 4.11), which shows higher impact of the attributes in prediction. Decision rules shown by Figure 6.4. In future software on the basis of model it can be easily understood by any user may be academicians, teachers, students, and management.

![Comparison with different used methods](image)

*Figure 7.1 Performance Comparison graph of different model*
Work covers all objectives and on the basis of work it corroborated that

(i) Fuzzy model best as compare to non-fuzzy model.
(ii) Weighted fuzzy decision tree give better performance.
(iii) Correlation shows that many factors may correlate to other.
(iv) Model is data sensitive. Model gives better result on numerous data set
(v) Researchers can add various further attributes in the model.
(vi) The proposed algorithm is efficient for extracting accurate and graspable set of rules from the students' dataset because it extracts only rules in which the originator satisfy the conditional probability within certain class.

Figure 7.1 shows comparisons between different model and base paper’s model. It is clear that weighted fuzzy decision tree accuracy wise give better performance and model can be used to discern innovative relationships that may exist in the data and possibly to improve the simplification of solution. One of the base paper analyzed the data set that containing information about students, such as gender, marks scored in the board examinations of classes X and XII, marks and rank in entrance examinations and results in first year. By applying the ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms on this data, they have predicted the general and individual performance of freshly admitted students in future examinations and the model was non weighted decision tree which gave 75.145% accuracy [5] in work it also shows with one of model with different attributes which near about similar. Presented weighted fuzzy decision tree gives 99% accuracy but it is data sensitive model. With the number of increased attributes and data set model accuracy may vary.

7.2. Tools Used In Thesis

WEKA: Waikato Environment for knowledge analysis (WEKA) is a popular suite of machine learning software (Chapter 3). It is a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions [50]. Present version was mainly designed as a tool for different application areas, in particular for educational purposes and research. Weka is free available under the general public license and ease of use due to its graphical user interfaces.
Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported) [47]. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

MATLAB: Matlab is basically a package for numerical computation and visualization, with a large number of add-on toolboxes for specialized tasks. As a numerical tool, it doesn't deal with categorical values, which is something to keep in mind for the target classes in general [45]. So that used fuzzy sets in chapter 4. Matlab has command line interface. Basically the matlab platform is optimized for solving engineering and scientific problems. The matrix-based matlab language is the world’s most natural way to express computational mathematics [28]. These matlab tools and capabilities are all strictly tested and designed to work together.

RAPIDMINER: RapidMiner is a software platform provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including data preparation, results visualization, validation and optimization [52]. Rapidminer provides data mining and machine learning procedures with data loading and transformation. With rapidminer researchers can perform data preprocessing and visualization, predictive analytics and statistical modeling, evaluation also (Chapter 5 and Chapter 6). Rapidminer is written in the Java programming language. Rapidminer provides a GUI to design and execute analytical process they consist of multiple operators [43]. Each operator is performing a single task within the process and the output of each operator forms the input of the next one. Rapidminer provides learning schemes, models and algorithms and can be extended using R and Python scripts. A huge amount of visualization techniques approximately 300 provided my rapidminer.
Performance prediction model have potential to provide a wide place e learning in India. In future researchers can place a comparison between performance e-learning and traditional learning. Among them predictive modeling has particularly interesting applications on predicting effectiveness of e-learning systems. Apart from them many open research challenges that exist in electronic learning is particularly centered on data privacy and ethics. At the time of deployment of data mining techniques within commercial e-learning systems requires a joint effort of ICT specialists, educationists and the learners.