Student Progression System using Descriptive and Predictive Analytics

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Software Engineering

Submitted By
Aanchal
(Roll No. 801631001)

Under the supervision of:
Harkiran Kaur
Lecturer
Computer Science and Engineering Department

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA – 147004

July 2018

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Student Progression System using Descriptive and Predictive Analytics*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Ms. Harkiran Kaur* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

(Aanchal Phutela)
(801631001)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Ms. Harkiran Kaur)
Lecturer, CSED

ACKNOWLEGDEMENT

I would like to thank god for blessing me with all the strength and resources required to complete this task. I would like to express my deepest gratitude to Ms. Harkiran Kaur, Lecturer, Department of Computer Science and Engineering for guiding me through the whole process, for her valuable suggestions and continuous help and providing me with your knowledge and experience. I am extremely thankful to her for sharing her knowledge and providing me her valuable time and timely assistance, which contributed a lot in successful completion of this thesis work. She has been a constant source of motivation for me, throughout this work.

I am heartily thankful to Dr. Maninder Singh, Associate Professor and Head, Computer Science and Engineering Department for his motivation, providing guidance and support and giving valuable suggestions for this work.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection.
Last but not the least I would like to thank my family for their support and encouragement without whose blessings none of this would have been possible.

Aanchal
(801631001)

ABSTRACT

Analytics is a practice of dividing a problem into simpler parts and then use their applications to make decisions better. It is only a way of thinking not a tool or technology. It has diverse applications across the globe as education, retail, marketing, gaming and health care. Data analytics play an important role in any organization, based on relevant facts that will allows making a better decision. Progression of students greatly affects the educational organization's future. Analysis of the academic dataset could reveal important insights, which if properly used can help students for their progression. In this work, the Descriptive and Predictive Analytics has been applied on the Academic dataset of students. Descriptive Analytics is the summarization of the past data and generates some useful patterns from that data. This work focuses on analyzing and querying large academic dataset for generating Student Progression using visualization and dashboards. Presently projects on Progression Systems exist but no descriptive or predictive analytics has been performed on these datasets. The proposed system collects data from different departments of University, store data into the large data warehouse of the University and generate validated set of Key Performance Indicators (KPIs), based on the past dataset of student Academic. These KPIs have been obtained after applying Statistical techniques on various sets of dimension on the academic datasets. After completion of this step, the focus was on predicting the performance of the student's by using cluster analysis approach. For this work, the cube designing has been done and three clusters have been fabricated by using k-means clustering algorithm. These clusters placed the similar attribute in one class and the objects with different attributes in other class, on which various models have been applied for predicting the performance of the students. The deep learning model has given the highest accuracy that is 99.02% in comparison to all other models such as Naïve Bayes, Decision Tree and Linear Discriminant Analysis (LDA).

TABLE OF CONTENTS

S. NO.	TITLE	PAGE NO.
	CERTIFICATE	i
	ACKNOWLEDGEMENT	ii
	ABSTRACT	iii
	TABLE OF CONTENTS	iv-v
	LIST OF FIGURES	vi-vii
	LIST OF TABLES	viii
CHAPTER 1- INTRODUCTION		**1**
1.1	Data Analytics	1-6
	1.1.1 Overview of Descriptive Analytics	
	1.1.2 Overview of Predictive Analytics	
	1.1.3 Overview of Prescriptive Analytics	
1.2	Key Performance Indicators (KPIs)	7
1.3	KPIs Classification and Identification methods	7-8
1.4	Organization of Thesis	9
CHAPTER 2 – LITERATURE REVIEW		**10**
2.1	Based on various techniques of Descriptive Analytics	10-11
2.2	Based on KPIs	11-13
2.3	Based on OLAP	13-16
2.4	Based on Predictive Analytics	16
CHAPTER 3- RESEARCH GAP AND PROBLEM STATEMENT		**17**
3.1	Problem Statement	17
3.2	Research Gap	17-18
3.3	Objectives	18
3.4	Methodology	18-19
CHAPTER 4 – PROPOSED WORK		**21**
4.1	Evaluation of applied Statistical technique and select the KPIs	21-26
4.2	Generating Cube by using OLAP technology	26-29
4.3	Descriptive Analysis of the Cube	29
4.4	Cluster Analysis	30-37
4.5	Predictive Analytics	37-43
4.6	Results and Inferences	43-45

CHAPTER 5 – CONCLUSION AND FUTURE WORK 45

REFERENCES 46-51

LIST OF PUBLICATIONS 52

List of Figures

FIGURE NO.	TITLE	PAGE NO.
Fig 1.1:	Dimensioning of cube	4
Fig 1.2:	Predictive Approach	5
Fig 1.3:	Filter Methods	9
Fig 1.4:	Wrapper Methods	9
Fig 1.5:	Embedded Methods	10
Fig 4.1:	Results of ANOVA test for regularity and Extra Curriculum Activities	23
Fig 4.2:	Results of ANOVA test for regularity and CGPA	24
Fig 4.3:	Results of ANOVA test for Semester in the course and CGPA	24
Fig 4.4:	Results of correlation test for regularity and Extra Curriculum Activities	25
Fig 4.5:	Results of correlation test for regularity and CGPA	25
Fig 4.6:	Results of correlation test for No_of_Sem and CGPA	26
Fig 4.7:	Academic dataset stored in MS-Access	27
Fig 4.8:	Cube query design	28
Fig 4.9:	Cube structure design	29
Fig 4.10:	Processing of cube	29
Fig 4.11:	Cube file created	30
Fig 4.12:	Cube visualization	30

Fig 4.13:	Cleaned Data	31
Fig 4.14:	Replace Missing Values model	32
Fig 4.15:	Clustering distance model	33
Fig 4.16:	Performance of clusters	34
Fig 4.17:	Graph shows the frequency of performance	35
Fig 4.18:	Scattered graph between cluster and performance	36
Fig 4.19:	Graph between project vs. research work	36
Fig 4.20:	Plot Graph for all Clusters	37
Fig 4.21:	Frequency of each label shown in histogram	38
Fig 4.22:	Predictive Analytics task sequence	40
Fig 4.23:	Decision Tree model	40
Fig 4.24:	LDA model	41
Fig 4.25:	Naïve Bayes model	41
Fig 4.26:	Deep Learning model	41
Fig 4.27:	Accuracy of Deep Learning model	44
Fig 4.28:	Confusion Matrix	44
Fig 4.29:	Prediction results of student performance	45
Fig 4.30:	Accuracy graph of models	45

List of Tables

S. NO.	TITLE	PAGE NO.
Table 4.1	CATEGORY-WISE CANDIDATE KPIs	20
Table 4.2	CONDENSED LISTS OF KPIs	25
Table 4.3	ATTRIBUTE PERFORMANCE VALUES	38

CHAPTER 1

INTRODUCTION

1.1 Data Analytics

Data analytics refers to the process of cleansing, transforming, classifying and modeling the dataset. Then, this data can be used to analyze organization behavior by recognizes patterns from it and extract information for analysis required for an organization. As the data is qualitative and quantitative in nature, it has been used to develop the productivity of the organization. The powerful illumination of given data has been given [30], which can be described as: Descriptive Analytics, Predictive Analytics and Prescriptive Analytics.

1.1.1 Overview of Descriptive Analytics:

Descriptive Analytics as the name propose, 'describe' or summarize the data into some constructive information and possibly formulate the data for further analysis that is comprehensible by humans. Descriptive Analytics is the starting stage of the analysis, where analysis has been performed on large dataset. The three types of Descriptive statistical techniques used for the analysis that are Univariate, Bivariate and Multivariate. Descriptive Analytics can be performed on the past datasets and has provided static view of data, as the data have increased various analytical techniques has been performed such as Classification, Clustering and Categorization for analysis.

Descriptive Analytics applies two major techniques, namely data aggregation and data mining to account past events [15].

i. **Data aggregation** is an information mining process that collects the information from various databases, compile that information and prepare it for the further processing or for human analysis.

ii. **Data mining** simply states, extorting or "mining" information from large quantity of data. For example: Data mining includes performing market analysis to identify new products or cross-selling to existing customers. Therefore, data mining also appropriately named as "knowledge mining from data," [15].

Generally IT systems have been divided into OLTP (Transactional) and OLAP (Analytical). OLTP stands for Online Transaction Processing, is a relational database used to support online transactions on the internet. The main emphasis of the OLTP systems is fast processing of query and data integrity. OLTP systems are the actual source of the data whereas in OLAP data comes from various OLTP databases. Also, the transactions in OLAP systems are very less and have very complex queries which include aggregations like sum, count, maximum or minimum. Also in OLTP the entity model (usually 3NF) schema is used whereas in OLAP multidimensional schema (usually star schema) is used.

Data have stored in information repositories and many different databases. So, one may need a repository where all the data from various databases have been gathered. From there the Data Warehouses (DW) is emerged, in which several heterogeneous data sources are organized beneath one unified Schema.

Bill Inmon defined Data Warehouse as "A warehouse is a subject-oriented, integrated, time variant and nonvolatile collection of data in support of management's decision making process" [33]. Also, Ralph Kimball illustrated that "A Warehouse is a copy of transaction data specifically structured for query and analysis" [20]. Data Warehouse provide various functionalities like cleaning of data, integrating of data and transformation of data which are important preprocessing steps of data mining.

Data Warehouse provides analysis of multidimensional data for which OLAP tools facilitate effective data mining. OLAP is an Online Analytical Processing which allows users to analyze the different dimensions of multi-dimensional data as shown in Fig 1.1. Basically it provides the analysis of the data, stored in large databases or data warehouses.

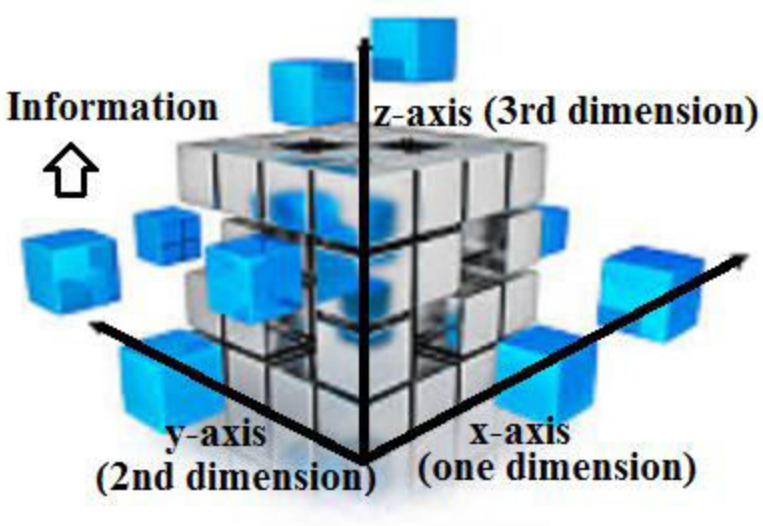


Fig 1.1 Dimensioning of cube

OLAP includes analysis techniques such as summarization, consolidation, and aggregation [15] and provides the capability to examine the information from various perspectives. For in-depth analysis various data analytics tools are requisite and analysis techniques include classification, clustering and the characterization of data [15]. Several other data mining functionalities, such as association, classification, prediction and clustering [15] were incorporated with OLAP operations to improve interactive knowledge.

OLAP consists of many operations which can be performed on the data, which include: slicing, dicing, roll-up, roll-down and pivoting [31]. OLAP and data warehouse are based upon multidimensional model. In this model, the data can be viewed in the form of data cubes that allows the users to analyze the data from multiple perspectives. They are defined using dimensions and facts.

i. Dimensions are real-world entities and Facts are the numerical measures, which can be calculated. For Example: Regularity, Projects and All Rounder Score are some dimensions and Present CGPA is a measure which has been calculated from these dimensions.
ii. A dimension also includes various attributes, levels and hierarchies. For instance, a student dimension's attributes could include first and last name.

iii. A hierarchy is many to one relationship between members of a table or between tables. For illustration, one feasible hierarchy in the date dimension is Day > Month > Quarter > Year. Every different hierarchy acts a level for cube. For example, year is one level, quarter is another level etc.

1.1.2 Overview of Predictive Analytics

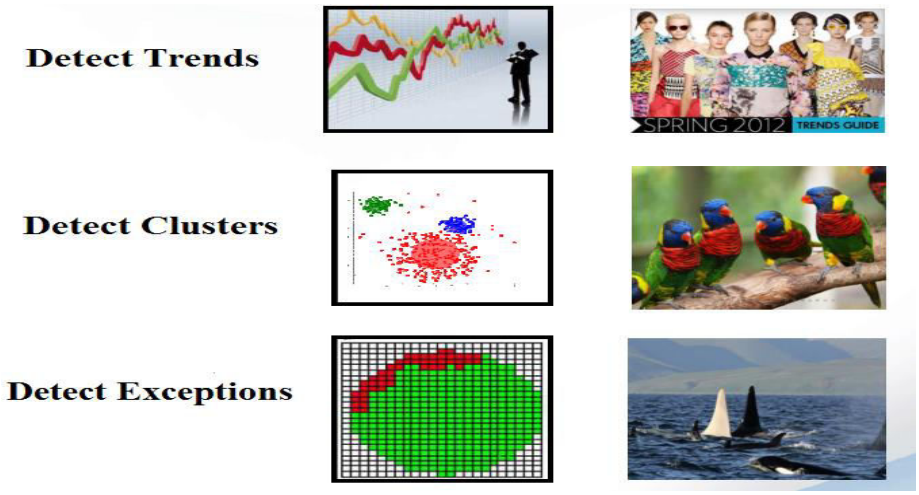


Fig 1.2: Predictive Approach [5]

Predictive analytics is an analytical technique which is used to predict the future outcomes. In this approach, various Statistical and Machine Learning approaches have been used to determine the outcomes of the future as shown in Fig 1.2 based on the past datasets. This analytical approach has been used to make prediction on academic dataset. To discover the future trends some Machine Learning algorithms and techniques have been used in this work. Predictions have been done by on exploratory data and finding patterns or associations in this data based upon the past trends and then concludes their associations. By applying various predictive models on multiple attribute of academic dataset, some valuable results have been achieved. Predictive Analytics only tell us the future outcomes or the possibility of future, depends on the accuracy given by the models have applied on different kinds of datasets.

Common applications of Predictive Analytics are recommendation engines, customer segmentation, socio-network analysis, medical science images, anomaly detection and

many more. Predictive Analytics mainly contains three broad buckets: Clustering, Prediction and Association. Using the techniques for each bucket, organizations may have deep insight about their data. In this work, the clustering has been used as an input for performing the Predictive Analytics on their dataset.

Clustering is a technique in which entities with similar traits have grouped together in one cluster and entities with divergent traits in another clusters and the analysis of these clusters is known as cluster analysis. This concept implements very straightforward idea that is, to split the data into different groups and make sure that every group contains similar objects. There are some common techniques for clustering:

i. Hierarchical Clustering: The data points that belong to a child cluster also belong to the parent cluster, called as Hierarchal Clustering. In this the hierarchy of clusters has been build as the two nearest clusters has been merged into one cluster until there is only one cluster left [22].
ii. Partitioning Clustering: There were basically two kinds of partitioning techniques such as: strict partitioning and strict partitioning with outliers. In strict partitioning each object belongs to only one cluster and in strict partitioning with outliers the objects belongs to one cluster but some of them can also belong to no cluster and are considered as outliers [22].
iii. Overlapping Clustering: The objects belong to more than one cluster usually involves hard clusters is known as overlapping clustering [22].

The main differences between hierarchical and partitioning clustering are, the hierarchal clustering algorithms were unable to deal with the big data whereas partitioning clustering algorithms can easily handle a huge amount of data. Also, the time complexity for partitioning clustering techniques is linear such as O(n) but it is quadratic for hierarchical clustering that is $O(n^2)$. This was the main reason for choosing the partitioning clustering algorithms over hierarchal clustering algorithms for this thesis work.

Some common clustering models are there which are used immensely are given below:

a. Centroid-based models: In this kind of models, the correlation of data points to the centroid of the clusters is identified and then clustering is done. These kinds of algorithms should be applied when one have former knowledge of their data. Also, this category of models runs iteratively to find local optima. K-means clustering algorithm falls in this category [22].

b. Distribution-based models: In this model, clustering is performed on the basis of the how probable the objects in the cluster belong to the same distribution (for instance: Normal, Gaussian). The distribution models frequently suffer from overfitting [22].

c. Connectivity-based models: In this model, the clusters are created based on the data points closer in data space, data points which are more similar and places in one cluster. These kinds of models are very easy to interpret whereas they are less scalable for handling big datasets [22].

d. Density-based Clustering: In this model, clusters are defined as the areas with diverse density in the data space and grouped the data points within these regions which are closely lay to each other [22].

1.1.3 Overview of Prescriptive analytics:

Prescriptive Analytics is an approach which prescribes the users to take different actions and direct them towards a solution. Mainly, this analytics provides the best advice for the given problem. Prescriptive Analytics is related to both types of analytics that is: Descriptive and Predictive Analytics. Prescriptive Analytics has been working towards the best course of actions based on the prior knowledge of dataset as described.

The decisions made using this approach can help an organization or a business to increase their profits and reduce their risk. Prescriptive Analytics generally gives us the answer of this question- what should we do? It is the new field of the analytics which tells not only what will happen but also tells why it will happen. This field gives the best advice for the future outcomes and also recommends the one or more possible courses of actions. This approach helps the companies to deliver the right product at right time.

1.2 Key Performance Indicators (KPI's)

Some common techniques engaged in Descriptive Analytics are observations, case studies, and surveys. In the proposed work the given dataset is analyzed by applying statistical methods on the academic dataset using IBM SPSS Statistics tool. The main agenda of this study is to create student progression system. For this purpose, the descriptive analytics must be performed on only those features of the academic dataset which have a huge impact on the projected goal and these features are commonly called Key Performance Indicators (KPIs).

In its simplest form, a KPI is a type of measurement that helps you to recognize how your organization or department is carrying out. A good KPI will help you and your team to recognize whether you're choosing the right path in the direction of your planned goals or not. A KPI must be effective if it should follow the SMART criteria. SMART refers to Specific, Measurable, Achievable, Relevant, and Time-bound [1].

The present study performs analytics on academic dataset of students of a University, which can be done by descriptive analytics (the preliminary stage of the data analysis) and makes the summary of historical data and generate some useful information from the past trends in the dataset.

The vital success of an organization is its ability to analyze the data, find some major facts in it and take some actions towards the changes. The task of finding major facts in these dataset is performed by validated selection of KPI. Further the corresponding actions are taken as per the data analytic technique applied on these KPIs contained in the dataset [16].

1.3 KPI Classification and Identification Method

The major issue in KPIs identification is that, there are N numbers of KPIs to choose from. If we are choosing the wide of the mark, then we are calculating something that doesn't line up with our objectives. Every organization needs to analyze the results for their respective business problems and consider just those set of KPIs which are most relevant for an organization's further monitoring, evaluating, planning and decision

making. Therefore, it is a prerequisite to identify a set of KPIs for a specific organization for its mission, vision and values on the subject of an organization's approach. A KPI should be: relevant, realistic, specific, attainable, measurable, and used to identify trends, timely, understood, agreed, reported, governed or resourced [16].

There are several methods and techniques you can use for selecting subset of features which helps your model to perform efficiently. These include: Pearson's Correlation, Linear Discriminant Analysis (LDA), ANOVA and Chi-Square Test. These statistical filter methods have been used as a preprocessing step for identifying KPIs. In this the selection feature is independent from any machine learning algorithm. Fig 1.3 describes the implementation of filter methods.

Fig 1.3: Filter Methods [21]

Some other methods include: Forward Selection, Backward Elimination and Recursive Feature elimination, these are Wrapper Methods. Fig 1.4 demonstrates the processing of wrapper methods, in which learning algorithms are used for the selection of best subset of features [21].

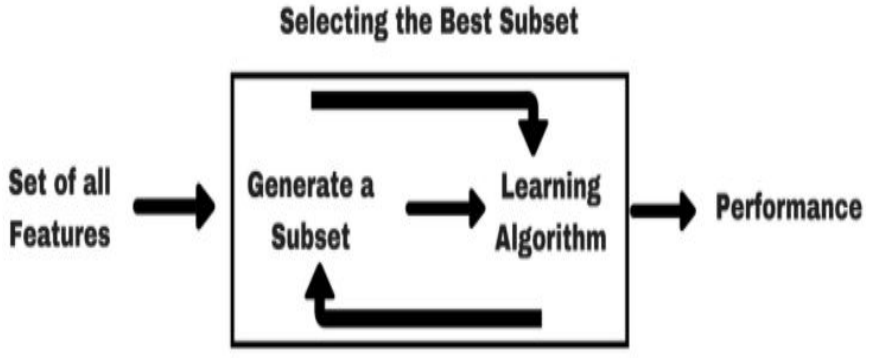


Fig 1.4: Wrapper Methods [21]

There are also some mehods which combine the qualities' of both techniques that are: Filter and Wrapper Methods that are known as Embedded Methods which includes LASSO and RIDGE Regression Methods. Fig 1.5 illustsrates the processing of embedded methods. In which the best subset of features are selected from set of all features.

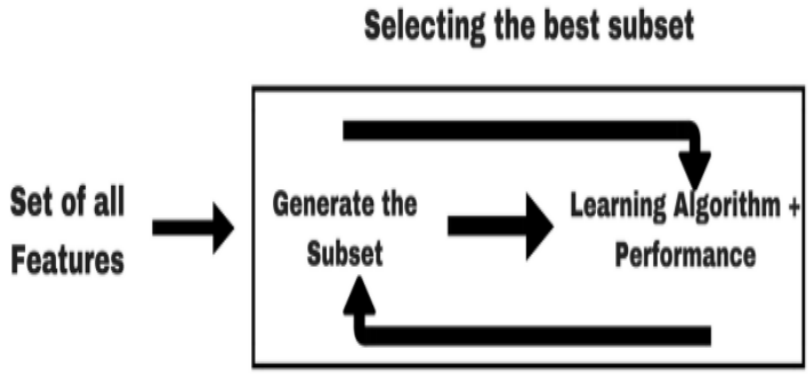


Fig 1.5. Embedded Methods [21]

1.4 Organization of thesis

The rest of the work is organized as follows:

Chapter 2:- This chapter includes the details of the literature survey has been done to study and differentiate the various analytics approaches used in different fields like in business, marketing, education and many more.

Chapter 3:- This chapter provides the details of problem formulation, the methodology used for it, resaerch gap and objectuve of the work.

Chapter 4:- This chapter gives the details of the work that have been implemented during the thesis and proposed the solution for the problem formulated in the previous chapter. This chapter also discuss the results and inferences of the work implemented.

Chapter 5:- This chapter concludes the work that has implemented in the implementation chapter and also tells the future scope of this study.

CHAPTER 2

LITERATURE REVIEW

Literature survey has been conducted on the basis of different parameters includes: different techniques of Descriptive Analytics, based on KPIs, on the basis of OLAP technolgy and also for Predictive Analytics. Various papers have referred for the appropriate knowlegde of the working field.

2.1 Based on various Techniques of Descriptive Analytics

Kandogan, E. in (2012), has explored Just-In-Time Descriptive Analytics techniques for consideration of high level structure of data. And this technique can also be applied in improving convenience visualizations. Just-In-Time is also useful in recognizing and recommending visual. In this paper [23], the author has developed simple algorithms that have detected cluster, outliers and trends, while users have explored the data and at attraction time. These algorithms were specific to Point-Based Visualization [23].

Song, S. K. et. al (2016), have introduced a prescriptive analytic system that is' InSite Advisory' which has helped researchers to advice their future research. This structure contains two parts that were: descriptive analysis and predictive analytics. This is a very useful tool which is never introduced earlier. It has analyzed existing global and domestic position of researchers in their specific research area and provides dictatorial advice according to 5W1H questions [32].

Various approaches of descriptive analytics have compared and while reviewing the literature of the descriptive analytical approaches some comparisons have done, as discussed below:

Mainly in this work, four techniques have compared includes: Decision Rules, Association Rules, Cluster Analysis and Summarization Rules. Every technique comprises some of the applications and challenges in different area of domains. Like Association Rule approach find the patterns in dataset and reveal how the objects have

associated with each other and web mining is especially good quality application of this approach [37]. But challenges faced in this approach were large dataset, redundancy in data, which makes it somewhat difficult. Another technique which has been mainly used for predictions is Cluster Analysis. In this approach the class with similar items has placed in one cluster and with different items has placed in another cluster [18]. The most common example of cluster analysis is whether analysis. But the main challenge for this technique is heterogeneity and size of the data. Exploratory Data Analysis is a superior application of Summarization approach but the uncertainty and partial data is a challenge [37]. Another technique studied was the Decision Rules approach that has used in Credit card companies and Telemarketing, and the choice, ranking and sorting of the decision rules are the main challenges of this approach.

2.2 Based on KPI's

Plandor, D. et al. (2012) proposed an article, in which an application for producing KPIs has been invented. First of all, the results of the company's business diagnostic process were initiated. Formerly, the results were ready for the KPIs analysis, which was stored in the database. The authors' [39] objective was to create a set of KPIs, for which they have used genetic algorithm. A brand new software application was developed as a tool for KPI collection [39].

Krathu, W. et al. (2013), proposed an analysis for inter-organizational relations (IOR's) that were important for collaborations between businesses. In this work, the KPIs for measuring success factors were proposed. For this work, a method was presented for identifying IOR's. The proposed methods took the semantics and data types of both data elements to generate precise results. This technique was applied on real-world industries and offered a set of inter-organizational KPIs. The KPIs offered were used for the evaluation of IOR's [19].

Park, J. et al. (2015), proposed a work for manufactured scheduling it is almost unmanageable for schedulers to study all the constraints. So, the authors [27] proposed an approach that was based on simulation. In this work, a new method was proposed that meet the features of any process by selecting appropriate KPIs. This method was verified

with empirical analysis whether they meet the requirements of KPIs or not. In this the outputs consequent from simulation-based with amendments of domain experts was associated and proposed a framework to identify the appropriate KPIs [27].

Key performance indicator (KPI) framework proposed by Joshi, S. M. et. al [24] have used by an educational institute to meet the developmental necessities across the world. In this work, a performance management system (PMS) was developed. Firstly, it categorized the KPIs in different areas and then its performance was evaluated by faculty to identifying the KPIs. The PMS helped to produce faculty ranking and also helped the institutes to raise their quality standards.

Wudhikarn, R et al. (2017), proposed a work for intellectual capital management for which the Delphi Method have used. Mainly, the research was focused on the identification of important indicators, which were significant for the business performance. Since the research was done on Delphi Method, it helped the organization to deal with the intellectual capital in the business logistics more effectively. This method was superior to entirely depending upon expert's advice and also helped in determining important KPIs in business logistics [41].

Suseno, D. et al.(2017), have proposed a work in which the bonus in enterprise resource planning using KPI has determined. This work, defined how to calculate the bonus by calculating the weight and analyzing the KPIs. In general, there were many ways to determine the bonus in the company but it leads to some drawbacks for the company. This is because there is no clear idea that determined the significant factors for the organization. SMART criteria had used for analysis and determine the bonus in the company. The authors have determined that the bonus depends upon KPI is more resourceful than seniority-based distribution of bonus among the employees. Another observation was that, if the employee can achieve the KPI goal he will get more bonuses [43].

Roberto, P. et al. (2017), had developed a system which is Business Indicator Management which helped them to meet the requirements of information accessibility and dexterity as well. The proposed work had given the integrated approach to manage

the KPIs, through which real-time information about organizations was obtained and further the actual situation of the companies could be analyzed. This helped the organization to increase their productivity and also decision making efficiency [38].

2.3 Based on OLAP

Mirabedini, S (2014), introduced the database of analytical samples that was tested in a university. In this paper, the author came with many problems in education and research institutions across the country. Business Intelligence Strategies, Analytical databases and online tools had been used to overcome this problem. By applying OLAP, they made it possible for managers to look at the different prospective and different level of detailed [26].

Bawane, G. R. et. al (2015), presented an algorithm, Apriory-cube, which is used to discover the association rules in Multi-dimensional datasets. The authors presented an advanced algorithm or traditional Apriory algorithm which was used to integrate OLAP and association rule mining. The authors also increased its efficiency by using new Hash technique. This integration developed a system which provided rules for further analysis to take decision regarding market trend [3].

Du, X. et. al (2015), presented a way to calculate sea water quality and also proposed a framework based on graph OLAP for Multi-Dimensional analysis. The quality of the water was evaluated by calculating the difference between water in different areas. For this work, the authors also constructed monitored area network for evaluation of sea water quality. For this purpose, the Multi-Dimensional analysis methods was used, based on graph OLAP. It was found that, if the density of the monitored sites network at certain monitoring months was the largest, the occurrence probability of large scale marine algal bloom was the biggest [12].

Salem, S. B. et. al (2015), proposed an approach to identify unessential dimensions which could be removed without losing the significant information. For this work, the Genetic Algorithm (GA) theory was used that is extensively used in the area of Artificial Intelligence (AI) to resolved optimization problems. The Multiple Correspondence Analysis (MCA) was also used, which was a statistical analysis method used to calculate

the every solution produced by GA. In this work, both techniques were used to diminish the size of the data warehouse, so that analysts could focus on the appropriate data while exploring the data cubes [35].

Dijiroun, R. (2016), proposed the creation of the cubes based on blending of cubes. This fusion of cubes contains user's based queries. In this paper, the author has deals with the dilemma or designing the data cube according to user query. A tool has developed Design-Cubes-query for presenting cube design and construction for set of existing cubes [13].

Dhanasree, K. et. al (2016), presented the various OLAP technologies and their accessing methods. The authors have designed new translated lattice called Peak Chrome lattice. In this work, the natural indexing was implemented on Peak Chrome lattice which results, reduction in indexing search space, distributed communication cost and search time [7].

Shi, J. et. al (2016), presented a database audit scheme in which they used OLAP technology. The data warehouse has been formed by using logical, conceptual model and ETL. After designing data warehouse, the author used OLAP technology for multi-level audit analysis. The use of this technology enhanced the flexibility and efficiency of database audit work [34].

Azabou, M. et. al (2016), presented new OLAP operators for textual documents. Mainly, the authors have presented three operators that are: To_Semantic operator, S_Drill operator and List_Discriptor operator. As OLAP technology gives us basically Slicing, Dicing, Drill Down, Drill Up operations which are applicable on documents so the authors have presented some new operators which were: To_Semantic operator, S_Drill operator and List_Discriptor operator, for document usage [2].

Fisun, M. et. al (2016), presented an information system for Multi-Dimensional data analysis and data mining. The integration of OLAP and data mining were the main goal of this research. The integrated system, from which intellectual information system was developed, based on object DBMS cache. Also, the system was represented servers and plug-ins developed with the help CSP technology [10].

Zarea, K. et. al (2017), have investigated the incidents of migraine headache and the aspects related to it. For this work, the statistical techniques were used like chi-square test, t-test and logistic regression. The data was analyzed in SPSS tool. The study revealed some factors which affect the growth of migraine, which were stress, sunlight, fatigue, loud noise and overheating. In the results, the authors concluded that fatigue is one of the major factors responsible for migraine headache and its value was 77.8% which was highest from all the factors [40].

Cuzzocrea, A. et. al (2017), focused on the evolving querying encrypted OLAP data problems. In this paper, the problem that was focused was evolved encrypted OLAP cubes and for this work ad-hoc algorithms were used for querying encrypted cubes. The authors also have developed some suitable transformation from/to the encrypted/decrypted domain [6].

Chen, W. et. al (2017), proposed an optimized dispersed OLAP system for big data. Architecture was designed, which contained four modules: Data acquisition, storage of data, OLAP analysis and visualization of data. The design architecture not only supported the Impala but also kylin, the MOLAP engine. After implementation, the performance of the system was evaluated and it was found that the efficiency of the system was drastically better than the traditional system [4].

Fu, L. (2017), proposed a new system called Recommender System-Online Analytical Processing (RS-OLAP), which integrated the functionalities of OLAP to RS. The four new algorithms for RS were proposed. The first one was top-rated items in user's frequent categories (TIUFC) also known as transaction based RS. The two more algorithms for RS were given that are pair-wise association RS (PARS) and RS for spatial items. In PARS, algorithm recommendation is developed based on the association rule mining. And, the last algorithm was proposed for spatial items in which, if the information contains only the location of the item then the top-related item in the same location will be recommended [9].

Tohir, A. S et. al (2017), have proposed a model to solve the problem of leaders or decision makers who were responsible to make any policies for the graduates. For this work, the data was presented by creating the data warehouse in multi-dimensional form that was (OLAP). For the proper presentation of the data, a good design of the data warehouse had been obtained. For this purpose, the authors [42] used nine steps for the designing of the data warehouse and then present the multi-dimensional data to the decision makers [42].

Gutiérrez-Batista, K. et. al (2018), proposed the formation, addition and accomplishment of a new dimension known as Contextual Dimension for the examination of textual data. This dimension was created from the texts obtained from social networks which were integrated in the form of multidimensional model. The creation of the dimension was automated with the help of data mining technique (hierarchical clustering algorithms) which was not dependent from the language of the texts. The methodology used the tools such as Multilingual Central Repository (MCR 3.0) and Wonder (3.0) and experimentation take place on real datasets from social networks [17].

2.4 Based on Predictive Analytics

Saini P. et. al (2014), implemented the Decision Tree based ID3 algorithm by using educational data. This data mining algorithm was used for the classification of the data objects. This algorithm designed decision tree by using breadth-first technique. This study concluded that, this algorithm was perfect for classification of educational dataset [36].

Jain P. et. al (2017), have applied data mining classification models to evaluate the car dataset. These models helped the customers to judge the best car segment. The authors [25] have found that accuracy obatined by decision tree model was 91.12% which was highest among all of these: K-NN, Random Forest, Naïve Bayes and Rule Induction[25].

CHAPTER 3

RESEARCH GAP AND PROBLEM STATEMENT

3.1 Problem Statement

In the last few years the Cube Technology has been applied on various kinds of datasets, as it is a very powerful technology for analyzing the data and helps in better decision making. In the academic dataset, the record of the students has been maintained and performance of the students could be analyzed simply by using OLAP technolgy. By analyzing academic dataset, some strategy was pre-planned for the organization success. In this work, also the predictive algorithms have been applied on academic dataset for predicting performance of the students in future .

3.2 Research Gap

In literature survey, it has been discussed that there exist many fields where the cube technology has been implemented. But this was seen from survey that, there was no Descriptive and Predictive Analytics has been performed on academic datasets. Several research gaps have been found during this study:

i. As observed in the literature survey, presently projects on progression system are available with other University but no descriptive or predictive analysis has been performed on Academic dataset.

ii. Some Descriptive techniques have been applied on student's dataset but no work has done for predicting their performance in future.

iii. Many cubes has been generated for graduates only to analyze the placement in the organisation, but nothing has done to analyse the performance of the students.

iv. Many business strategies have been planned by analyzing the data in multidimensional view, but no planning has been done for academic progression.

v. Many methods have been used in education field but data have not been analyzed in the form of clusters.

3.3 Objectives

The objectives of this thesis are:

i) To conduct literature survey on various datasets on which Descriptive and Predictive Analytics has been applied.
ii) To study the academic dataset and Extract the condensed KPIs from it.
iii) To study and implement Cube technology on extracted dataset or KPIs to obtain visualizations.
iv) To implement the clustering algorithm on the academic dataset and predictive models have applied.
v) To implement various Machine Learning models on the obtained set of clusters to predict the performance of the students.
vi) To assess and compare the performance of the suggested models with available algorithms.

3.4 Methodology

This work focuses on analyzing and querying large academic dataset for generating Student Progression using cube technology and visualizing these progressions in the form of dashboards. After the creation of cubes clustering has been done by using k-means algorithm and further predicting models are applied for predicting student performance. To accomplish the proposed work, the below steps have followed:

i) Conducting literature survey of descriptive and predictive analytics from which various ideas like perform cube technology, clustering and predictive models in education field has generated for analysis of academic data. And based on the review, Descriptive and Predictive techniques have been used to analyze the academic dataset.

ii) Feature selection has been performed before applying analytics on dataset for this, extraction of KPIs has been done. For extracting KPIs, the two tests have been used including correlation test and ANOVA test by using SPSS tool. These tests have been performed on candidate KPIs and condensed list of KPIs has been obtained as shown in table 2.

iii) The cube technology has been applied by using Cube-it-Zero tool on academic dataset and cubes have been generated for visualization.

iv) Further the cluster analysis has been applied on academic dataset by using k-means clustering algorithm and three clusters have been obtained on which prediction has been applied.

v) The result of cluster analysis helps to calculate the performance of the students. After obtaining clusters, various prediction models have been applied including Naïve Bayes, LDA, Decision Tree and Deep Learning.

vi) The results of various models have compared on the basis of their accuracy and the model with highest accuracy has been used for making predictions

CHAPTER 4

PROPOSED WORK

The proposed work applies statistical filter methods: ANOVA test and correlation testing on academic dataset, to obtain validated set of KPIs and further in this, Descriptive and Predictive methods have been implemented such as: OLAP (Cube Technology) and Cluster Analysis on academic dataset, to obtain visualizations.

For this work, these steps were used for obtaining validated set of KPIs for Student Progression System. These steps are as follows:

4.1 Evaluation of applied statistical technique and select the KPIs

This step further involve following sub steps. These steps are: a) describe candidate KPIs, b) describe Null Hypothesis and Alternate Hypothesis for the selected candidate KPIs in first step, c) perform descriptive analytics to recognize their correlation, d) formulation of condense list of KPIs.

Step a. Describe candidate KPIs

This study has used academic dataset of the students from the database which includes different characteristics of students like state, number of semesters, regularity of students, Extra curriculum activities, projects done, research work done, their grades or CGPA, marks in 10^{th}, marks in 12^{th} and all rounder score. These are the candidate KPIs for the preferred domain as shown in table 4.1.

TABLE 4.1 CATEGORY-WISE CANDIDATE KPIs.

Sr. No.	**Candidate KPIs**	**Category**
1.	No. of Backlogs	Quantitative KPI

2.	Extra curriculum activities	Leading KPI
3.	Regularity	Actionable KPI
4.	CGPA	Outcome KPI
5.	State	Quantitative KPI
6.	Projects	Quantitative KPI
7.	Research Work	Qualitative and Quantitative KPI
8.	All Rounder Score	Leading KPI
9.	Number of Semester in the course	Quantitative KPI
10.	Number of Subjects	Quantitative KPI

Step b. Describe Null Hypothesis (H_0) and Alternate Hypothesis (H_1) for the selected candidate KPIs in first step

In this work, the filter statistical approaches were used for feature selection (significant factors), including ANOVA, Correlation Test and from the list of given features. These tests have applied on various combination of candidate KPIs. Some of them have been showcased in the coming sections. Let us take the Null and Alternate Hypothesis for two factors that are:

H_0: First factor have no significant effect on Second factor.

H_1: First factor have significant effect on Second factor.

This test has been performed under 0.05 significant levels.

a. ANOVA Test

One-way Analysis Of Variance(ANOVA) has been performed on these factors, which has determined whether there is any statistically significant differences between the

means of two or more independent (unrelated) groups or not. If the results have been less than the p-value then the null hypothesis has been rejected and alternate hypothesis has been accepted and vice- versa. ANOVA test has been performed on several pair of factors as shown below:

Set 1: Regularity and Extra Curriculum Activities

The subsequent hypothesis would be:

H_0: Regularity has no significant effect on extra curriculum activities.

H_1: Regularity has significant effect on extra curriculum activities.

ANOVA

Regularity

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10.085	2	5.042	19.433	.000
Within Groups	12.195	47	.259		
Total	22.280	49			

Fig 4.1: Results of ANOVA Test for Regularity and Extra Curriculum Activities

Now, according to the Fig 4.1, the result shows that p-value was less than 0.05 then, the null hypothesis has been rejected and the alternate hypothesis has been accepted. It means that Regularity has significant effect on Extra Curriculum Activities.

Set 2: Regularity and CGPA of students.

The corresponding hypothesis for this set them would be

H_0: Regularity has no significant effect on CGPA of students.

H_1: Regularity has significant effect on CGPA of students.

ANOVA

CGPA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	17.189	3	5.730	12.860	.000
Within Groups	20.494	46	.446		
Total	37.683	49			

Fig 4.2: Results of ANOVA Test for Regularity and CGPA

Now, according to the Fig 4.2, the result shows that p-value is less than 0.05 then, the Null Hypothesis has been rejected and the Alternate Hypothesis has been accepted. It means that Regularity have significant effect on CGPA.

Set 3: Number of semester in the course and CGPA of students.

The subsequent hypothesis for this set would be:

H_0: Number of semester has no significant effect on CGPA of students.

H_1: Number of semester has significant effect on CGPA of student.

ANOVA

CGPA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.220	2	.110	.138	.871
Within Groups	37.463	47	.797		
Total	37.683	49			

Fig 4.3: Results of ANOVA Test for semester in the course and CGPA

Now, according to the Fig 4.3, the results shows that p-value is greater than 0.05 then, the Null Hypothesis will be accepted and the Alternate Hypothesis will be rejected. It means that Number of Semester in a Course has no significant effect on CGPA of students.

b. Correlation Test

Two features have been used at a time and also the correlation between them has been explored, as the value of correlation test told about the dependency of features upon each

other. If the correlation test has given the positive value then it means features are directly proportional with each other and if its value is negative then it means features were inversely proportional with each other. In IBM SPSS tool, this technique has been implemented on following set of candidate KPIs.

Set 1: Regularity and Extra Curriculum Activities.

In Fig 4.4, the observed negative correlation between these factors was observed which identifies the inversely proportional relationship among them. It means that when regularity increases the participation of students in Extra Curriculum Activities decreases and when regularity decreases the participation of students in Curriculum Activities increases.

Correlations

		Extra_Curriculum_activities	Regularity
Extra_Curriculum_activities	Pearson Correlation	1	-.550**
	Sig. (2-tailed)		.000
	N	50	50
Regularity	Pearson Correlation	-.550**	1
	Sig. (2-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

Fig 4.4: Results of Correlation Test Regularity and Extra Curriculum Activities

Set 2: Regularity and CGPA.

The positive correlation between these factors has been observed as shown in Fig 4.5, which identifies the directly proportional relationship among them. It means that when regularity increases CGPA also increases and when regularity decreases CGPA also decreases.

Correlations

		Regularity	CGPA
Regularity	Pearson Correlation	1	.639**
	Sig. (2-tailed)		.000
	N	50	50
CGPA	Pearson Correlation	.639**	1
	Sig. (2-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

Fig 4.5: Results of Correlation Test Regularity and CGPA

Set 3: Number of Semester in a course and CGPA.

The observed positive correlation between these factors was observed as shown in Fig 4.6, which identifies the directly proportional relationship among them. It means that when numbers of semester increases CGPA will also increases and when numbers of semester decreases CGPA will decreases. But the Fig has shown that the significant value is greater than 0.05, which has been recognized that there is no significance on number of semesters on the CGPA of the students.

Correlations

		No_of_Sem	CGPA
No_of_Sem	Pearson Correlation	1	.075
	Sig. (2-tailed)		.604
	N	51	50
CGPA	Pearson Correlation	.075	1
	Sig. (2-tailed)	.604	
	N	50	50

Fig 4.6: Results of Correlation Test for No_of_Sem and CGPA

TABLE 4.2 CONDENSED LISTS OF KPIs

Sr. No.	**Condense KPIs**	**Category**
1.	No. of Backlogs	Quantitative KPI
2.	Extra curriculum activities	Leading KPI
3.	Regularity	Actionable KPI
4.	CGPA	Outcome KPI
5.	Projects	Quantitative KPI
6.	Research Work	Qualitative and Quantitative KPI

7.	All Rounder Score	Leading KPI

Table 4.2, described the condensed list of KPIs retrieved after applying step1 on set of candidate KPIs. By applying statistical methods on them, the results have shown that only some KPIs were significant for meeting the objective. So the descriptive techniques was applied only on the condense list of the KPIs which has been increased the performance of the progression system and make the decision making process easier.

Further in this work, the following steps were followed for Analyzing Student Progression System by using these condense list of KPIs. The steps are given below:

4.2 Generating Cube by Using OLAP technology.

This step further involve following sub steps.

The steps are: a) create Database and connect the source of the data, b) Design a data source Query, c) Design the Cube Structure and Process the Cube, e) Visualization of the cube.

Step a. Create Database and connect the source of the data

This study utilizes Academic Dataset of the students from the Academic Database of the organization which includes different key factors of students. The data have been stored in database (MS – Access) as shown in Fig 4.7 and the dataset which was stored in the database has been connected for its further processing.

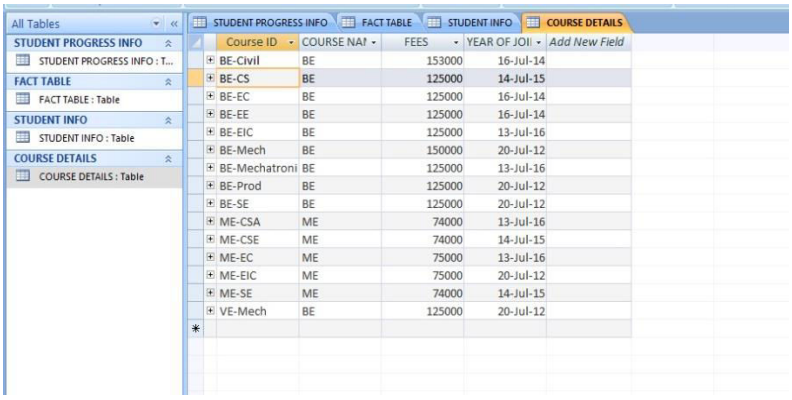


Fig 4.7: Academic dataset stored in MS – Access

Step b. Design a data source Query

All the tables were loaded from the database for query designing. Then we have to design a Multidimensional Expressions (MDX) query from these columns and choose them for the cube designing as a dimension or a measure. The query was automatically generated by the tool, as shown in the Fig 4.8. After designing the query, the designing of the cube structure has been performed.

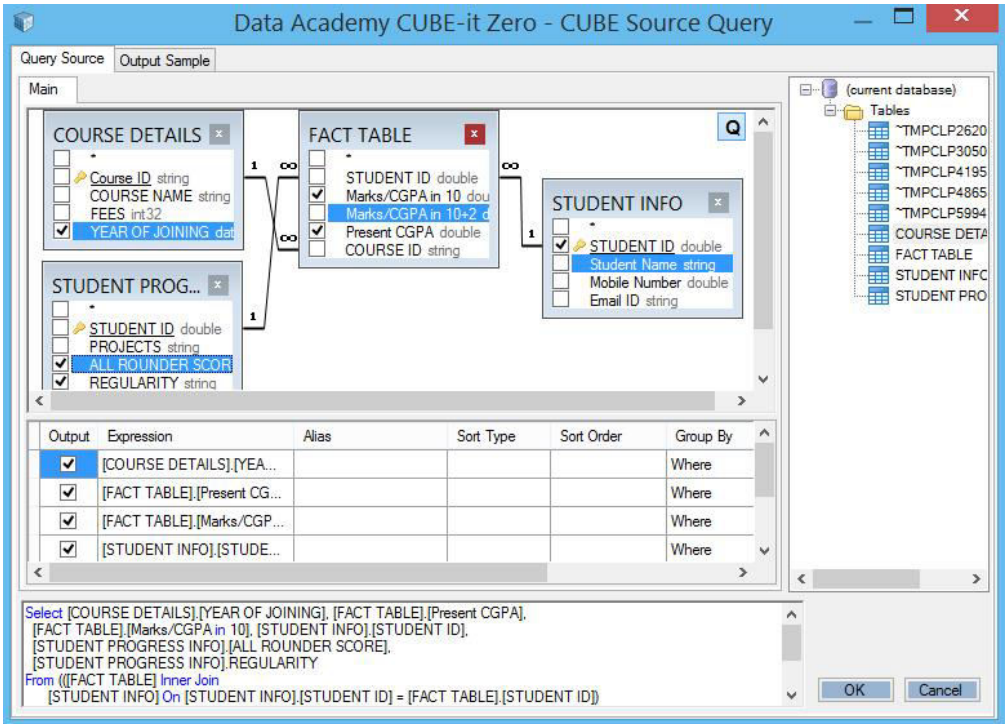


Fig 4.8: Cube query design

Below the MDX query has been showcased for the dimensions selected in the Fig 4.8.

```
Select [FACT TABLE].[Marks/CGPA in 10], [FACT TABLE].[Present CGPA],
  [STUDENT INFO].[STUDENT ID], [COURSE DETAILS].[YEAR OF JOINING],
  [STUDENT PROGRESS INFO].[ALL ROUNDER SCORE],
  [STUDENT PROGRESS INFO].REGULARITY
From (([COURSE DETAILS] Inner Join
     [FACT TABLE] On [COURSE DETAILS].[Course ID] = [FACT
TABLE].[COURSE ID])
    Inner Join
    [STUDENT INFO] On [STUDENT INFO].[STUDENT ID] = [FACT
TABLE].[STUDENT ID])
  Inner Join
  [STUDENT PROGRESS INFO] On [STUDENT PROGRESS INFO].[STUDENT ID]
= [FACT TABLE].[STUDENT ID]
```

Step c. Design the Cube Structure Process the Cube

As shown in the Fig 4.9, the dimension and measure were selected for the cube and also selected the type of aggregation (i.e. sum, count, maximum, minimum or average) for the cube designing. After selection of dimension and measure the cube structures have generated and once the cube has generated, processing of cube has been performed.

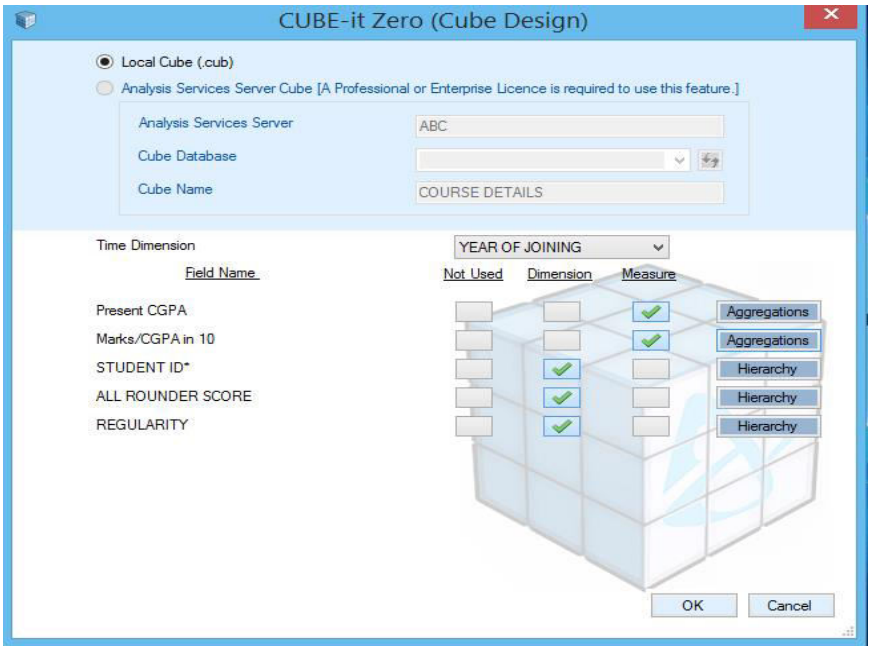


Fig 4.9: Cube structure design

The cube has been generated and after cube designing, processing of cube has been done as shown in Fig 4.10. Once the cube was processed, a .cub file has been created as shown in Fig 4.11, which has been used for the visualization of cube.

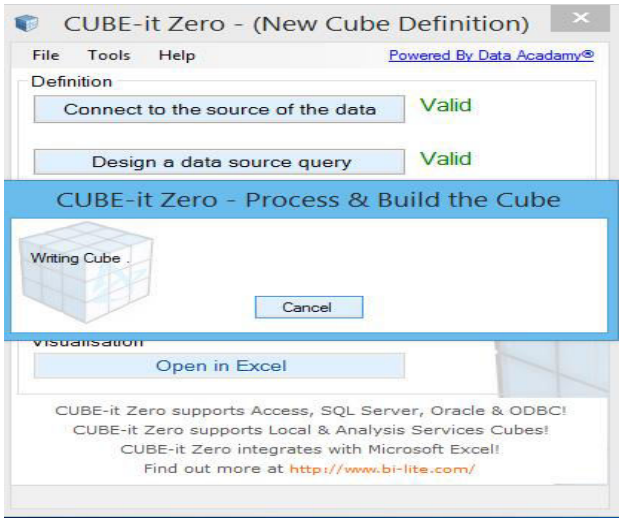


Fig 4.10: Processing of cube

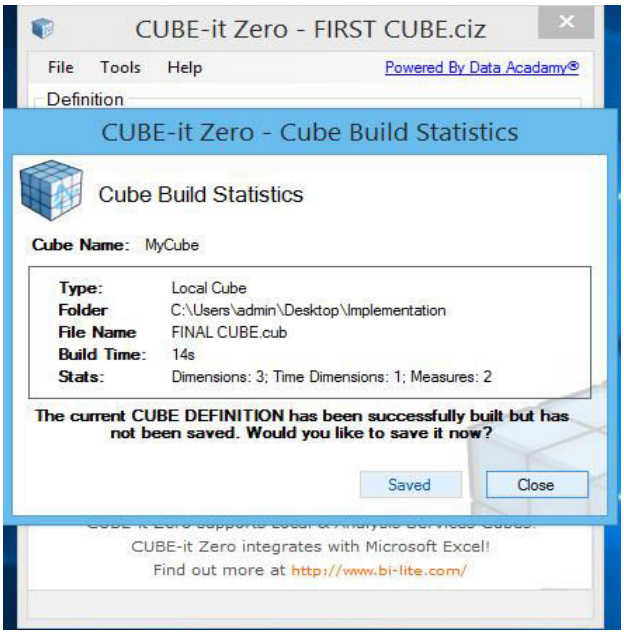
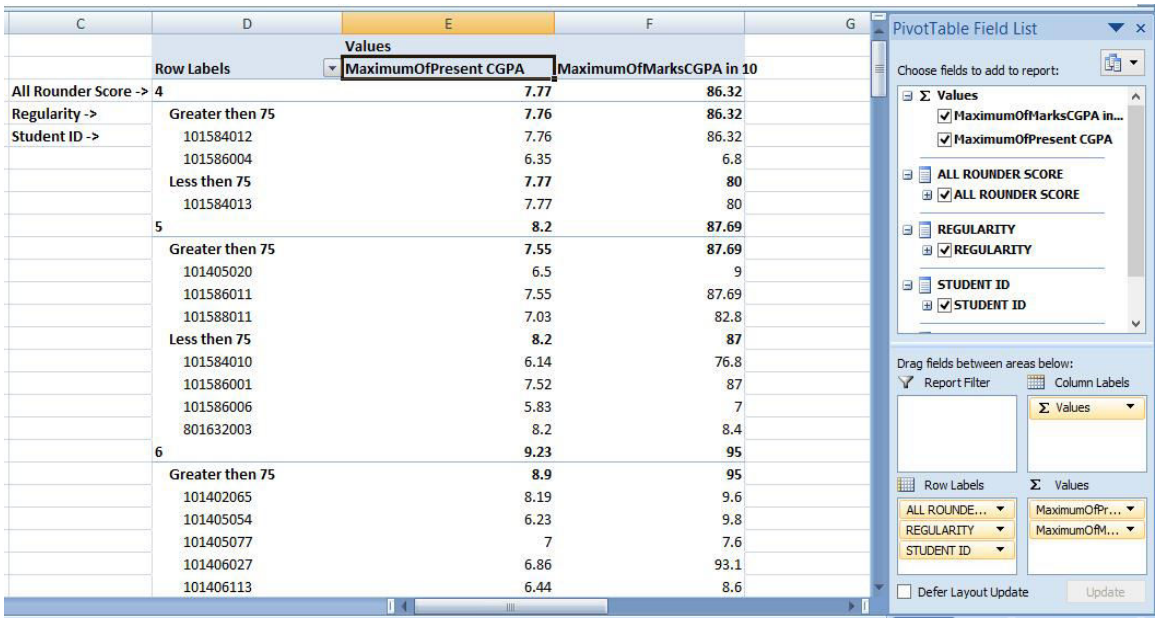


Fig 4.11: Cube file created

4.3 Descriptive Analysis of the cube.

The cube has been generated and ready for visualization. For its visualization the .cub file which was saved and has been opened in the MS Excel, where the cube has been shown in the form of Pivot table as given in Fig 4.12. This pivot table shows multiple dimensions and corresponding measure against them.



	Row Labels	Values: MaximumOfPresent CGPA	MaximumOfMarksCGPA in 10
All Rounder Score ->	4	7.77	86.32
Regularity ->	Greater then 75	7.76	86.32
Student ID ->	101584012	7.76	86.32
	101586004	6.35	6.8
	Less then 75	7.77	80
	101584013	7.77	80
	5	8.2	87.69
	Greater then 75	7.55	87.69
	101405020	6.5	9
	101586011	7.55	87.69
	101588011	7.03	82.8
	Less then 75	8.2	87
	101584010	6.14	76.8
	101586001	7.52	87
	101586006	5.83	7
	801632003	8.2	8.4
	6	9.23	95
	Greater then 75	8.9	95
	101402065	8.19	9.6
	101405054	6.23	9.8
	101405077	7	7.6
	101406027	6.86	93.1
	101406113	6.44	8.6

Fig 4.12: Cube Visualization

4.4 Cluster Analysis.

In this segment some clusters have been created with the help of academic dataset and have used these data clusters for predicting performance of the students.

This step further involve following sub steps.

The steps are: a) data cleaning; two models have been prepared b) first model Replace Missing Values, c) second model performs the clustering on the database.

Step a. Data cleaning

The two models have been prepared, one that finds and replace the missing values with the zero values and second model performed the clustering on the database. Prior to data cleaning, conversion is performed on the database so that accurate clusters have been created.

Firstly, the string values are converted into numeric. Then, the string was replaced with numerical values that are 0 and 1 as shown in Fig 4.13. This has been done to make the data in a form so that it has been processed.

Projects → 1 for Y, 0 for N

Research work →1 for Y, 0 for N

Where Y: Yes, N: No

Regularity→ Less than 75 = 0, Greater than 75=1, Equals to 75 = 1

	A	B	C	D	E	F	G	H
1	REGULARITY	PROJECTS	REASEARCH WORK	Present CGPA	ALL ROUNDER SCORE	Marks/CGPA in 10	Marks/CGPA in 10+2	Performance
2	0	1	1	6.4	6	83.7	87.0	Average
3	1	0	1	6.9	6	91.0	75.8	Average
4	1	0	0	6.5	4	83.4	86.4	Average
5	1	0	0	6.5	4	9.8	75.0	Average
6	1	1	0	7.1	7	79.3	79.6	Average
7	1	1	0	6.9	6	9.6	81.6	Average
8	1	1	0	6.4	6	7.2	81.8	Average
9	0	0	0	7.4	3	93.4	90.2	Average
10	1	1	0	6.2	6	8.8	77.2	Average
11	1	0	1	6.7	6	9.8	93.1	Average
12	1	0	1	7.2	7	9.2	84.0	Average
13	1	1	1	7.1	9	76.2	75.8	Average
14	0	1	1	7.9	7	9.4	88.4	Average
15	1	1	0	7.6	7	95.0	90.6	Average
16	1	1	0	6.7	6	10.0	80.8	Average
17	1	0	0	8.2	6	9.6	82.8	Above Average
18	1	1	0	9.2	8	95.0	94.4	Above Average
19	0	1	0	8.2	6	10.0	87.0	Above Average
20	0	1	0	6.6	4	9.6	84.6	Average
21	1	1	0	7.0	7	9.4	93.7	Average
22	1	1	0	7.4	7	9.6	80.0	Average
23	1	1	0	7.5	7	89.3	74.0	Average

RapidMiner Data

Fig 4.13: Cleaned Data

Following that a new column “Performance” was prepared based on the ranges of the CGPA. For this, minimum and maximum values of the CGPA was found then based on this the new attribute “label” is assigned with values. On the basis of performance, the CGPA grades were classified into three groups namely, Below Average, Average and Above Average. These groups include a range of CGPA: Below Average - CGPA below 6, Average - CGPA between 6 and 8, Above Average - CGPA between 8 and 10. Their ranges and other details are given below: Minimum CGPA- 5.63 and Maximum CGPA- 9.48

Now, this new column has been created the base for the clusters. This task has been performed by using this column as the label attribute for the clustering.

Step b. First model Replace Missing Values

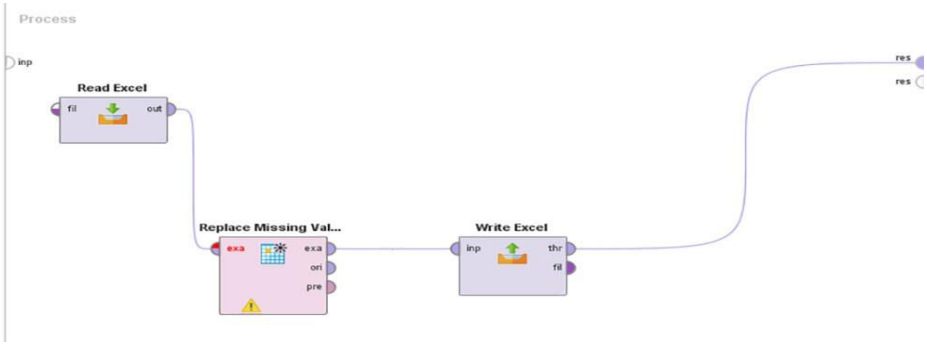


Fig 4.14: Replace Missing Values Model

In Fig 4.14, the model was used to replace the missing values in the dataset if there were any. The “Read Excel” operator was used to read the excel file. After that, the “Replace missing Values” operator was used to substitute the missing values. The last operator writes the resultants excel file into the data folder. After preparing the dataset the clustering model has been applied on it.

Step c. Second model performs the Clustering on the database

K-means clustering was used for this step. Every clustering algorithm aims mainly two things. Primarily, it separates the items in the dataset into different groups in a way that data items in one group are more related to the data items in the same group. Secondarily,

it arranges the data items into groups so that items in one group are different from another group's items. The K-means operator has been used to represents an accomplishment of k-Means [11]. This operator has created a cluster attribute if not present yet.

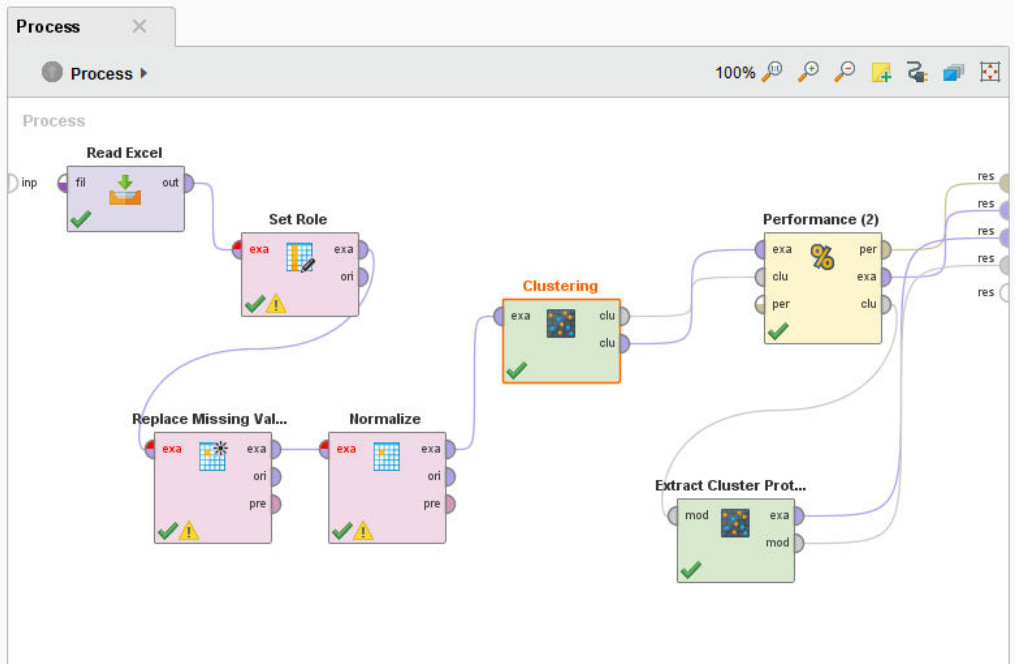


Fig 4.15: Clustering Distance Model

To perform clustering some operators have been used as shown in Fig 4.15. Each of them has performed some kind of task, is defined below:

a. Read Excel: This operator has been used to examine the excel file.
b. Set role: To change the role of the attributes this operator has been used. The new attribute was used as targets attribute and set its role to batch. Batch is a special role. An attribute with the batch role indicates the membership to a specific batch.
c. Replace Missing Values: This operator has been used to substitute the missing values.
d. Normalize: This operator normalizes the values of the selected Attributes. Normalization has done to change the representation, not the values. It has done in a specific range to scale the values.

e. Clustering**:** This is the main operator which performs the very important task, which was the K-means clustering. This clustering algorithm follows the following steps:

i. Specify the value of k

ii. Randomly assign the data points to clusters

iii. Compute centroid for the clusters by using Euclidean Distance.

Euclidean Distance (D) for computing centroid is given below:

$$\text{Dist}_{(x,\,y)} = \sqrt{(a_2 - a_1)^2 + (b_2 - b_1)^2} \qquad (1)$$

iv. Re-locate each point closest to cluster centroid

v. Re-compute cluster centroids

vi. Repeat steps 4 and 5 until no enhancements were possible.

f. Data to Similarity: This operator measures the similarity of each given Exampleset with every other same Exampleset.

g. Performance: This operator has been used for evaluating the performance of the centroid-based clustering methods. This operator describes a list of performance criteria values based on cluster densities as shown in Fig 21.

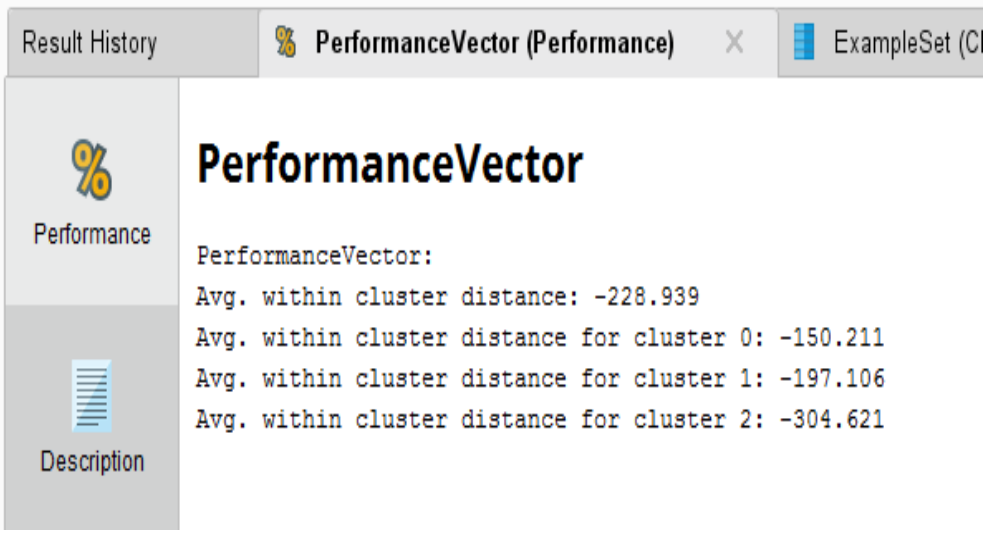


Fig 4.16: Performance of clusters

The average distance between the clusters found was -228.939. As shown in Fig 4.16, the smaller the distances are, the better, the clustering works. Also, all the distances are in negative as the distances are multiplied by -1 to make them useful for optimization. The Cluster Distance Performance measures the average distance between the centroids, therefore the distance is multiplied by -1. If the distance would not be multiplied by -1, then the optimization operator would select the higher average distance between the centroids. Some graphs have been plotted based on the clustering performance which makes easy to visualize the performance of students.

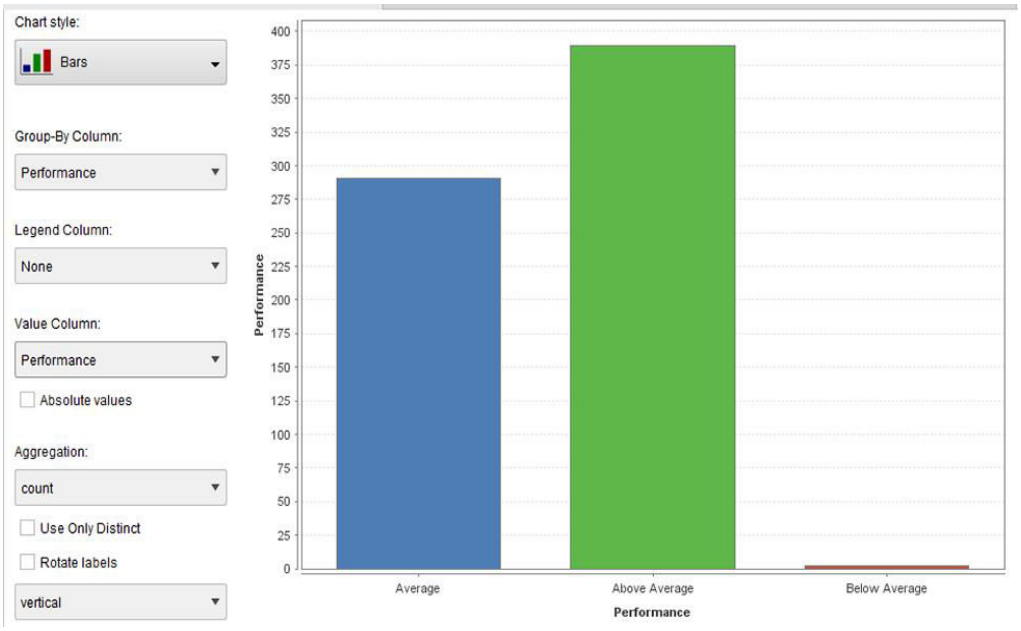


Fig 4.17: Graph shows Frequency of the Performance

The Fig 4.17 shows the frequency of the Performance attribute which shows the frequency of students who scored above average CGPA. There are only a few students that received below average but the performance of most of the students was above average.

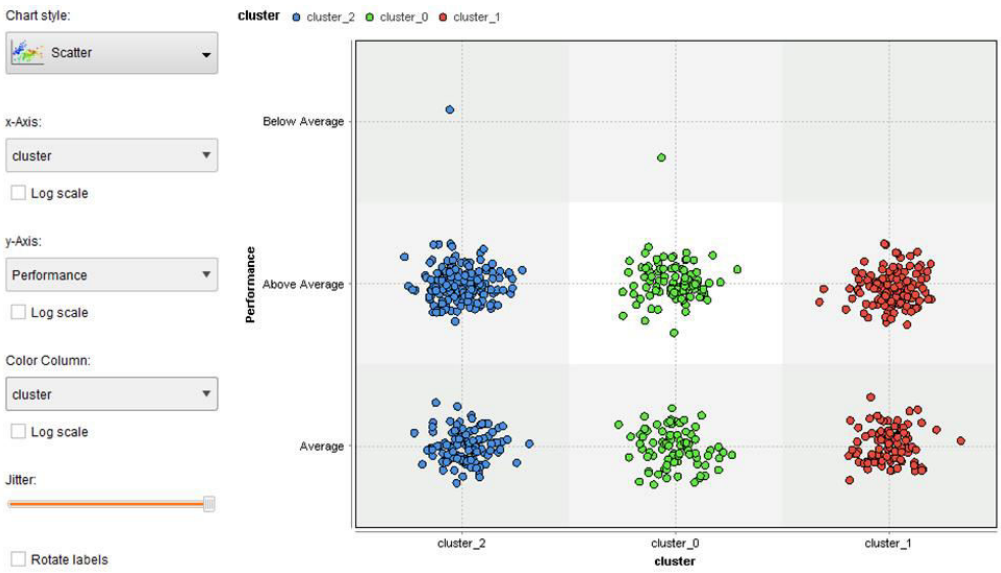


Fig 4.18: Scatter graph between Cluster and Performance

The Fig 4.18 shows the clustering based on "Performance" attribute. It has been stated as a label attribute in Rapid Miner. Fig 4.18 also illustrates that, the cluster 2 have maximum number of students which have above average performance and very rare students which have performance below average. Also, the graph demonstrates that cluster 1 only holds students which have average and above average performance. Cluster 1 does not have any student whose performance was below average.

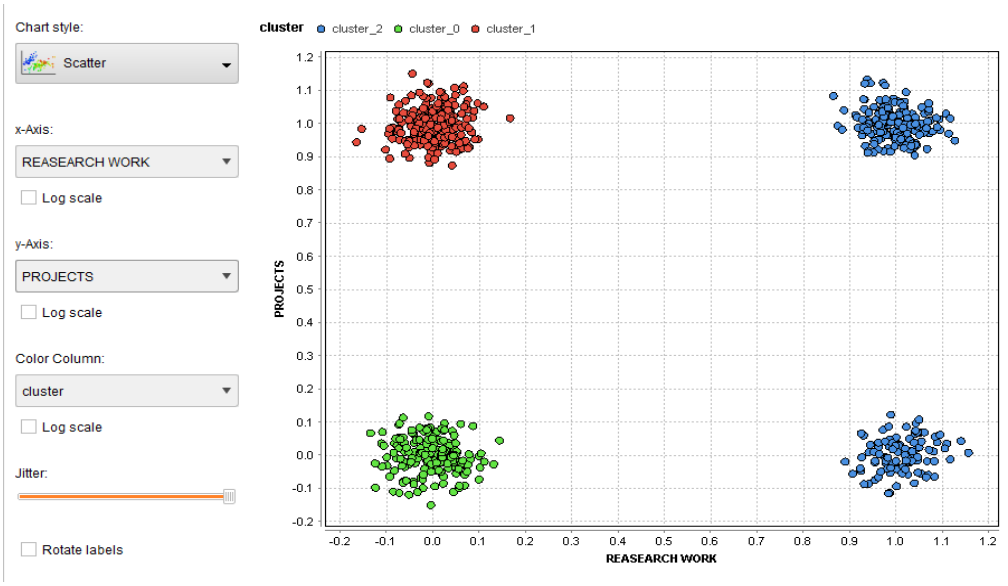


Fig 4.19: Graph between Projects vs. Research work

The Fig 4.19 represents the clusters based on projects and research work. It can be observed that if the students did not worked on a research work they were put in cluster 0 as they were worked on none of the projects as well. And a student who has worked on research work and not on projects are put into cluster 2 or also the student who has worked on projects or not the research work were placed in cluster 2. The cluster 1 contains the students who worked on both projects and research project as well.

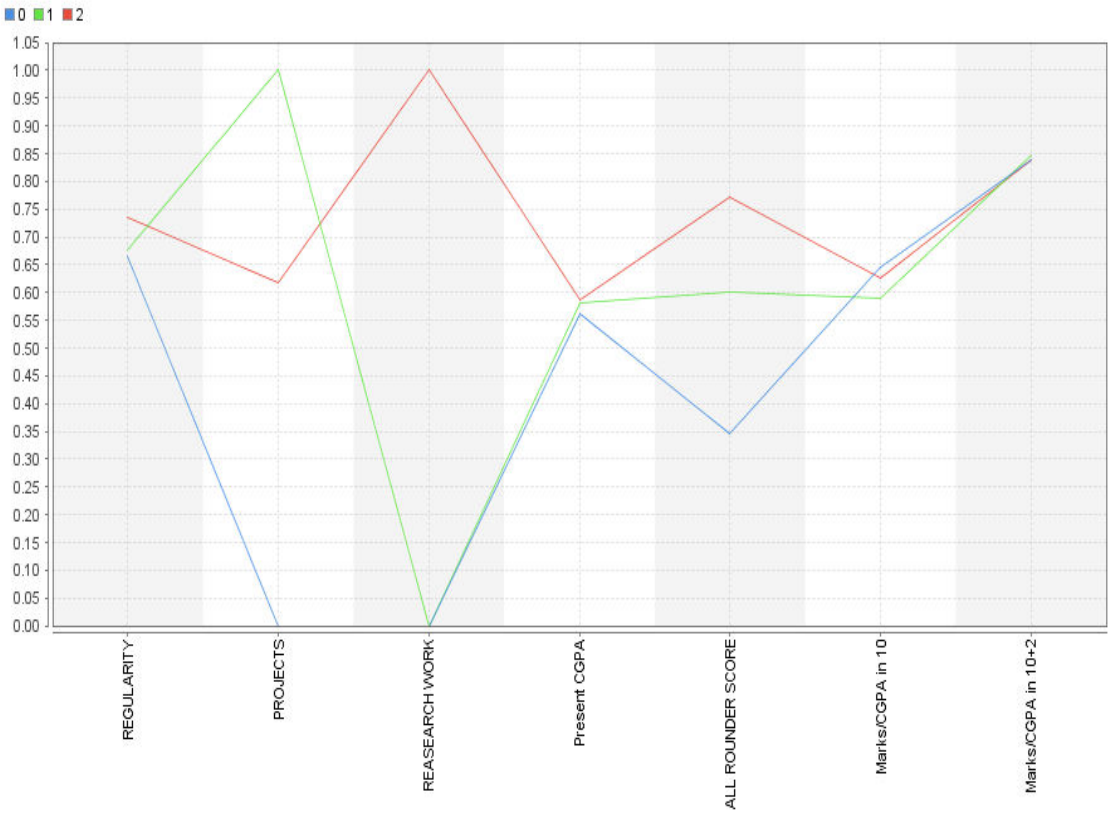


Fig 4.20: Plot Graph for all Clusters

In Fig 4.20, graph shows that the students belong to cluster 0 had the awful results in projects, research work, percent CGPA, all-rounder score, and also in regularity. They had most appreciated results marks in 12^{th} and marks in 10^{th}. The students belonging to cluster 1 had best results in projects and marks in 12^{th}. Whereas cluster 0 had unpleasant results in projects, regularity and research work. The students belonging to cluster 2 had most appreciated results in regularity, research work, present CGPA and all rounder score. But were average results in projects, marks in 12^{th} and marks in 10^{th}.

The average performance was observed in projects and 10th CGPA. Coming on the results needs to give recommendations for the overall improvement of the study process, the "Histogram Color" has been used to observe how the students belonging to the different cluster were distributed by overall performance.

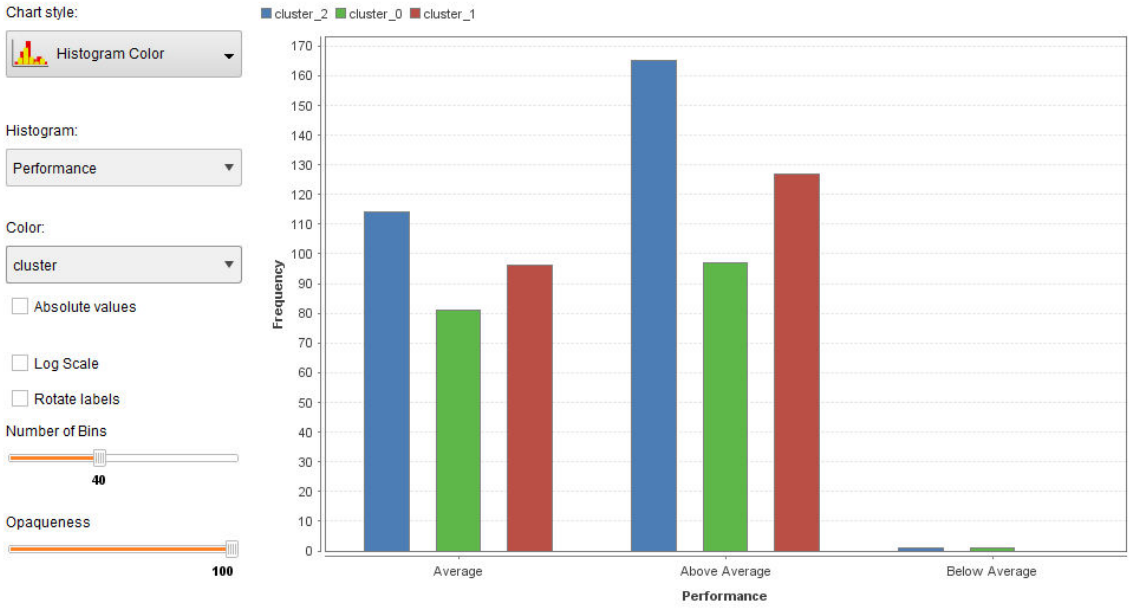


Fig 4.21: Frequency of each Label shown in histogram

It can be seen from Fig 4.21 that most of the students that belong to "average" class are in cluster 2. Moreover, the students from cluster 1 very rarely belong to the "Below average" class but they belong very well in average and above average class. This represents that the performance of students in cluster 1 would be increased by giving focus on the 10th CGPA and Present CGPA, as seen from the above Fig 25 the students belongs to cluster 1 were have average results in their present CGPA and marks in 10th. So, from observation it was recommended that these grades should be pointed out as an important at the starting of the studies.

4.5 Predictive Analytics

The result of cluster analysis helped to calculate the performance of the students. Primarily, the cube designing has been done and then designed three clusters by using k-means algorithm. These clusters placed the similar attribute in one class, on which

various models have been applied for making prediction. These models are Naïve Bayes, LDA, Decision Tree and Deep Learning. For applying these models a new attribute "Performance" was added which was calculated based on the values in other fields. It includes three values low, average, and high that shows the student performance. The three values are based on the following conditions:

TABLE 4.3: ATTRIBUTE PERFORMANCE VALUES

Less Value	**Average Value**	**High Value**
1. Regularity is 0 and Project work is 0 2. Present CGPA is less than 6 3. All-rounder score is less than 5	1. Regularity is 0 and Project work is 1 2. Present CGPA is between 6 and 8 3. All-rounder score is between 5 and 7.	If criteria of Less and average is not satisfied that means Regularity is 1, CGPA is greater than 8, All-rounder score is greater than 7, project work is 1, research work is 1, then High Participation can be recorded from student's side in discussion board.

The dataset contains a target value which is predicted and it is of two types: Classification and Regression. Classification type dataset have discrete set of value like yes or no, whereas regression type dataset contains continuous value in their target attribute. As the target attribute of the academic dataset was of classification type and the models includes in it were: Naïve Bayes, LDA, Decision Tree, Support Vector Machine (SVM) and Deep Learning which had applied on academic dataset. But these four classification models such as Naïve Bayes, LDA, Decision Tree and Deep Learning have applied on data as they were suitable for academic dataset but SVM was not performed on academic data because this model could not be performed on polynomial type of dataset.

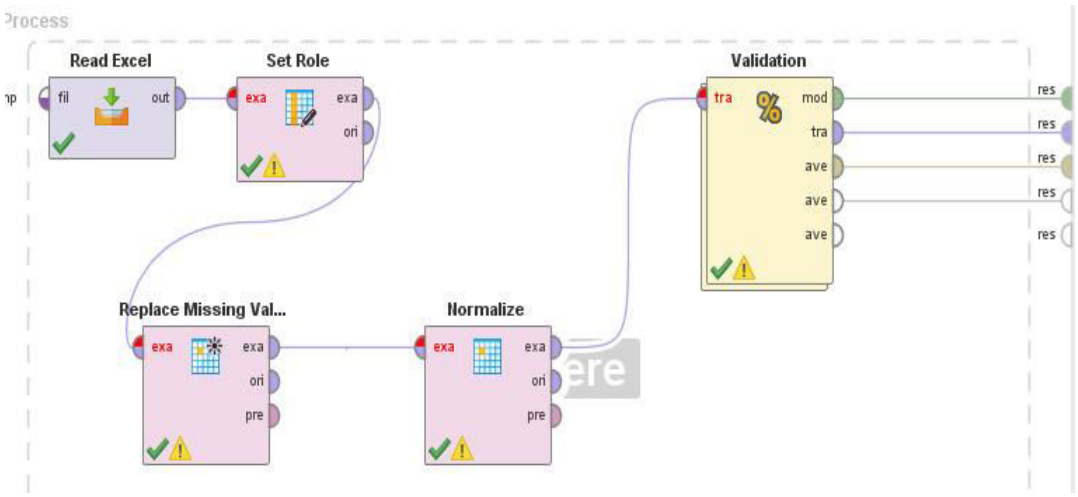


Fig 4.22: Predictive Analytics task sequence

As Fig 4.22 demonstrates, the Predictive Analytics task sequence for applying predictive models it includes a split validation operator which is only a nested operator. It has two sub processes: a training sub process and a testing sub process. In training and testing part the data has been split into 70 and 30 ratio. That means training has been done on 70% of dataset and testing has conducted on 30% of data. The model has been learned from the training set and then has applied on the testing set. Various models have been applied on academic datasets as shown below:

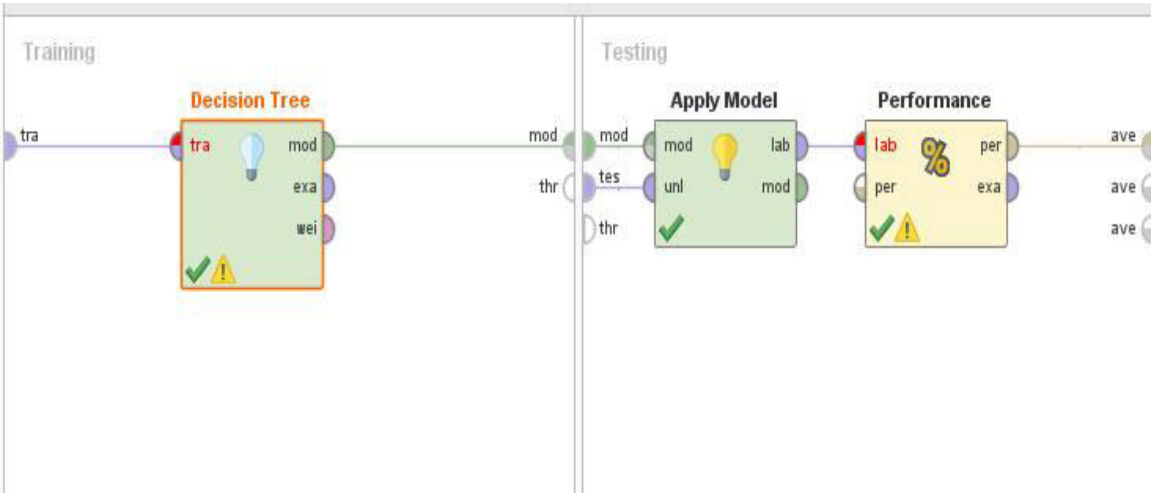


Fig 4.23: Decision Tree model

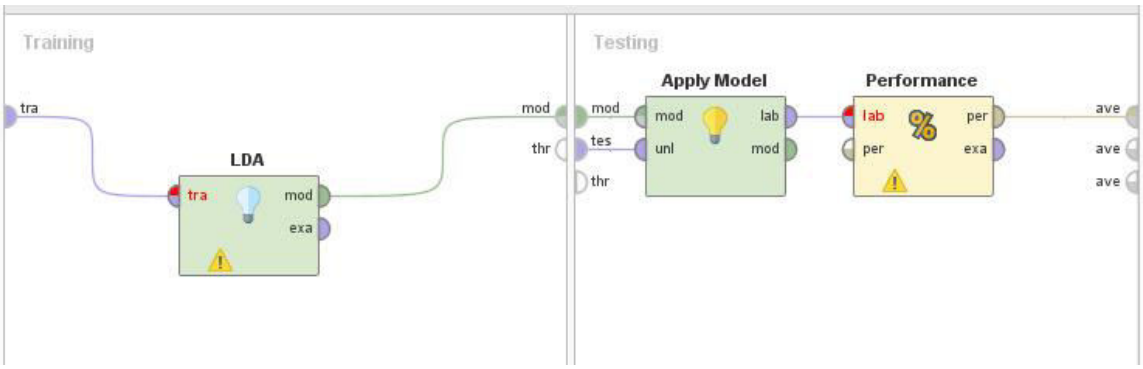


Fig 4.24: LDA model

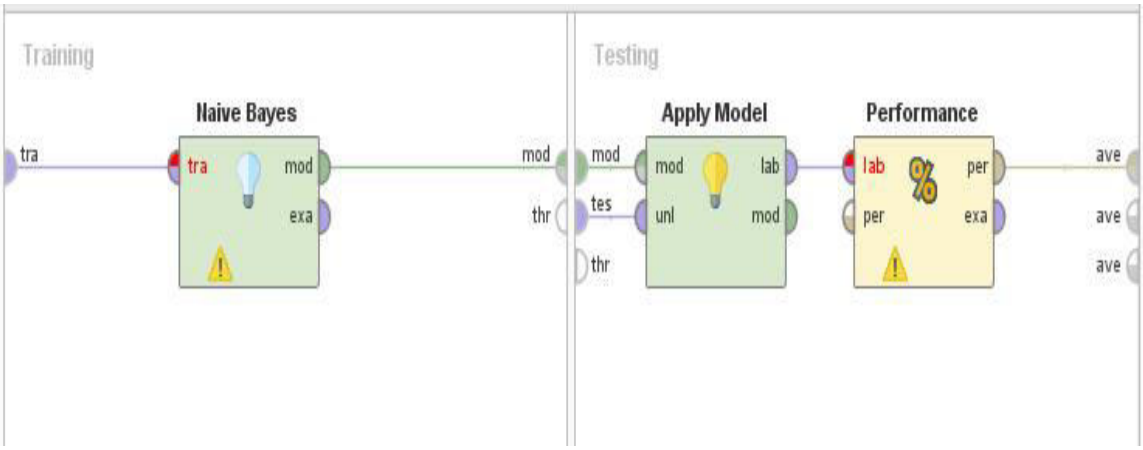


Fig 4.25: Naïve Bayes model

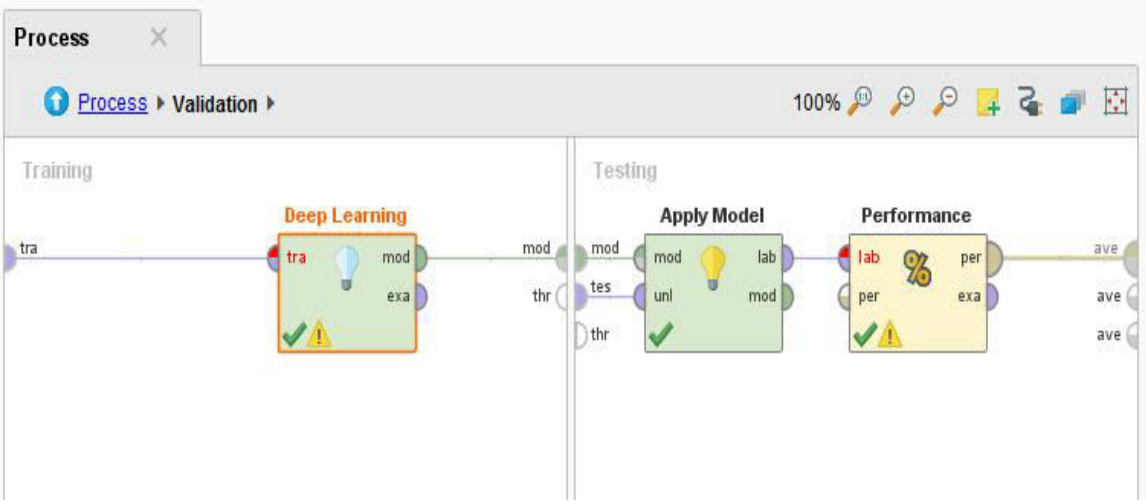


Fig 4.26: Deep Learning model

A. Decision Tree

Decision tree generates the classification or regression models in the form of a tree structure in which the rectangular boxes are known as the nodes. Root node is the topmost node of the tree and internal nodes have a child. If the internal nodes don't have any child then it is known as leaf node or terminal node. There are five commonly used algorithms for decision tree includes: - Iterative Dichotomiser 3 (ID3), Classification and Regression Tree (CART), Chi-squared Automatic Interaction Detection (CHAID), C4.5 algorithm and J48 [36]. In this work, the C4.5 algorithm has been used for performing decision tree model. As it is the successor of the ID3 algorithm and can be used for the classification. It was used over other algorithms because it could handle both continuous and discrete data points, it could handle the data with missing values and also one can go back to when the tree was created and remove the nodes which were not helpful for coming on decisions.

ID3 algorithm was developed by Ross Quinlan in 1983 who was a data mining computer science researcher [36]. ID3 algorithm has performed attribute selection by using Entropy and Information Gain idea.

Entropy is degree of randomness of elements or in other words it is measurement of uncertainty**.**

$$\text{Entropy}(A) = \sum_{n=1} -p(I)\log_2 p(I) \qquad (2) \qquad [36]$$

Where: A is set, Entropy (A) is information entropy of A

p(I)= proportion of A belonging to class I

Information gain has been used to select a particular attribute to become a decision node in ID3 algorithm.

$$\text{Gain}(G, A) = \text{Entropy}(G) - \sum ((|G_v|/|G|) \times \text{Entropy}(G_v)) \qquad (3) \qquad [36]$$

Where:

G is records collection.

A: - attribute

Gain (G, A) is the gain of G after split on A

v is all the possible values of the attribute A.

G_v is the number of elements for each v.

$\sum$ is the summation of (($|G_v|/|G|$) x Entropy (G_v)) for all the items from the set of v

B. Naïve Bayes

The Naive Bayesian classifier is based on Bayes' theorem which is given as; "Bayes' theorem elaborates the probability of occurrence of an event, based on earlier knowledge of conditions that might be related to the event". Naive Bayesian model is easy to build as there is no requirement of estimated iterative parameters which has been made useful for very large datasets [14].

Bayes theorem provides a way of calculating the posterior probability, $P(c|a)$, from $P(a)$, $P(y)$, and $P(y|a)$ [23]. Naive Bayes classifier assumes that the effect of the value of a predictor (y) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(A|Y) = P(Y|A)\,P(A)/P(Y) \quad (4) \quad [25]$$

$$P(A|Y) = P(Y1|A) \times P(Y2|A) \times \ldots\ldots\ldots\ldots \times P(Yn|A) \times P(A)$$

Where P (A|Y) is Posterior probability and P (Y|A) is Likelihood and P(A) class prior probability and P(Y) predictor prior probability [25].

C. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a classification method that was developed by R. A. Fisher [28] in 1936. The dataset sometimes includes a number of features, so it is

difficult to work on it or visualize that dataset. This is the place where Dimension Reduction algorithm plays a crucial role. There are various methods used for Dimension Reduction that includes: Principal Component Analysis (PCA), LDA and Generalized Discriminant Analysis (GDA).

The goal of LDA (one of the dimension reduction technique) was to reduce the dimensional space. This has been achieved by performing following three steps: i. calculate the reparability between the different classes which is known as between-class matrix ii. Calculate the within-class matrix (the distance between mean and samples of each class) iii. To fabricate the lower dimensional space which decreases the within class variance and increases the between-class variance.

D. Deep Learning

Deep learning is based on a multi-layer artificial neural network technique that takes few ideas of Artificial Intelligence. For instance: technology behind driverless cars. It helps in reducing the decision-making capabilities of humans. To feed the data into model deep leaning has been used. After the model training, in testing the data which was fed into the system has been used to make decision about other data. The system training has done through neural networks. Deep learning is important because it emphasis on developing these neural networks and they are known as Deep Neural Networks. Deep Neural Networks consists of three layers that are: input layer, hidden layer and output layer. Fig 27(d) shows the deep learning model has been used for making predictions.

4.6 Results and Inferences

Deep Learning model has given the accuracy that is 99.02% as shown in Fig 4.27, which was highest from all the models includes Naïve Bayes, Decision Tree, LDA and Deep Learning applied on academic dataset. A confusion Matrix was generated based upon trained data and the tested data. As confusion matrix contains the actual class and the predicted class. As shown in Fig 4.28, the actual value is 127 for average and predicted

value is also 127 for average class but the model also predicts 1 value for high class but it doesn't lie in high class. In this way the values can be analysed from confusion matrix.

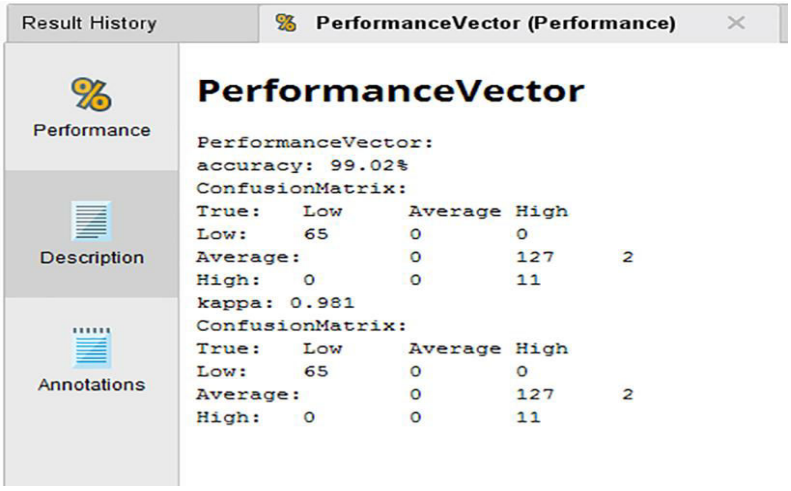


Fig 4.27: Accuracy of Deep Learning Model

accuracy: 99.51%

	true Low	true Average	true High	class precision
pred. Low	65	0	0	100.00%
pred. Average	0	127	1	99.22%
pred. High	0	0	12	100.00%
class recall	100.00%	100.00%	92.31%	

Fig 4.28: Confusion Matrix

ExampleSet (205 examples, 5 special attributes, 7 regular attributes) Filter (205 / 205 examples): all

Row No.	Performance	prediction(Performance)	confidence(Low)	confidence(Average)	confidence(High)	REGULARITY	PROJECTS
1	Low	Low	0.692	0.308	0.000	1	0
2	Average	Average	0.009	0.991	0.000	1	1
3	Low	Low	1.000	0.000	0.000	0	0
4	Average	Average	0.009	0.960	0.032	1	0
5	Average	Average	0.000	1.000	0.000	1	1
6	Average	Average	0.006	0.993	0.001	1	0
7	Average	Average	0.000	1.000	0.000	1	1
8	Average	Average	0.001	0.999	0.000	1	1
9	Average	Average	0.000	1.000	0.000	1	1
10	Low	Low	0.627	0.373	0.000	1	0
11	Low	Low	1.000	0.000	0.000	0	0
12	Low	Low	1.000	0.000	0.000	0	0
13	Low	Low	1.000	0.000	0.000	0	1
14	Average	Average	0.000	0.999	0.001	1	1
15	Average	Average	0.000	1.000	0.000	1	1

Fig 4.29: Prediction results of student performance

The students performance has been predicted as shown in Fig 4.29, by using training dataset. The model has used 70% of the data for training and 30% for testing.

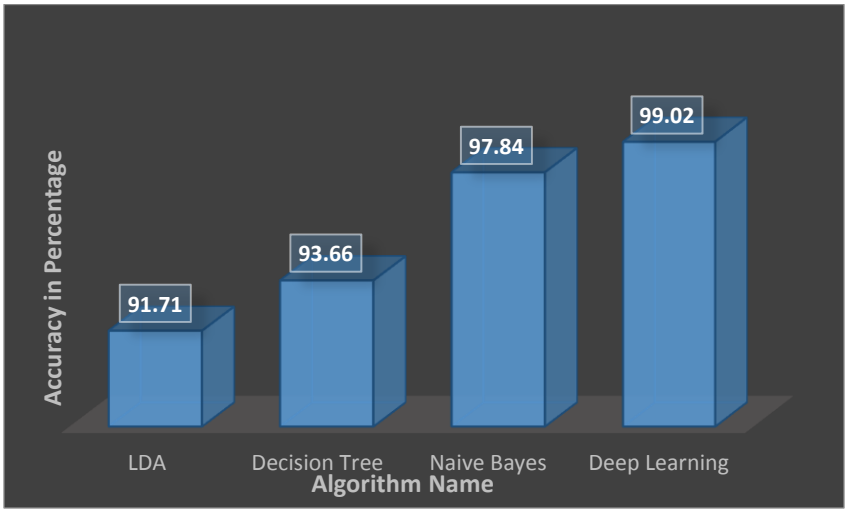


Fig 4.30: Accuracy graph of Models

Fig 4.30 depicts the accuracy graph of all the models used to predict the performance of the students. Fig 4.30 illustrates that Deep Learning model has given the maximum accuracy whereas LDA has given the least accuracy results among all of them. Accurate result means good predictive power of that model. That's why the Deep Learning model has been used for predicting the performance of the students. This model predicts the performance with highest accuracy.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In the proposed work, primarily the Statistical techniques have been implemented such as Hypothesis Testing and Correlation testing to generate the optimized and most significant set of KPIs for the Student Progression System. Secondarily, the descriptive and predictive analytics have been implemented to generate cube structure, visualization and predicted the performance for student progression system. The condensed list of KPI's has been used for this implementation. The OLAP technique has been applied on the academic dataset to obtain some cubes for making analysis better. The cube generated by using cube technology helps to visualize the multiple dimensions of the students like course name, all rounder score and regularity of students and it gives an aggregation for all these dimensions. Also, the predictive analytics has been performed on academic dataset, in which k-means clustering technique has been applied. In this work, the three clusters have been designed by using k-means clustering algorithm. The result of cluster analysis helped to find performance which includes a range of CGPA: Below Average - CGPA below 6, Average - CGPA between 6 and 8, Above Average - CGPA between 8 and 10. These clusters have placed the similar attributes in the same class, on which various models have been applied for making prediction. These models include Naïve Bayes, LDA, Decision Tree and Deep Learning. For applying these models a new attribute "Performance" is created which differentiated the student records as low, average, and high performances. It has been found that the accuracy of the Deep Learning model was 99.02%, which was highest among all the models.

REFERNCES

[1] P. R. M. d. Andrade and S. Sadaoui, "Improving business decision making based on KPI management system," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017.

[2] M. Azabou, "Analyzing Textual Documents with New OLAP Operators," in *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, 2017.

[3] G. R. Bawane, "Integration of OLAP and Association rule mining," in *IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15*, 2015.

[4] W. Chen, "An Optimized Distributed OLAP System for Big Data Wenhao," in *2017 2nd IEEE International Conference on Computational Intelligence and Applications, ICCIA 2017*, 2017.

[5] M. Corcoran, The Four types of Data Analytics, 2012.

[6] A. Cuzzocrea, "Querying Encrypted OLAP Data Alfredo," in *2017 IEEE 41st Annual Computer Software and Applications Conference* , 2017.

[7] K. Dhanasree, "A Survey on OLAP," in *2016 IEEE International Conference on Computational Intelligence and Computing Research*, 2016.

[8] F. Dehne, VOLAP: A Scalable Distributed Real-Time OLAP System for High-Velocity Data, VOLAP: A Scalable Distributed Real-Time OLAP System for High-Velocity Data, 2017.

[9] L. Fu, "A Recommendation System Using OLAP Approach," in *Proceedings - 2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016*, 2017.

[10] M. Fisun, "Implementation of the Information System of the Association Rules Generation from OLAP-cubes in the Post-relational DBMS Caché," in *Computer Sciences and Information Technologies - Proceedings of the 11th International*

Scientific and Technical Conference, CSIT 2016, 2016.

[11] C. Elkan, "Using the Triangle Inequality to Accelerate k-Means," in *Proceedings of the Twentieth International Conference on Machine Learning*, Washington DC,, 2003.

[12] R. Djiroun, "A Data Cube Design and Construction Methodology Based on OLAP Queries," in *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, 2016.

[13] X. Du, "Multidimensional analysis of seawater quality data based on graph OLAP Xiangjun," in *2015 8th International Symposium on Computational Intelligence and Design*, 2015.

[14] D. J. Hand, "Idiot's Bayes---Not So Stupid After All?," 2001.

[15] J. Han and M. Kamber, Data Mining:Concepts and Techniques, Morgan Kaufmann Publishers, 2012.

[16] G. Hagan, "HR- The Appraisal Process".

[17] K. Gutiérrez-Batista, "Building a contextual dimension for OLAP using textual data from social networks," *Expert Systems with Applications,* 2018.

[18] Granat, Djorgovski, A. R. Brunner, R. Mahabal and R. Williams, Statistical Challenges in Astronomy, springer.

[19] R. kimball, The Data Warehouse Toolkit: The Defi nitive Guide to Dimensional Modeling, Third Edition, John Wiley & Sons, Inc., Indianapolis, Indiana, 2013.

[20] S. Kaushik, "Analytics Vidhya Learn everthing about analytics," 03 Nov 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/.

[21] S. KAUSHIK, "Introduction to Feature Selection methods with an example (or how to select the right variables?)," 2016.

[22] S. M. Joshi, "Developing Key Performance Indicators framework for evaluating performance of engineering faculty," 2016.

[23] P. Jain, "A Case Study on Car Evaluation and Prediction: Comparative Analysis using Data Mining Models," *International Journal of Computer Applications*

(0975, vol. Volume 172 – No.9, 2017.

[24] S. Mirabedini, "The Research on OLAP for Educational Data Analysis," *International Research Journal of Applied and Basic Sciences,* 2014.

[25] J. Park, "New key performance indices for complex manufacturing scheduling," 2015.

[26] A. R. Othman, "A Model Classification Technique for Linear Discriminant Analysis for Two Groups," *IJCSI International Journal of Computer Science Issues,* vol. Vol. 9 , no. Issue 3, No 2, May 2012.

[27] S. Mujawar, "Data Analytics Type, Tools,and Their Comparison," 2015.

[28] Sultana.A, G. Priya and Razia, "OLAP (Online Analytical Processing)".

[29] S. K. Song, "Prescriptive analytics system for improving research power," in *Proceedings - 16th IEEE International Conference on Computational Science and Engineering, CSE 2013*, 2013.

[30] R. P. Singh, Design and Research of Data Analysis System for, 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), 2016.

[31] J. Shi, "Research on Database Audit Scheme Design of Life Insurance Industry Based on OLAP Technology," in *2016 International Conference on Logistics, Informatics and Service Sciences (LISS)*, 2016.

[32] S. B. Salem, "Reducing the Multidimensionality of OLAP Cubes with Genetic Algorithms and Multiple Correspondence Analysis," *Procedia Computer Science,* 2015.

[33] P. Saini, "Decision Tree Algorithm Implementation Using Educational Data," *International Journal of Computer-Aided technologies (IJCAx) ,* Vols. Vol.1,No.1, 2014.

[34] R. Yadav, K. Garg and P. Khurana, "Issues and Challenges associated with Association Rules Mining Algorithms," 2014.

[35] K. Zarea, "Epidemiology and associated factors of migraine headache among iranian medical students: A descriptive-analytical study," *Clinical Epidemiology*

and Global Health, 2017.

[36] A. S. Tohir, "On-Line Analytic Processing (OLAP) modeling for graduation data presentation," in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2017.

[37] D. Suseno, "Determining bonus in Enterprise Resource Planning at Human Resource Management module using Key Performance Indicator," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017.

[38] R. Wudhikarn, "Determining key performance indicators of intellectual capital in logistics business using Delphi method," in *International Conference on Digital Arts, Media and Technology (ICDAMT)*, 2017.

[39] D. Plandor, "Generating KPI sets using genetic algorithms," in *Proceedings of the 13th International Carpathian Control Conference (ICCC)*, 2012.

[40] P. Roberto, "Improving Business Decision Making based on KPI Management System," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017.

[41] D. Oreški, "Estimating profile of successful IT student: Data mining approach," in *40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2017.

[42] E. Kandogan, "Just-in-Time Annotation of Clusters, Outliers, and Trends in Point-based Data Visualizations," in *IEEE Symposium on Visual Analytics Science and Technology* , 2012.

[43] W. Krathu, "Identifying inter-organizational key performance indicators from EDIFACT messages," in *IEEE 15th Conference on Business Informatics*, 2013.

LIST OF PUBLICATIONS

[1] A. Phutela and H. Kaur, "Statistical Dimension Identification and Implementation for Student Progression System," *International Journal of Innovative Technology and Creative Engineering,* vol. Vol.8 No.5, May 2018 2018. (Published)

[2] A. Phutela and H. Kaur, "Applying Descriptive and Predictive Analytics on Academic Dataset," in *International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering - (ICRIEECE)*, 2018. (Accepted and Registered)

[3] A. Phutela and H. Kaur, "Commentary upon descriptive data analytics," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018. (Published)