An approach for examining Student Satisfaction in an Indian University

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

Master of Technology

in

Computer Science and Engineering

Submitted By

Himika

Roll No. 801532021

Under the supervision of:

Ms. Sukhchandan Randhawa

Lecturer, CSE Department

Dr. Maninder Kaur

Assistant Professor, CSE Department

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2017

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*An approach for examining Student Satisfaction in an Indian University*", in partial fulfillment of the requirements for the award of degree of Master of Technology in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Ms. Sukhchandan Randhawa* and *Dr. Maninder 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.

(Himika)

(801532021)

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

(Ms. Sukhchandan Randhawa)

Lecturer,
CSE Department

(Dr. Maninder Kaur)

Assistant Professor,
CSE Department

ACKNOWLEDGEMENT

First I would like to thank the Almighty, who has always guided me to work on the right path of the life. This work would not have been possible without the encouragement and able guidance of my supervisors **Ms. Sukhchandan Randhawa** and **Dr. Maninder Kaur**. I thank my supervisor for their time, patience, discussions and valuable comments. Their enthusiasm and optimism made this experience both rewarding and enjoyable.

I am equally grateful to **Dr. Maninder Singh**, Associate Professor and Head, Computer Science & Engineering Department, a nice person, an excellent teacher and a well – credited researcher, who always encouraged me to keep going with work and always advised me with his invaluable suggestions.

I will be failing in my duty if I don't express my gratitude to **Dr. S.S. Bhatia,** Senior Professor and Dean of Academic Affairs, Thapar University, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

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, which made my stay at Thapar University memorable.

Last but not least, I would like to thank my family whom I dearly miss and without whose blessings none of this would have been possible. To my parents, I own thanks for their wonderful love and encouragement. I would also like to thank my brother, since he insisted that I should do so. I would also like to thank my close friends for their constant support.

Date: July, 2017

Place: Thapar University, Patiala

(Himika)

ABSTRACT

Student Satisfaction is a measure to know the contentment levels of the students. Universities nowadays strive to be the best in the global market. The techniques followed for computing student satisfaction vary for each University. In general, a few factors are considered for determining the satisfaction levels. These factors include technology, admin support, co-curricular activities in the University, hostel facilities, library and safety at the campus etc. Higher the satisfaction levels, higher is the retention of the students at the University. Students are treated as customers and happy customers are always beneficial. A lot of study and research has been done in this area in the foreign Universities. Analysis of the data has been done using various tools but none of the approaches used prediction models for satisfaction levels.

This research work includes the analysis and prediction of student satisfaction rate using machine learning on the data gathered from the University. Various machine learning models are applied and an ensemble approach is opted for prediction. Feedback from the University students, reviews from social media are collected and analyzed. Sentiment analysis is applied on the reviews and the results conclude that a more positive response is received through the reviews on social media.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Student Satisfaction, [1] is a measure that defines the student's review regarding the college. The opinion of the student regarding the University is modified by their experience at the college and the educational value received by them. The students at the University are spending resources such as effort, money and time and they aim to achieve quality education. Satisfaction level, at a particular Institute or University alters the motivational level of the students and ultimately changes the rate of retention of the students at a particular University. The educational sector has seen a lot of reforms and the approach of teaching and learning has incredibly changed. Satisfaction level of the students has helped to bring these changes and the educational sector has grown for good. Initially, only the modes of teaching were limited to black boards or white boards but with the advancement of technology, presentations are introduced which help to create a livelier environment in the class. The advancement of technology has led to increase in the interaction level between the faculty members and the students, which again help to improve the satisfaction level of the students. Student satisfaction and the overall experience of the student with the University is a highly contentious topic in the academic literature. The viewpoints of different authors on this topic are quite different from each other. Student satisfaction is a well-researched topic and inspite of the fact that much research has been done on this topic there are still issues which remain unsolved. Student satisfaction occurs when the facilities provided by the University meets or exceeds the student’s expectations. Student satisfaction rate is being calculated which ultimately alters the retention of students in the University. So, student satisfaction and retention are always related.

1.1 Importance of Student Satisfaction

Nowadays, so many services are available to students that were not accessible decades before and the adaptation to these technologies is also easy. The standards have risen and are increasing day by day [2]. On the other hand, with these advancements, the competition has also increased and providing the best services is also becoming a challenge. The feeling of competition has ultimately led to a more complicated lifestyle and leads to higher stress

level among the students. Thus, the task of achieving high level of student satisfaction is a tedious task for Universities these days.

Nowadays, the education sector has also become a commercial sector. To know the contentment of the people related to this sector has become an important agenda. It is very important to focus on student satisfaction in a particular University as it can the help the institutions offering higher education to know their strengths and know their weaknesses and work on those areas. Student Satisfaction not only involves considering the viewpoint of the students, but various other factors which actually affect the contentment of the students. These factors should be studied well and should be known that how these can alter the satisfaction level of the students within a University. It helps the top management of the University to do:

- Strategic planning of the actions.
- Devise and empower various methods for student retention.
- Keep a progress track of the University to achieve the campus goals.
- Improve Institutional marketing strategies.
- Meeting specific accreditation requirements.

1.2 Factors impacting Student Satisfaction

While discussing about the basic classroom environment [3], only a limited number of factors had an impact over the satisfaction level of the students. These factors included:

- The relationship status between the faculty and the students.
- The course curriculum.
- Resources available to students.

As there was, more advancement in the technology many more factors came up in the picture. These included:

- The availability of faculty at the University.
- Availability of career advisors.
- Social life of students on campus.
- Overall relationship between the students and the faculty members.

Thus, the factors will had a great influence over the satisfaction levels are depicted in Fig. 1.1.

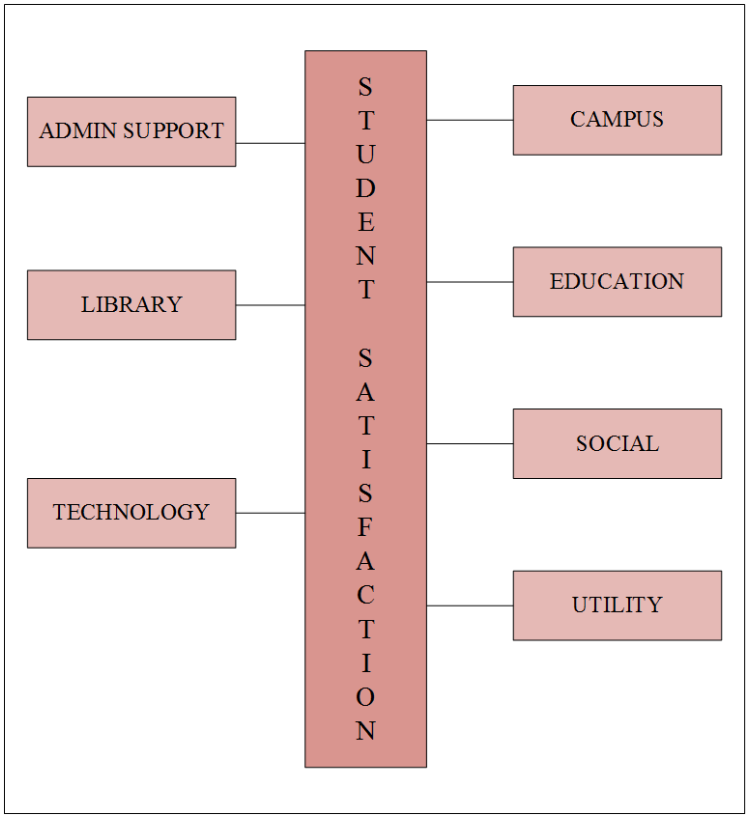


Fig. 1.1: Factors affecting student satisfaction.

The above mentioned factors had an impact when the learners and the instructors have a face to face interaction with each other, but while dealing with the long distance courses a different set of challenges come up. So, while dealing with distance courses various other factors need to be kept in mind while determining the student satisfaction. These factors include:

- **Instructor issues:** The instructor is the main component of in student satisfaction. The availability of the instructor and the response time have a great impact over the levels of satisfaction. It is mandate for the instructor to communicate with the students undertaking the course so that a certain level of understanding is assured at both ends. The involvement of the instructor, keeps the students motivated. The feedback that is attained by the instructor also plays an important role in determining the satisfaction level of the students. It is necessary for the instructor to communicate with his students regularly.

- **Communication:** It is important for the learners and the instructors to communicate among themselves. Communication helps to build a healthy environment between the learners and the instructor. Instructors should initiate and try to complete the course work in stipulated time. People learning distant courses generally have the anxiety levels higher as compared to those learning through physical classrooms. Instructors should try to keep a check over the student's progress and keep him updated about his performance. This technique followed by the instructors help the students to keep them motivated thereby reducing the anxiety levels.
- **Technology:** The success level of the course is related to the technology used in that course. Students enrolled in a particular course must be familiar with the technology used. The availability of tools also changes the levels of satisfaction among the students. Students having unlimited access to the tools outperform students having limited access. The frustration levels also increases if the students do not have an access to the technology.
- **Course Management:** The students learning a particular course online should have a well-structured course. An instructor should be available to them which can help the students and assist them to complete their tasks. Multiple resources such as course textbooks, online library and technical support should be available for them.
- **Institution Website:** It is important for the institute to develop a good website for the students enrolled in online courses. The website should have a rich supply of material and should be well structured. The material should be easy to read and download for the course website. The website should be designed in such a way that the student does not require knowledge of any particular technology and the navigation within the website should be easy and relatable. The links provided in the website should be easily accessible and should offer relevant information. The complete schedule of the course should be readily available at the website for the students to access it. The website should update the events of the University regularly.

- **Interactivity:** Not only the interaction of the instructor and the student is important but the interaction among all the students undertaking that course is also important. It helps to build a healthy environment for competition and keep the students motivated. They strive to work hard and perform better. Collaborative learning tools should be available which help the students to work in groups. Students are able to share and discuss their view points with each other online and discuss about future perspectives about the work.
- **General Information:** Distant learners should be updated for any change in course material. They should keep themselves motivated, organized and committed towards the work. They are solely responsible for their performance and work. The grades obtained by the students also alter the levels of satisfaction among the students. Good grades of the students yield better satisfaction levels as compared to low grades.

The general structure of the student satisfaction is depicted in Fig. 1.2.

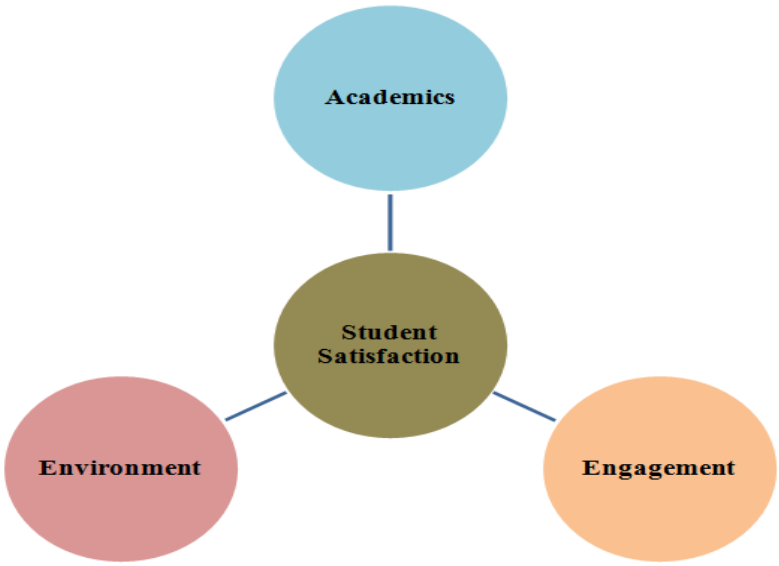


Fig. 1.2: General categorization of student satisfaction.

1.3 Role of Machine Learning in Student Satisfaction

Student satisfaction is gaining importance these days owing to its importance. It is beneficial and important for the University to compute student satisfaction levels. Due to various techniques in the field of machine learning it is possible to predict student

satisfaction levels. *Machine Learning* (ML) has emerged as growing technology and it aims to analyze processed data [4]. ML can be applied to many fields and many applications. These applications include namely web page ranking which makes the user efficient to search queries. Applications like Collaborative Filtering, Automatic Translation, named Entity Recognition and Speech Recognition also can be efficiently solved through machine learning. ML came into existence to solve the problems that were excessively complex or beyond the capabilities of humans to solve. These problems include the problems having large data sets. The change is demanded every now and then and ML provides flexibility to solve the problems [5].

There are different types of learning as depicted in Fig.1.3.

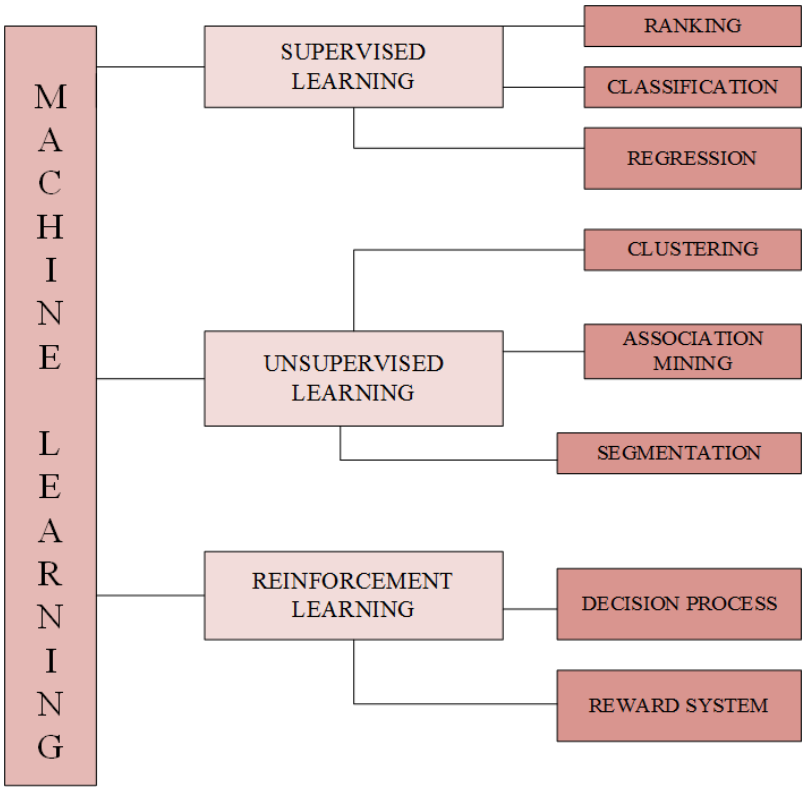


Fig. 1.3: Types of Machine Learning

The interaction of the agent with the actual world can be classified into *Supervised Learning (SL)*, *Unsupervised Learning (UL)* and *Reinforcement Learning (RL)* [6] as depicted in Fig.1.3. Supervised learning is based on labeled data. Supervised learning is about classification and it also maps to the most suitable decision. The classes are predefined in supervised learning. Unsupervised Learning is learning based on unlabeled

data. The hidden structures are recognized and it about clustering the similar patterns together. With the evolution of technology, came into existence a new type of learning known as RL. It widely deals with ML and holds a great significance in the branch of Artificial Intelligence (AI). The growing RL combines the field of Dynamic Programming (DP) and SL to generate a ML system, which is very close to approaches used by human learning. Key feature of RL is that it explicitly considers the whole problem of a goal directed agent interacting with an uncertain environment. Agent must decide itself which action is best to perform according to the given environment. It must take into account its own experience and must update its own strategy according to the appraisal it receives [7].

1.3.1 Sentiment Analysis

Sentiment Analysis or *Opinion Mining* is a technique that helps to know about the perceptions of the users. The users can be customers, students or the viewers. It holds its roots in Natural Language Processing (NLP). An approach is followed in which the data is classified into positive, negative or neutral category [8] depending upon the words that makes up the review. The reviews can also be scaled on 1-5 where 1 represents the extremely dissatisfied and 5 represents extremely satisfied. Sentiment Analysis can be done on movie reviews, customer reviews, organization services and products in the market etc.

Social media has played a very important role in extracting the reviews. Various micro-blogging sites such as Twitter, Facebook have enabled the users to express themselves freely about various issues. They express their views regarding movies, products in the market, various services offered by the company or an institution. Sentiment Analysis has its own set of benefits. The companies have taken a lot of advantage as they can easily know about the reviews of the people regarding their services and products. It helps them to devise new marketing strategies and improve the existing ones. It helps to improve their customer services. On the other hand, sentiment analysis is also beneficial for the individuals. They can know about the review about a particular movie before actually seeing it. It helps them to take right decisions while buying their products.

Sentimental Analysis can also be challenging in some ways:

- It is environment dependent. The review that is negative to one situation can be positive to the other.

- Some people are not expressive which can mislead the customers.
- At various micro-blogging sites, where people express informally, they usually tend to combine multiple expressions.
- The structure of sentence and placement of words also plays an important role.
- English is the most frequent language used and training sets are easily available. In other languages, it becomes difficult to train the models and the datasets are not easily available.

1.4 Thesis Organization

Thesis is organized into various chapters. Outline of thesis is discussed as below:

Chapter 1 illustrates Student Satisfaction, its importance and the factors affecting it. It further continues and discusses about Machine Learning, its types and Sentiment Analysis.

Chapter 2 describes the various approaches followed for Student Satisfaction. It also details about the techniques followed for Sentiment Analysis.

Chapter 3 elucidates the research gaps found in the existing work, problem formulation and the research objectives.

Chapter 4 describes the methods and procedures followed in this research work. It describes about data collection, the proposed questionnaire and the research methodology followed.

Chapter 5 concludes and validates the results of the followed approach. The data collected during the survey is analyzed. The prediction is done using machine learning models. The results for the sentiment analysis for the feedback and the reviews are also documented.

Chapter 6 concludes the work done and states the future scope of this research work.

Chapter 2
Literature Survey

In this chapter, different techniques are illustrated which are applied to compute the student satisfaction rate throughout the world. Nowadays, student satisfaction is becoming a topic of huge concern. All the Universities around the globe work hard to achieve a higher rate of satisfaction. Various parameters are considered that takes into account all the aspects of the University. Student Satisfaction and Student Retention are closely related. Therefore, to retain students in the University it is mandate to keep them satisfied. Social media is an emerging platform and it has helped a lot to review about the University. Likes or ratings offered at social media are not sufficient to determine the satisfaction levels. The comments given by the participants also need to be deeply studied. *Sentiment Analysis* and *Opinion Mining* has helped a lot to investigate about the reviews being offered at social media.

2.1. Student Satisfaction Literature

Up till now, various research methodologies have focused on analyzing various parameters that alter the satisfaction levels of the students. In the following section, the approaches followed for computing satisfaction levels on different parameters is being discussed. **Castillo. L** [9] investigates a multiagent programming course, by introducing the concept of virtual laboratory. The courses taken into account has been opted by many students. Due to the introduction of the virtual laboratory, it became easier for the students to access it. Owing to the extensive use of laboratory, huge amount of data has been collected for analyzing. This research work emphases on the detection of patterns, depicting both the failed and successful behaviour and offers data driven assessment. It also helped in analyzing the trend in the changed behaviour of the students [10, 11]. The students must interact with the agents in a limited manner, connected by a server through their own multiagent system and solve various virtual problems.

Kamis. L et.al. [12] aims to determine satisfaction levels of the students based on the physical amenities presented by Politeknik Sultan Salahuddin Abdul Aziz Shah (PSA). These facilities include the basic classroom facilities, hostel rooms and cafeteria etc. The researchers aim to conclude the perception of the participants regarding sustainability. The

perception of the students on the available facilities offered to them at PSA also needs to be studied. The satisfaction rates of the students due of the availability of physical resources in PSA needs to be determined. The research methodology included the analysis of data using Statistical Package for the Social Science (SPSS). Descriptive mean and frequency were the parameters, which helped to find the students' perception over the concept of sustainability. Cluster sampling was also used to classify the characteristics.

Zamakhsari.Z and Ridhzuan.A [13] tried to explore parameters like satisfaction and participation towards online learning through student's perspective at the Universiti Technologi MARA Pahang. Due to lack of student's interest towards online courses, it becomes difficult for the University to implement courses through online learning approach. Firstly, the researchers aim to detect the form of online activities among students. Secondly, tried to examine the methods taken by faculty members to inspire the students to participate in online activities. Thirdly, to assess the relationship between participation and satisfaction level of the students. This research concluded that quiz and online test is the activity that gained the student's attention to the maximum. The students also searched for online notes and be a part of online discussions.

Mendonca.M et al. [14] developed a Fuzzy Cognitive Map (FCM) that evaluates the excellence level of students at Federal Technological University of Parana (UTFPR-CP). This study associates various other intelligent techniques as well. It aims to find out those possible areas that can alter the satisfaction level of the students. Various topics such as teacher training, the overall built and condition of classrooms and laboratories are discussed. Many educational Institutions fail to recognize their flaws due to unavailability of an assessment tool. The proposed tool aims to benefit the educational Institutions in long term.

Alshammari. M et al. [15] discussed how important it is to analyze student satisfaction based on learning style adaptation. Satisfaction plays an important role in motivating the students that ultimately improves their learning. This approach involves building of such paths that was implemented through adaptive e-learning. These techniques led to improved levels of satisfaction among the students and motivated them.

Nikolic. S et al. [16] analyzed the importance of laboratory in student satisfaction levels. The research is done in the Electrical Engineering department of Australian University.

Multiple surveys were conducted on all the available laboratories and the results were analyzed to improve the performance of the laboratories [17]. The surveys conducted had an overall impact over the overall satisfaction. It also had an impact over the laboratory conditions that included laboratory notes, the computers that are provided and the other engineering equipment that are being provided to the students. The research concluded that there is an incredible increase of about 32% in the satisfaction level of the students from the year 2007 to 2013.

Sohoraye.M *et al*. [18] aims to determine whether facebook likes are sufficient to determine student satisfaction in Open Distance Learning (ODL). The assessment of the satisfaction is done through Online Social Networks (OSN). Social media (OSN) has become a large platform for promoting ODL. It is often believed that student satisfaction can be determined by facebook likes only, but this is not the case as student satisfaction depends upon many other factors also. This work shows the complete analysis of some of the ODL institutions, by taking into consideration the data obtained from their facebook pages. Quantitative method is chosen for the analysis to be done. The Institution on which the analysis is done is chosen based on their association with the open distance learning. The work also proposed that there is a repetition of questions from the students on the pages and student prefers doing that instead of looking up at the websites. Multiple solutions have been proposed for the Institutions, which use social media as a platform to promote their ODL programs. These include meeting the locals, keeping the knowledge of students by dealing with them, learning about the culture and their native language. It also proposed to select good ambassadors for representing them. The survey concluded that to measure student satisfaction, facebook likes are not sufficient. It is important for the Institutions to also focus on the comments made on these pages and not solely depend upon the likes on pages.

Armenski.G *et al*. [19] finds student satisfaction for courses like *Computer Architecture and Organization* by the use of multiple e-Learning tools. Among the multiple tools used, the first being is the interactive online learning tool where the students are allowed to evaluate their knowledge. The other tools allowed the student to visualize their concepts, regarding the technology that is being taught to them. A hypothesis is set which decided that the latter tool i.e. the visual tool will be preferred by the student tool as compared to

the former one. The paper further continues and discusses about the visual simulators which are EDUCache Simulator and HADES simulator. The survey evaluated the students by asking them multiple questions regarding the courses undertaken by them, the grades that they got in that particular course. It even inquired them about their experience with the online tests and about their laboratory experience while dealing with the visual simulators. The survey concluded that the students find the course of Computer Architecture and Organization difficult. Contrary to the hypothesis, it was found that the students prefer the online learning tool over the visual tool. About 76% of the students who undertook the survey were positive regarding the set of tools used, which encourages the authors to conduct multiple surveys and extend their work to other visual simulators as well.

Zhang. H et al. [20] evaluated teaching quality through student satisfaction. Multiple factors such as attitude of teachers, their teaching style, the content taught and the methodology followed by them are kept in mind to determine the overall student satisfaction levels. The complete survey is divided into first call indexes, which further sub divides into second-class indexes. Teaching attitude falls into first class indexes. Preparing the lectures well to be delivered, giving appropriate homework and checking it regularly all falls into second class indexes under the first class index teaching attitude. They also offer some suggestions to improve the process of teaching

Choudhary.M.A [21] discussed the factors that have influenced the engineering students' performance and they tried to discuss how student satisfaction is related with teaching along with other University experiences. The research is done at the University of Engineering and Technology, Taxila. The research concluded that the people who took part in this survey were satisfied with services such as academic and learning facilities but were not satisfied with the facilities like as computing, career opportunities, hostel facilities and recreational facilities. It also concluded that the students participating in co-curricular activities have outperformed in every field as compared to those who have not. While comparing within the class, it is found that the female respondents have spent more time for academics as compared to male candidates. A special consideration is given to check the social activities done by students of the University. It is found that most of the students are not involved in any political campaigns and have not interacted with students belonging to other race. Most of the students have not smoked either which is a good indication for

the University. Around 70% of the people want to pursue their higher studies i.e. either Masters or Doctorate degree. This survey has really helped the University to improve the interaction between the faculty and the students.

Table 2.1: Existing Student Satisfaction Techniques

Year	**Problem Statement**	**Research Parameters**	**Approach**
Castillo.L [9] 2016	To determine satisfaction levels for a Multiagent Programming course by the introduction of a virtual laboratory for the course.	Data driven assessment by detection of patterns.	Students must interact with agents through a server.
T.Kamis *et al.* [12] 2015	To determine satisfaction level among diploma in Medical Electronic Students.	Determine student perception: • On sustainability • Existing physical facilities	Cluster Sampling was used to identify the characteristics. Questionnaires were used to determine the level of student satisfaction based on the research parameters.
Z.Zamakhsari *et al.*[13] 2015	To determine student participation and satisfaction towards online learning at University Technology MARA Pahang	To identify the type of online learning activities used among students. To investigate the approaches used by lecturers to encourage student participation in online learning. To examine the relationship between student's participation and satisfaction towards online learning.	Data was collected through questionnaire and the questions were scaled from 1 to 5. Descriptive measures, mean and standard deviation were used to determine the level of satisfaction based on research parameters.

M.Alshammari [15] 2015	To determine students' satisfaction in learning style based adaptation.	To enhance traditional e-learning systems. The use of Adaptive e-learning systems is discussed.	Data was collected using questionnaire. Satisfaction was measured using e-learning satisfaction (ELS) tool.
S.Nikolic *et al.* [16] 2015	Determine student satisfaction level, to improve laboratory experience at the Electrical Department of Australian University.	Aims to examine: • Overall Satisfaction • Laboratory Notes • Learning Experiences • Computer facilities. • Engineering equipment • Condition of the laboratory.	Data was continuously collected from the students by making them rate the questions on the scale 1 to 5, from the year 2007-2013.
P.Cunningham [18] 2014	To determine whether facebook likes are enough to measure the level of satisfaction of the students in Open Distance Learning (ODL)?	The objectives of the study are: • To compile a list of possible issues evoked by learners. • To analyze comments posted to discern trends. • To make recommendations based on the way responses are made and how they can be better addressed.	The focus of the study is not only the number of "likes" but on how effectively comments made have been looked into and the speed at which these are treated.
G. Armenski *et al.* [19] 2014	To determine the satisfaction level for the use of e-learning tools for the course of Computer Architecture and Organization.	To compare the educational tools belong to two different approaches: • Visual simulator tools • An interactive adaptive environment for	Analyzed two different Visual tools i.e. EduCache Simulator and HADES simulator. Data was collected by a survey asking questions regarding the course and

		knowledge assessment.	comparing the educational tool and the visual tool.
A. Carbone and J. Ceddia [22] 2013	To improve those areas which have critically low satisfaction rate in Physical Sciences at Monash University.	Keeping in mind the students' perspective, what are the common areas in physical science units that are most in need of improvement?	The comments of the students were studied and categorized into multiple categories by two independent researches and the process was repeated multiple times.
U. Udwhg *et al.* [23] 2012	To determine the factors that influences the level of performance and student satisfaction for engineering students at The University of Engineering and Technology, Taxila.	All the parameters were considered which included: • Academic facilities • Lab facilities • Student housing • Career advising • Interaction of students with the professor.	About 25% students from a total population of 3200 were selected and were given a comprehensive questionnaire to answer.

2.2. Sentiment Analysis Literature

Until now, various research approaches have focused on sentiment analysis. Multiple methodologies have been proposed to deal with reviews and the following section discusses a few of them.

Mullen.T and Collier.N [24] introduced an approach that makes use of Support Vector Machines (SVMs) to determine sentiment analysis. The reviews are collected for various movies from Opinions. The hybrid SVMs have shown better performance as compared to the simple SVMs. The value phrases are obtained and the values corresponding to these phrases are found. The first dataset contained approximately 1380 movie reviews that were almost 50% positive and the remaining were negative. The results were compared for 3 fold cross validation and 10 fold cross validation. The second dataset was recorded through

Pitchfork Media. It had 100 records and due to limited number of records 5-fold cross validation was applied to this dataset.

Kouloumpis. E et.al [25] classifies the tweets available at Twitter by using linguistic features. They have used multiple categories of twitter data in their experiments. The dataset is divided into *Hashtag dataset*, *Emoticon dataset* and *iSieve dataset*. The hashtag dataset is formed by filtering out tweets from the total set of tweets collected. Out of 97 million tweets collected only 4 million tweets had hash tags in them. These tweets are then analyzed to get an idea for the positive, neutral and negative tweets. The emoticon dataset only had ":)" which represented positive review and ":(" that represented a negative review. The *iSieve* dataset had only 4000 tweets and the value associated with them depicted their nature. Features that included part of speech did not prove very helpful in sentiment analysis in the field of microblogging. Collected hashtag and emoticon dataset definitely proved beneficial in fetching better results.

Go.Alec et.al [26] introduced a method that classifies the twitter messages into either positive or negative. This type of research is really helpful as it can help the customers to know the reviews about the product. Machine learning algorithms are applied to twitter data to classify them. The training data for the models constituted of tweets with emoticons. Various machine learning algorithms applied in this research are Naive Bayes, Support Vector Machine and Maximum Entropy. The results are analyzed on multiple features such as Unigram, Bigram and the combination of the both. The classifiers applied have shown a great accuracy.

Lin.C and He.Y [27] designed a framework that makes use of Latent Dirichlet Allocation (LDA), that classifies the textual data into different sentiments. This approach is referred to as Joint Sentiment/Topic Model (JST). While comparing it to machine learning algorithms, they act quite differently to this approach. Machine learning algorithms make use of labelled data and JST makes the use of unsupervised data. Movie reviews are selected for the checking the proposed model's performance. The results show that the proposed model is at par with the previous models that make use of supervised techniques. This approach has a limitation that it does not consider the ordering of words that leads to improper classification of tweets. The researchers also aims to use bigrams and trigrams in their proposed model in future.

Prabowo.R and Thelwall.M [28] designed a new approach that makes use of the existing approaches for sentiment analysis. The proposed approach makes use of supervised learning, rule based classification and machine learning algorithms. The results are compared for different existing classifiers and the hybrid approach that is designed. This proposed approach is applied over multiple datasets that include product reviews, movie reviews and comments available at MySpace. The performance of existing classifiers is also kept in mind. An approach is designed that helps the existing classifiers to improve their accuracy.

Martineau.J and Finin.T [29] proposed an approach that takes a step forward than the existing machine learning algorithms. They have presented a methodology known as Delta TFIDF that aims to give ratings to the words even before they are classified. The weights to each word are assigned by the number of occurrences of that word given in that document. In this approach, the weights given to the words depends upon their difference between the positive score and negative score. The results show that Delta TFIDF, have outperform the simple approaches in the terms of sentiment polarity and subjectivity detection.

Boiy.E and Moens.M.F [30] used machine learning approach to analyze sentiments of web texts in multiple languages. The reviews and texts for sentiment analysis is collected through multiple blogs. The machine learning models are trained with the dataset that already has classified reviews. The reviews are classified into positive, neutral and negative. Various problems are encountered while training machine learning models. First is the nature of the text gathered i.e. noisy data was available. Second, is the sentiment score allotted to each entity. The third is the size of training dataset i.e. very small. The results conclude that while dealing with English text the accuracy recorded is maximum. The accuracy of machine learning models while dealing with Dutch and French texts is comparatively less than that of English. The accuracy of the latter texts is less as compared to the former one because of multiple variety of texts. The training set needs to be improved and the machine learning models needs to be trained repeatedly to achieve better accuracy results. The authors have also discussed various active learning techniques that can efficiently reduce the manual annotations of the examples.

Gamon.M [31] discussed a method that allows automatic categorization of feedback received from customers. The data that is obtained is usually very noisy. It is observed that linear SVM can be trained really well by opting for feature reduction techniques for large feature vectors. The analysis done on feedback data has shown a decrease in accuracy as compared to that movie data. The reviews are also available in multiple languages that have also contributed in altering the accuracy levels.

Chapter 3
Problem Definition

In this specific chapter, the gaps that existed in the present work are discussed. This chapter states problem in hand and lists out the various objectives that are required to be met for solving the problem.

3.1 Gap Analysis

In the literature review that is discussed, various gaps are encountered. The various loopholes seen in the existing work is depicted below:

1. There is no common tool or platform that can compute student satisfaction in every University.
2. Till now, the Universities which have computed student satisfaction have only analyzed various parameters using different analytical tools and no prediction has been done yet.
3. Social media is not taken into consideration for computing student satisfaction for any University.
4. Sentiment Analysis is not done on multiple reviews collected through surveys.
5. Various Universities collaborate with other foreign universities and offer exchange programs for the students. Sentiment Analysis can be done for English, but the performance of the classifiers reduce drastically for other foreign language.

3.2 Problem Statement

Student satisfaction is a parameter for analyzing the contentment levels of the students in the University. It is very important in today's competitive world. The satisfaction level affects the retention level of the students and helps the University to achieve a better position in the market. Various factors such as academics, research, hostel, admin support, technology are responsible for modifying the satisfaction levels in the University. These factors need to be considered for calculating the satisfaction levels. Machine learning approaches are beneficial for the prediction of data and are considered for determining the

satisfaction levels. The proposed research also focusses on the reviews obtained through social media platforms.

3.3 Research Objectives

1. To study the approaches and techniques that are applied in the past for computing student satisfaction.
2. To design a questionnaire and collection of data in the University.
3. To focus on various parameters that modify the satisfaction levels within the University.
4. To analyze satisfaction levels in the University.
5. To apply different machine learning approaches for prediction of student satisfaction.
6. To analyze the reviews collected within the University and check their contribution in student satisfaction levels.
7. To take into account social media to analyze the reputation of the University and determine the satisfaction levels of the students.

Chapter 4
The Proposed Methodology

This section of thesis details about data collection technique and research methodology followed. Questionnaire is used as a mode for collection of data and the proposed questionnaire is discussed. The pros and cons for the selected mode of survey is also discussed. This section illustrates about the data, reviews obtained and the classification of reviews done. The reviews are classified into positive, neutral and negative classes.

4.1. Data Collection

Data can be broadly classified into *Primary Data* and *Secondary Data*. When the information is already available and data is only meant to be extracted, then it is referred to as *Secondary data*. On the other hand, when no prior information is available and the data needs to be collected from scratch, it is referred to as *Primary Data*. Secondary data can be collected through sources like government publications, census reports, personal records and through online sources [32]. Primary data, on the contrary needs to be collected through personal observations, one or one or group interviews, questionnaires [32]. J.W. Creswell [33] have divided the qualitative collection procedures into four broad categories viz. Documents, Interviews, Observations and audio- visual materials.

In this research work, questionnaire is used as a data collection technique. The main aim of the study is to calculate *Student Satisfaction Rate* and analyze the weak and the strong parameters of the University. The satisfaction level of the students at the University is the focus of the research.

4.1.1 Structure of Questionnaire

The questionnaire proposed is well divided into multiple sections and it considered various factors related to the University. Some of the parameters considered are the academics, teacher student relationship, hostel, guide availability, gender, research and co-curricular activities. 438 students from the University participated in this survey. For understanding the importance of the satisfaction level in the University, it is well distributed on a scale of 5 as: 1-Very Dissatisfied, 2- Somewhat Dissatisfied, 3- Neutral, 4- Somewhat Satisfied, 5- Very Satisfied. Once the first draft of the questionnaire is prepared, a printed set of

questionnaire is handed over to 20 students of the University to ensure whether they are easily able to comprehend to the questions. The feedback received through them is improvised and the questionnaire is again refined and distributed.

Questionnaire is selected as a mode to collect primary data by considering its advantages [34] and disadvantages [35]. The advantages of using questionnaire is:

- It is considered to be more practical and an easy way to gather information.
- Validity and reliability of the data remains unaltered even if the dataset is researched by any number of people.
- The actual dataset collected can be modified according to the research.
- It is scalable and can fetch speedy results.
- Large amount of information can be collected in a short interval of time.
- The dataset can be researched and can be modified according to the aim of the study.
- The research done can be compared with the existing research and methodologies.

This mode of survey has its own set of disadvantages. They are:

- The participant may have some issues in interpreting the questions due to which some questions can be skipped.
- The students may feel that there is lack of personalization.
- The environment in which the students are filling the questionnaire can play an important role.
- The participants might be dishonest and partial.

4.2 The Proposed Questionnaire

After incorporating the changes in the questionnaire, it is circulated in the University through online means and thereon the responses are collected. The proposed questionnaire is as follows:

SECTION 1. GENERAL INFORMATION		
What is your gender? • Male • Female	What is your course? • Graduation • Post-Graduation • Doctorate	Which year? • 1 • 2 • 3

		• 4
What are you? • Day Scholar • Hosteller	Which hostel do you live in?	What is your current CGPA?
What is the mode of teaching that you prefer? • White Boards • Presentations • Combined	Are your teachers audible in your class? • Yes • No	Is white board visible? • Yes • No
If you opt for higher education, would you choose this college again? • Yes • No	Does the strength of your class affects the quality of teaching? • Yes • No	Is discipline well maintained in your class? • Yes • No
Are any malpractices observed within the college campus? • Yes • No	Have you been a member of student club/ organization? • Yes • No	Is guide available to you? • Yes • No
SECTION 2. OVERALL SATISFACTION LEVELS **Satisfaction Rating** **1 – Very Dissatisfied 2- Somewhat Dissatisfied 3- Neutral 4- Somewhat Satisfied 5- Very Satisfied.** **How satisfied are you with:**		
The orientation program/ awareness program provided when you were a new student at this college (eg. New student orientation activities, different block locations, timetable, schedule of classes)?	Out of class experiences organized (e.g. Members of student clubs, participating in sports, participating in organized cultural or social activities)?	Your informal participation out of class (e.g. attending plays, hearing speakers, having informal student discussions)?

Fees structure of your college?	The academic procedure that is being followed?	The extent to which faculty follows the methods like class presentations, assignments or discussions to increase the productivity of students?
The safety and security at your campus?	The quality of teaching?	Your opportunity to participate in an independent research project with a faculty member?
The syllabus that is structured for your course?	The classroom/ lab facility?	Your opportunity to participate in a study abroad program?
The opportunities to meet the faculty outside the classroom?	How well are you able to correlate while the teacher is teaching?	The approach with which the college authorities deal the grievances of the students?
The availability of courses to make progress towards your degree?	The schedule of your classes ?	The campus resources for students (such as scholarships, reductions in fees with low yearly income or good grades scoring students)?
The use of technology in your class (like projectors, different android applications)?	With the campus bookstore?	The internet facilities at your campus?
Accessibility of your teachers (by emails, Phone calls)?	The student financial aid services (like funding of projects)?	The medical services?
Overall quality of examination procedure?	The campus library services?	The recreational facilities (gym, swimming pool) in campus?

The placement drives in campus?	The food services provided in campus?	
SECTION 3. NUMERAL QUESTIONS		
What is the strength of your class? • <30 • 30-70 • 70-100 • 100-150 • >150	When are most of your classes scheduled? • 8am-4pm • 10am-5pm • 12 noon- 7pm	How many faculty/staff members do you know well whom you can ask for a personal or a professional advice ? • Zero • 1-3 • 4-6 • 7-9 • 10 or more
SECTION 4. HOSTEL FACILITIES Satisfaction Rating **1 – Very Dissatisfied 2- Somewhat Dissatisfied 3- Neutral 4- Somewhat Satisfied 5- Very Satisfied.** How satisfied are you with:		
The overall hostel facilities?	The cleanliness of hostel rooms?	The bedding provided in hostels?
The basic requirements in hostel (like water purifiers, washing machines)?	The hostel mess food?	The timings of food availability in the mess of your hostel?
The Internet facilities in your hostel?	The other basic facilities in your hostel (gym, lift)?	
SECTION 5. ESTIMATE THE NUMBER OF TIMES **1. Zero times 2. 1-2 times 3. 3-4 times 4. 5-6 times 5. 7 or more**		

Worked on academic research with faculty outside the class? (1 project = 1 time)	Participated with faculty members on activities other than course work (eg. Student organization, out of class activities)?	Attended a lecture outside class (University sponsored speakers/ presentations)?
Attended a cultural event (art exhibition, plays, dance or theatre performances)?	Participated in spiritual/ religious activities on/off campus (worship, meditation, prayer)?	Attended an athletic event?
Participated in community events or organizations (like IEEE, ACM)?	How many class sessions do you skip in an average week ?	
SECTION 6. HOW MANY HOURS DO YOU SPEND? • Zero • 1-5 hours • 6-10 hours • 11-14 hours • 15 or more		
Working (for pay) on/ off campus?	Studying / doing homework/ team projects outside a class?	Voluntarily practicing in community service ?
Exercising or practicing in clubs or sports ?		
SECTION 7. Express your feedback in words.		
// Add comment here		

4.3 Data Collection and Preprocessing

438 students participated in the survey. The collected dataset contains 78 features as each question in the questionnaire is represented as a feature in the dataset. These participants are pursuing their graduation, post graduation and doctorate degrees from the University. The data is collected and then preprocessed before actually feeding it into machine learning models and sentiment classifiers. The data is preprocessed by removing redundant entries,

filling in missing values, removing outliers and by converting nominal values into ordinal values. Fig 4.1 depicts the raw data that is collected during the survey. Fig 4.2 depicts the processed data that is ready to be given as input to the machine learning models.

Fees structure of your college?	The safety and security at your campus?	The syllabus that is structure	The orientation program	If you are starting college again,	What is the strength of your	What is the mode of teaching that you prefer?
1	4	3	4	Yes	70-100	Combined
3	5	4	2	No	30-70	Combined
2	2	3	3	Yes	30-70	Combined
2	5	4	4	Yes	30-70	Combined
2	4	3	4	No	>150	Combined
3	4	4	2	Yes	<30	Presentations
2	4	2	3	No	<30	Combined
4	5	3	5	Yes	<30	White Boards
4	4	4	4	Yes	30-70	Combined
3	5	5	4	Yes	30-70	Combined
2	5	3	3	No	30-70	Combined
1	2	3	2	Yes	<30	Combined
1	3	4	5	Yes	100-150	Combined
2	3	4	5	Yes	70-100	Combined
1	3	3	3	No	100-150	White Boards
4	3	2	4	Yes	70-100	Combined
1	2	2	4	Yes	70-100	Combined

Fig.4.1: Raw data received during the survey.

Few of the conversions of the nominal values into ordinal values are depicted below:

Nominal Values	Changed values
Yes	1
No	0

Nominal Values	Changed values
<30	1
30-70	2
70-100	3
100-150	4
>150	5

Nominal Values	Changed values
White Board	3
Presentations	4
Combined	5

Fees structure of your college?	The safety and security at your campus?	syllabus that is structured for your course?	orientation program / awareness	starting college again, would you like to	What is the strength of your class?	What is the mode of teaching that you prefer?	satisfied are you with the quality of teaching ?
1	4	3	4	1	3	5	2
3	5	4	2	0	2	5	3
2	2	3	3	1	2	5	3
2	5	4	4	1	2	5	4
2	4	3	4	0	5	5	1
3	4	4	2	1	1	4	3
2	4	2	3	0	1	5	2
4	5	3	5	1	1	3	3
4	4	4	4	1	2	5	5
3	5	5	4	1	2	5	4
2	5	3	3	0	2	5	4
1	2	3	2	1	1	5	3
1	3	4	5	1	4	5	3
2	3	4	5	1	3	5	3
1	3	3	3	0	4	3	4

Fig 4.2: Processed data for machine learning models.

The feedback is taken from the students of the University through the questionnaire. To understand the problems and the view point of students reviews are also collected from various social networking sites like facebook. The students feel free to express themselves at social media. The data received though feedback and social media is cleansed by removing unstructured textual data. Fig 4.3 shows the feedback received from the participants. Fig 4.4 depicts the cleansed feedback that can be classified by the classifier to get proper results. Fig 4.5 depicts various reviews collected through social media. Fig 4.6 illustrates the cleansed data to be fed to the classifier.

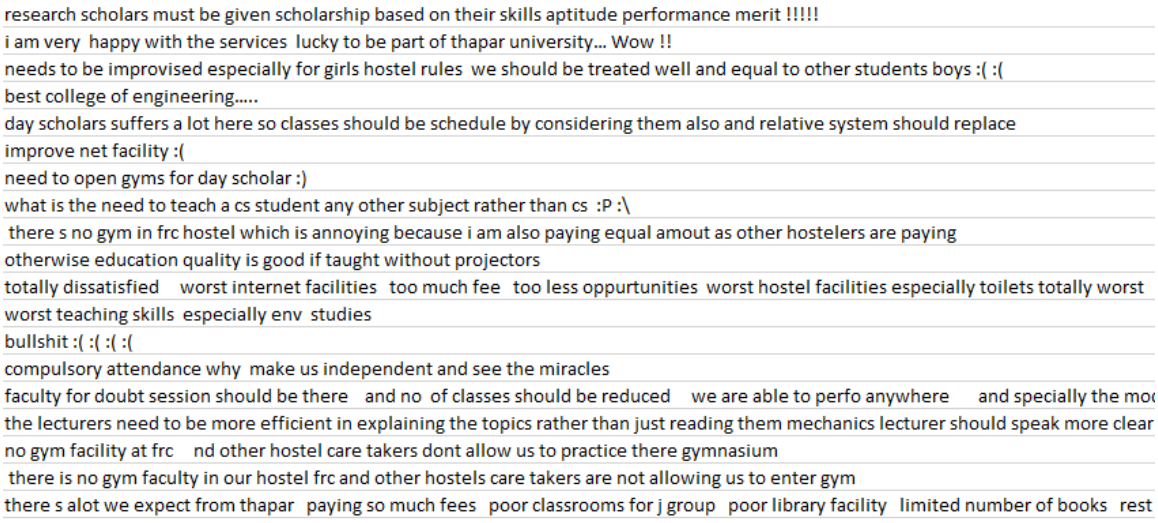


Fig 4.3: Feedback received during the survey.

The data obtained as feedback through the survey is raw data. The data is cleansed by removing noisy data, punctuation marks before feeding them into classifiers for the analysis.

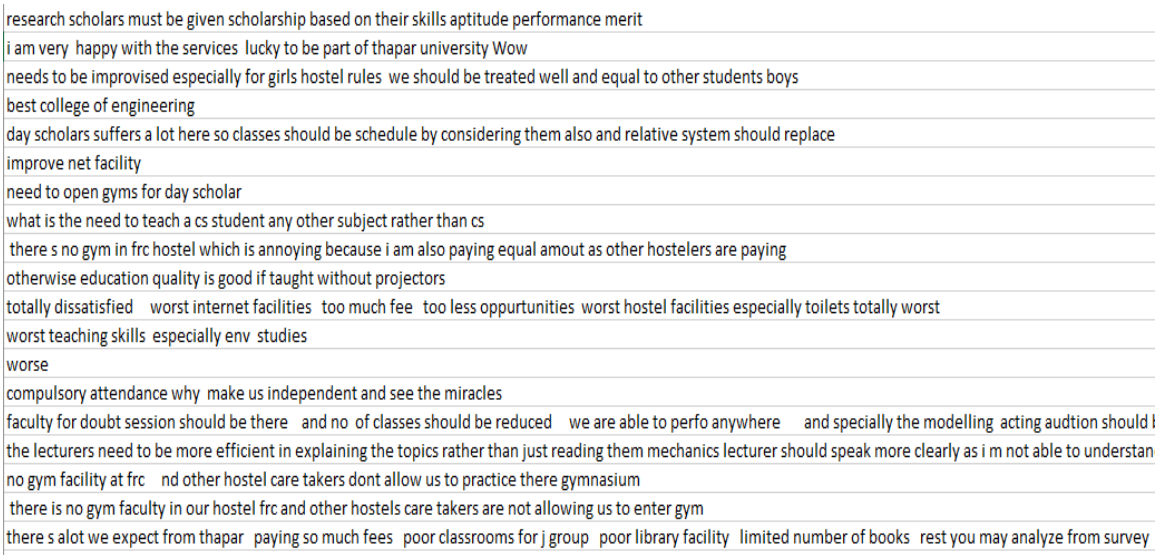


Fig 4.4: Cleansed feedback for the classifier

Thapar university is an excellent place of learning where you can persue your dreams in a thorough professional and disciplined manner
Its really an asset to patiala.
Lack of Thapar Culture..Works more like a school
The university is great with strong alumni association +the new director is very supportive!
Infrastructure is good
75% compulsory attendance and 8am-5pm classes
TU has also come up with minor courses in finance,business administration etc.
The reputation of the university is also great!
also good canteen is average library facilities are excellent
excellent sports facilities are very good..... faculties were also good as compare to other college and this is a best college in Patiala.....
) No , board percentage: If you are that guy, who screwed up his board percentage and got a fairly decent marks in jee- mains,Thapar is fit for you.
Thapar university is really welcoming, i can speak of what i have been through, i can not say about other colleges. So, if you have jee marks 150+, i suggest you to join thapar."
. Sports facilities are best with swimming pool, synthetic athletic track, 4-5 lawn tennis courts, badminton court, gym facilities in every hostel.
Lack of college support in things like entrepreneurship
Placements are good
Infinite events that happen here help in improving organising skills,help in personality development...
Lots of societies..
Always miss this place. Some of best lessons I learned here. Meet with great people here who inspired me alote.
An amazing place to study! Great college, best faculty and a wonderful learning experience

Fig.4.5: Reviews obtained through social media.

Social media is taken into consideration by collecting reviews. The data obtained as reviews is cleansed by removing noisy data, punctuation marks before feeding them into classifiers for the analysis.

Thapar university is an excellent place of learning where you can persue your dreams in a thorough professional and disciplined manner
Its really an asset to patiala
Lack of Thapar Culture Works more like a school
The university is great with strong alumni association the new director is very supportive
Infrastructure is good
75 compulsory attendance and 8am 5pm classes
TU has also come up with minor courses in finance business administration
The reputation of the university is also great
also good canteen is average library facilities are excellent
excellent sports facilities are very good faculties were also good as compare to other college and this is a best college in Patiala
No board percentage If you are that guy who screwed up his board percentage and got a fairly decent marks in jee mains Thapar is fit for you
Thapar university is really welcoming i can speak of what i have been through i can not say about other colleges So if you have jee marks 150 i suggest you to join thapar
Sports facilities are best with swimming pool synthetic athletic track 4 5 lawn tennis courts badminton court gym facilities in every hostel
Lack of college support in things like entrepreneurship
Placements are good
Infinite events that happen here help in improving organising skills help in personality development
Lots of societies
Always miss this place Some of best lessons I learned here Meet with great people here who inspired me alote
An amazing place to study Great college best faculty and a wonderful learning experience
Great place to realize dreams specially mentioning Computer Science and Engineering Department with Elective Focus on
High Performance Computing Machine Learning Data Analytics Computer Animation Gaming Cyber Information Security Software Engineering
Got full support and motivation from the administration to demonstrate and explore the teaching and research potential

Fig. 4.6: Cleansed reviews obtained through social media.

4.4 METHODOLOGY USED

The complete research methodology is divided into two main modules. The first module is the *Prediction Module* and the second module is the *Sentiment Analysis* module as illustrated in Fig.4.7.

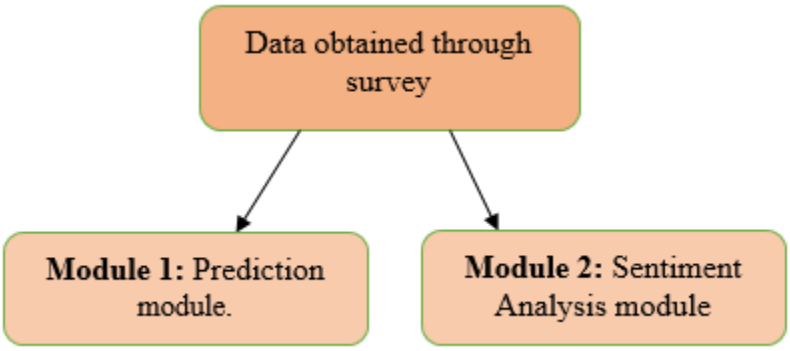


Fig.4.7: Broad classification of the research methodology.

The detailed methodology of module 1 and module 2 is represented in Fig.4.8. The questionnaire is designed and circulated through online means. The data is collected and saved in .csv format. The dataset has 78 features. The data is preprocessed by removing all the missing values, errors and by converting nominal to ordinal values.

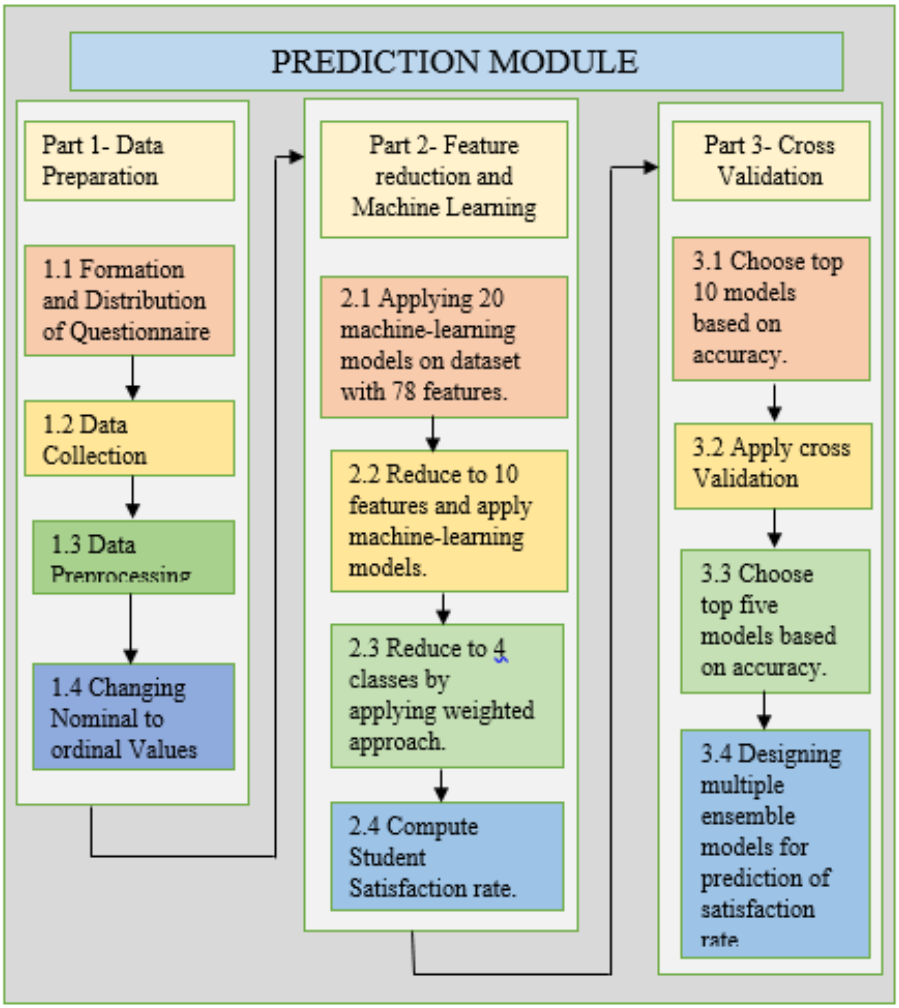


Fig. 4.8 (a): Module 1 of proposed framework.

Prediction Module consists of three submodules that includes (1) Data Preparation (2) Feature reduction and Machine Learning and (3) Cross Validation. The first sub module deals with the data. It includes the formation and circulation of questionnaire. The data collected is cleansed by applying different techniques. The second sub module deals with feature reduction. The original dataset included 78 features that were reduced to 10 features and then further reduced to 4 features. Satisfaction rate is calculated for the University and machine learning models are applied for the prediction of data. The third sub module chooses top 10 models based on their accuracy and 10- fold cross validation is applied to choose top 5 models. Ensemble approach is applied to design a model that can be used for prediction of data.

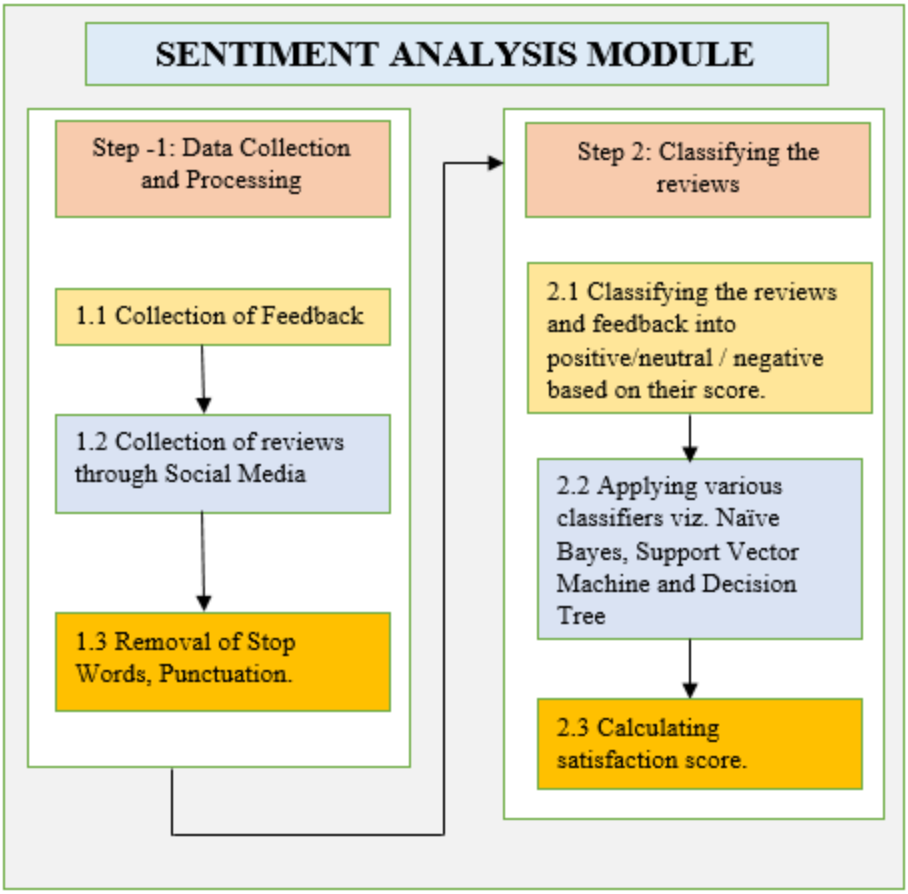


Fig.4.8 (b): Module 2 of the proposed framework.

Machine learning models are applied on this dataset as described in Table 4.1. The results are recorded. The dataset is evaluated on parameters like correlation, R-Squared, Root Mean Square Error and Accuracy.

Table 4.1: Different Machine Learning models applied.

Model No.	**Model Name**	**Type**	**Libraries**	**Tuning Parameters**
1	Bagged CART [36]	Dual Use	ipred, plyr, e1071	None
2	Bagged MARS using gCV Pruning [37]	Dual Use	Earth	Degree
3	Boosted Generalized Linear Model [38]	Dual Use	plyr, mboost	mstop,prune
4	Boosted Linear Model	Dual Use	bst, plyr	mstop, nu
5	CART	Dual Use	Rpart	Cp
6	Conditional Inference Random Forest [39]	Dual Use	Party	Mtry

7	Conditional Inference Tree [40]	Dual Use	Party	Mincriterion
8	Extreme Learning Machine [41]	Dual Use	elmNN	nhid, actfun
9	Gaussian Process with Polynomial Kernel [42]	Dual Use	Kernlab	degree,scale
10	Generalized Additive Model using Splines [43]	Dual Use	Mgcv	select,method
11	Glmnet [44]	Dual Use	h2o	Ntrees, max_depth, min_rows, learn_rate, col_sample_rate
12	K-Nearest Neighbour [45]	Dual Use	Kknn	kmax, distance, kernel
13	Multilayer Perceptron [46]	Dual Use	RSNNS	Size
14	Multivariate Adaptive Regression Spline [47]	Dual Use	Earth	nprune,degree
15	Parallel Random Forest [48]	Dual Use	e1071, randomForest, foreach	Mtry
16	part DSA	Dual Use	partDSA	cut.off.growth, MPD
17	Partial Least Squares [49]	Dual Use	Pls	Ncomp
18	Boosted Tree [50]	Dual Use	party, mboost, plyr	mstop,maxdepth
19	Self Organizing Map [51]	Dual Use	kohonen	xdim, ydim, xweight, topo
20	Stochastic Gradient Boosting [52]	Dual Use	gbm, plyr	n.trees, interaction.depth, shrinkage, n.minobsinnode

Table 4.1 describes different type of models used. In this current work, Dual Use models are used which means that they can be applied to both Regression and Classification data. The libraries column in the table describes those libraries which must be installed before running these machine learning models. Tuning Parameters are used to achieve the best accuracy for the models. After this the dataset is reduced to 10 features as described in Table 4.2. The original features are grouped under this 10 features and dataset is condensed. Machine learning models are again applied to this condensed dataset and the results are recorded on these evaluation parameters.

Table 4.2: Reduced dataset with 10 features.

Class No.	Parameter	Description
Class- 1	Gender	It includes the gender information of the participant.
Class- 2	Academics	It contains all the parameters related to academics such as syllabus of the course.
Class- 3	Faculty	It contains all the parameters that are related to faculty.
Class- 4	General	All the general facilities e.g. gym facilities, safety facilities etc. are grouped together and placed in this class.
Class- 5	Research	The research parameters are put together in this class.
Class- 6	Extra-Curricular	It contains the parameters which tells about the extra curricular activities at the campus.
Class- 7	Hostel	All the hostel facilities are included in this class such as hostel living, food and bedding facilities etc.
Class- 8	Technology	It includes all those parameters that are related to technology such as the use of latest version of the softwares in labs.
Class- 9	Guide Availability	It gives us the information, telling whether the guide is allocated to the student or not.
Class- 10	Recreational	All the recreational parameters are included in this class such as meditation classes.

The dataset is again reduced to 4 classes as described in Table 4.3 and Weighted approach [53] is applied to these 4 classes as described. A score related to each participant is computed using this approach. An overall satisfaction score is computed. Cross Validation [54] is applied on top 10 models and top 5 models are chosen based on their accuracy. Ensembling [55] is done for the top five models and the best model can be used for the prediction of the satisfaction score.

Table 4.3: Weights assigned to different classes.

Group	Name	Weight Assigned	Grouping
1	Gender	0.1	It includes Class 1 in it.
2	General	0.2	It includes Class 4, Class 6, Class 7 and Class 10 in it.
3	Academics	0.3	It includes Class 2 and Class 3 in it.
4	Research	0.4	It combines Class 5, Class 8 and Class 9 in it.
		Total weight = 1	

Fig 4.9 shows the classification of reviews into positive, neutral and negative. 1 represents negative, 2 represents neutral and 3 represents positive reviews.

f1	f2	f3	f4	f5	f6	f7	f
2	research	scholars	must	be	given	scholarshi	b
3	i	am	happy	with	the	services	l
2	needs	to	be	improvise	especially	for	g
3	best	college	of	engineering			
1	day	scholars	suffers	a	lot	here	s
1	improve	net	facility				
2	need	to	open	gyms	for	day	s
1	what	is	the	need	to	teach	a
1		there	s	no	gym	in	f
2	otherwise	education	quality	is	good	if	t
1	totally	dissatisfie	worst	internet	facilities	too	r
1	worst	teaching	skills	especially	env	studies	
1	bullshit						
2	compulso	attendanc	why	make	us	independ	a
2	faculty	for	doubt	session	should	be	t
2	the	lecturers	need	to	be	more	e
1	no	gym	facility	at	frc	nd	c
1		there	is	no	gym	faculty	i
1	there	s	alot	we	expect	from	t
3	good						

Fig. 4.9: Classification of reviews.

After classifying the reviews and the feedback, weighted approach is used to compute the satisfaction score. The feedback and reviews are given an equal score of 0.5. Positive and Neutral percentage obtained through classifiers are taken into account for attaining the satisfaction= score. The score can be obtained as given in Eq. (1):

$$Y = \left[w_1 * \frac{Positive_f + Neutral_f}{100} + w_2 * \frac{Positive_r + Neutral_r}{100}\right] * 5 \quad (1)$$

where Y represents the satisfaction score, w_1 represents the weight assigned to the feedback obtained i.e. 0.5 and $Positive_f$ represents the positive percentage obtained and $Neutral_f$ represents the neutral percentage attained through feedback data. w_2 represents the weight assigned to the reviews gathered i.e. 0.5 and $Positive_r$ represents the positive percentage obtained and $Neutral_r$ represents the neutral percentage attained through the reviews.

After applying all the steps mentioned in the above methodology, a complete analysis of the parameters will be obtained. An overall satisfaction score of the University will be attained by considering all the parameters. Detailed analysis of models based on evaluation parameters like correlation, R-Squared, Root Mean Square Error and Accuracy will be

obtained. A satisfaction score based on sentiment analysis of feedback and reviews is also achieved.

Chapter 5
Experimental Results

This section includes the analysis of responses gathered, prediction through machine learning models and sentiment analysis of feedback received and reviews collected through social media. The results are found by conducting a survey within the college premises and feedback is also collected through social media.

5.1. Survey Results

The complete dataset contains 78 features, which focus on multiple parameters such as academics, hostel facilities, co-curricular activities and research.

5.1.1. Analysis of the data captured

There are a few parameters that need to be given special attention by the University for analyzing the data captured during the survey. The parameters that are considered and analyzed are described as follows:

How satisfied are you with the academic procedure that is followed?

	Overall		**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N=36	8.2191	N=10	8.9285	N= 26	7.9754
2	N= 105	23.9726	N= 19	16.9642	N= 86	26.3803
3	N= 171	39.0410	N= 47	41.9642	N= 124	38.0368
4	N= 103	23.5159	N= 28	25	N= 75	23.00613
5	N= 23	5.2511	N= 8	7.1428	N= 15	4.60122
	MEAN = 2.93		**MEAN = 3.044**		**MEAN = 2.898**	

How satisfied are you with out of class experience?

	Overall		**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 34	7.7625	N= 12	10.7142	N= 22	6.7484
2	N= 68	15.5251	N= 16	14.2857	N= 52	15.9509
3	N= 134	30.5936	N= 34	30.3571	N= 100	30.6748
4	N= 142	32.42	N= 38	33.9285	N= 104	31.9018
5	N= 60	13.6986	N= 12	10.7142	N= 48	14.7239
	MEAN = 3.287		**MEAN = 3.196**		**MEAN = 3.319**	

How satisfied are you with your informal participation out of class?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 36	8.2191	N= 7	6.25	N= 29	8.8957
2	N= 58	13.2420	N= 14	12.5	N= 44	13.4969
3	N= 158	36.0730	N= 45	40.1785	N= 113	34.6625
4	N= 139	31.7351	N= 36	32.1468	N= 103	31.5950
5	N= 47	10.7305	N= 10	8.9285	N= 37	11.3496
	MEAN = 3.235		**MEAN = 3.25**		**MEAN = 3.230**	

How satisfied are you with the fees structure of your college?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 219	50	N= 57	50.892	N= 162	49.6362
2	N= 123	28.082	N= 27	24.1071	N= 96	29.4478
3	N= 70	15.981	N= 23	20.5357	N= 47	14.4171
4	N= 18	4.1095	N= 4	3.5714	N= 14	4.2944
5	N= 8	1.8264	N= 1	0.8928	N= 7	2.1472
	MEAN = 1.796		**MEAN = 1.794**		**MEAN = 1.797**	

How satisfied are you with the Safety and security at your campus?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 25	5.7077	N= 3	2.6785	N= 22	6.7484
2	N= 44	10.0456	N= 14	12.5	N= 30	9.2024
3	N= 81	18.4931	N= 26	23.214	N= 55	16.8711
4	N= 174	39.7260	N= 47	41.964	N= 127	38.9570
5	N= 114	26.027	N= 22	19.642	N= 92	28.2208
	MEAN = 3.703		**MEAN = 3.633**		**MEAN = 3.726**	

How satisfied are you with the syllabus structured for the course?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 62	14.1552	N= 10	8.9285	N= 52	15.9509
2	N= 83	18.9497	N= 18	16.0714	N= 65	19.9386
3	N= 127	28.9954	N= 36	32.1428	N= 91	27.9141
4	N= 140	31.9634	N= 43	38.3928	N= 97	29.7546
5	N= 26	5.9360	N= 5	4.4642	N= 21	6.4417
	MEAN = 2.965		**MEAN = 3.133**		**MEAN = 2.907**	

How satisfied are you with the orientation Program?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 30	6.8493	N= 8	7.1428	N= 22	6.7484
2	N= 47	10.7305	N= 11	9.8214	N= 36	11.0429
3	N= 95	21.6894	N= 28	25	N= 67	20.5521
4	N= 140	31.9634	N= 37	33.0357	N= 103	31.5950
5	N= 126	28.7671	N= 28	25	N= 98	30.0613
	MEAN = 3.650		**MEAN = 3.589**		**MEAN = 3.671**	

How satisfied are you with the quality of teaching?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 60	13.6986	N= 12	10.7142	N= 48	14.7239
2	N= 87	19.8630	N= 20	17.8571	N= 67	20.5521
3	N= 184	42.0091	N= 42	37.5	N= 142	43.5582
4	N= 98	22.3744	N= 35	31.25	N= 63	19.3251
5	N= 9	2.0547	N= 3	2.6785	N= 6	1.8404
	MEAN = 2.792		**MEAN = 2.973**		**MEAN = 2.730**	

How satisfied are you with classroom/lab facility?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 42	9.5890	N= 8	7.1428	N= 34	10.4294
2	N= 51	11.6438	N= 12	10.7142	N= 39	11.9631
3	N= 127	28.9954	N= 29	25.8928	N= 98	30.0613
4	N= 157	35.8447	N= 46	41.0714	N= 111	34.0490
5	N= 61	13.9269	N= 17	15.7815	N= 44	13.4969
	MEAN = 3.328		**MEAN = 3.464**		**MEAN = 3.282**	

How satisfied are you with the schedule of your classes?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 97	22.1461	N= 21	18.75	N= 76	23.3128
2	N= 79	18.0365	N= 21	18.75	N= 58	17.7914
3	N= 118	26.9406	N= 35	31.25	N= 83	25.4601
4	N= 105	23.9726	N= 28	25	N= 77	23.6196
5	N= 39	8.9041	N= 7	6.25	N= 32	9.8159
	MEAN = 2.794		**MEAN = 2.812**		**MEAN = 2.788**	

How satisfied are you with extent to which faculty follows methods like presentations?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 32	7.3059	N= 6	5.3571	N= 26	7.9754
2	N= 85	19.4063	N= 21	18.75	N= 64	19.6319
3	N= 175	39.9543	N= 41	36.6071	N= 134	41.1042
4	N= 120	27.3972	N= 39	34.8214	N= 81	24.8466
5	N= 26	5.9360	N= 5	4.4642	N= 21	6.4417
	MEAN = 3.052		**MEAN = 3.142**		**MEAN = 3.021**	

How satisfied are you opportunity to participate in an independent research project with faculty member?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 109	24.8858	N= 21	18.75	N= 88	26.9938

2	N= 91	20.7762	N= 17	15.1785	N= 74	22.6993
3	N= 148	33.7899	N= 44	39.2857	N= 104	31.9018
4	N= 62	14.1552	N= 22	19.6428	N= 40	12.2699
5	N= 28	6.3926	N= 8	7.1428	N= 20	6.1349
	MEAN = 2.563		**MEAN = 2.812**		**MEAN = 2.478**	

How satisfied are you with the opportunity to participate in a study program?

Overall			**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 134	30.5936	N= 28	25	N= 106	32.5153
2	N= 83	18.9497	N= 22	19.6428	N= 61	18.7116
3	N= 147	33.5616	N= 34	30.3571	N= 113	34.6625
4	N= 55	12.5570	N= 21	18.75	N= 34	10.4294
5	N= 19	4.3378	N= 7	6.25	N= 12	3.6809
	MEAN = 2.410		**MEAN = 2.616**		**MEAN = 2.340**	

How satisfied are you approach with which the college authorities deal with the grievances of the students?

Overall			**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 64	14.6118	N= 10	8.9285	N= 54	16.5644
2	N= 83	18.9497	N= 22	19.6428	N= 61	18.7116
3	N= 177	40.4109	N= 44	39.2857	N= 133	40.7975
4	N= 86	19.6347	N= 31	27.6785	N= 55	16.8711
5	N= 28	6.3926	N= 5	4.4642	N= 23	7.0552
	MEAN = 2.842		**MEAN = 2.991**		**MEAN = 2.791**	

How satisfied are you with opportunity to participate in internships in companies or student teaching experiences?

Overall			**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 59	13.4703	N= 13	11.6071	N= 46	14.1104
2	N= 61	13.9269	N= 10	8.9285	N= 51	15.6441
3	N= 173	39.4977	N= 43	38.3928	N= 130	39.8773
4	N= 109	24.8858	N= 35	31.25	N= 74	22.6993
5	N= 36	8.2191	N= 11	9.8214	N= 25	7.6687
	MEAN = 3.004		**MEAN = 3.1870**		**MEAN = 2.9314**	

How satisfied are you with the availability of courses to make progress towards degree?

Overall			**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 51	11.6438	N= 7	6.25	N= 44	13.4969
2	N= 67	15.2968	N= 16	14.2857	N= 51	15.6441
3	N= 152	34.7031	N= 39	34.8214	N= 113	34.6225
4	N= 131	29.9086	N= 42	37.5	N= 89	27.3006
5	N= 37	8.4474	N= 8	7.1428	N= 29	8.8957
	MEAN = 3.082		**MEAN = 3.25**		**MEAN = 3.024**	

How satisfied are you with the use of technology in your class?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 30	6.8493	N= 3	2.6785	N= 27	8.2822
2	N= 65	14.8401	N= 18	16.0714	N= 47	14.4171
3	N= 137	31.2875	N= 35	31.25	N= 102	31.2833
4	N= 154	35.1598	N= 42	37.5	N= 112	34.3558
5	N= 52	11.8721	N= 14	12.5	N= 38	11.6564
	MEAN = 3.303		**MEAN = 3.410**		**MEAN = 3.266**	

How satisfied are you with the accessibility of your teachers?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 26	5.9360	N= 4	3.5714	N= 22	6.7484
2	N= 39	8.9041	N= 5	4.4642	N= 34	10.4294
3	N= 123	28.0821	N= 35	31.25	N= 88	26.9938
4	N= 157	35.8447	N= 41	36.6071	N= 116	35.5828
5	N= 93	21.2328	N= 27	24.1071	N= 66	20.2453
	MEAN = 3.575		**MEAN = 3.732**		**MEAN = 3.521**	

How satisfied are you with the overall quality of the examination procedure?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 36	8.2191	N= 8	7.1428	N= 28	8.5889
2	N= 63	14.3835	N= 11	9.8214	N= 52	15.9509
3	N= 140	31.9634	N= 33	29.4642	N= 107	32.8220
4	N= 147	33.5616	N= 48	42.8571	N= 99	30.3680
5	N= 52	11.8721	N= 12	10.7142	N= 40	12.2699
	MEAN = 3.264		**MEAN = 3.401**		**MEAN = 3.217**	

How satisfied are you with the placement drives in campus?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 25	5.7077	N= 4	3.5714	N= 21	6.4417
2	N= 41	9.3607	N= 13	11.6071	N= 28	8.5889
3	N= 164	37.4429	N= 35	31.25	N= 129	39.5705
4	N= 149	34.0182	N= 37	33.0357	N= 112	34.3558
5	N= 59	13.4703	N= 23	20.5357	N= 36	11.0429
	MEAN = 3.406		**MEAN = 3.553**		**MEAN = 3.349**	

How satisfied are you with the campus bookstore?

	Overall		Day Scholar		Hosteller	
	N= 438	%	N=112	%	N= 326	%
1	N= 48	10.9589	N= 12	10.7142	N= 36	11.0429
2	N= 76	17.3515	N= 15	13.3928	N= 61	18.7116
3	N= 143	32.6484	N= 38	33.9285	N= 105	32.2085
4	N= 118	26.9406	N= 28	25	N= 90	27.6073

5	N= 53	12.1004	N= 19	16.9642	N= 34	10.4294
	MEAN = 3.118		**MEAN = 3.241**		**MEAN = 3.076**	

How satisfied are you with the student financial aid services?

	Overall		**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 81	18.4931	N= 20	17.8571	N= 61	18.7116
2	N= 91	20.7762	N= 17	15.1785	N= 74	22.6993
3	N= 187	42.6940	N= 49	43.75	N= 138	42.3312
4	N= 62	14.1552	N= 22	19.6428	N= 40	12.2699
5	N= 17	3.8812	N= 4	3.5714	N= 13	3.9877
	MEAN = 2.641		**MEAN = 2.758**		**MEAN = 2.601**	

How satisfied are you with the campus library services?

	Overall		**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 52	11.8721	N= 14	12.5	N= 38	11.6564
2	N= 57	13.0316	N= 15	13.3928	N= 42	12.8834
3	N= 117	26.7123	N= 30	26.7857	N= 87	26.6871
4	N= 143	32.6484	N= 38	33.9285	N= 105	32.2085
5	N= 69	15.7534	N= 15	13.3928	N= 54	16.5644
	MEAN = 3.273		**MEAN = 3.223**		**MEAN = 3.291**	

How satisfied are you with the food services in the campus?

	Overall		**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 40	9.1324	N= 8	7.1428	N= 32	9.8159
2	N= 86	19.6347	N= 17	15.1785	N= 69	21.1656
3	N= 139	31.7351	N= 41	36.6071	N= 98	30.0613
4	N= 129	29.4520	N= 35	31.25	N= 94	28.8343
5	N= 44	10.0456	N= 11	9.8214	N= 33	10.1226
	MEAN = 3.116		**MEAN = 3.214**		**MEAN = 3.082**	

How satisfied are you with the campus resources for students?

	Overall		**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 89	20.3196	N= 16	14.2857	N= 73	22.3926
2	N= 77	17.5799	N= 17	15.1785	N= 60	18.4049
3	N= 156	35.6164	N= 42	37.5	N= 114	34.9693
4	N= 84	19.1780	N= 27	24.1071	N= 57	17.4846
5	N= 32	7.3059	N= 10	8.9285	N= 22	6.7484
	MEAN = 2.755		**MEAN = 2.982**		**MEAN = 2.677**	

How satisfied are you with the internet facility at your campus?

	Overall		**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 118	26.9406	N= 21	18.75	N= 97	29.7546
2	N= 91	20.7762	N= 18	16.0714	N= 73	22.3926
3	N= 116	26.4840	N= 34	30.3571	N= 82	25.1533

4	N= 81	18.4931	N= 25	22.3214	N= 56	17.1799
5	N= 32	7.3059	N= 14	12.5	N= 18	5.5214
	MEAN = 2.584		**MEAN = 3.75**		**MEAN = 2.463**	

How satisfied are you with the medical services at your campus?

Overall			**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 41	9.3607	N= 4	3.5714	N= 37	11.3496
2	N= 70	15.9817	N= 17	15.1785	N= 53	16.2576
3	N= 154	35.1598	N= 46	41.0714	N= 108	33.1288
4	N= 117	26.7123	N= 30	26.7857	N= 87	26.6871
5	N= 56	12.7853	N= 15	13.3928	N= 41	12.5766
	MEAN = 3.175		**MEAN = 3.312**		**MEAN = 3.128**	

How satisfied are you with the recreational facilities?

Overall			**Day Scholar**		**Hosteller**	
	N= 438	%	N=112	%	N= 326	%
1	N= 38	8.6757	N= 8	7.1428	N= 30	9.2024
2	N= 65	14.8401	N= 11	9.8214	N= 54	16.5644
3	N= 166	37.8995	N= 43	38.3928	N= 123	37.73
4	N= 112	25.5707	N= 30	26.7857	N= 82	25.1533
5	N= 57	13.0316	N= 20	17.8571	N= 37	11.3496
	MEAN = 3.194		**MEAN = 3.383**		**MEAN = 3.128**	

The results depict that the students are the University are highly satisfied with out of class experiences, with the safety and security at the campus, with the orientation program and classroom/lab facility. The opportunities to participate in internship program, to access teachers outside class, technology used in the campus and the overall examination procedures helped to raise their satisfaction levels. While comparing between hostellers and day scholars, it is found that the hostellers are more satisfied with the safety and out of class experiences as compared to that of day scholars. Participants are moderately satisfied with academic procedure, informal participation out of the students outside class, the structure of the syllabus, the quality of teaching, schedule of classes, with the college authorities to deal with grievances, campus bookstore, campus library, with the food services of the campus and medical and recreational facilities. While comparing day scholars and hostellers, it is found that the former are more satisfied in services like academic procedure, quality of teaching than the latter. The students at the University are dissatisfied with fees structure of the college and the Internet facility of the college. Day scholars are far more satisfied with the Internet facilities as compared to that of hostellers.

5.1.2. Prediction through models

This section details about the various machine-learning models employed over the dataset with 78 features, dataset with 10 features and dataset with 4 features. The training and testing ratios are changed to check the performance of the models over the dataset. Table 5.1 to Table 5.5 concludes about the results of various machine learning models over the dataset with 78 features. The dataset is evaluated by using various parameters i.e. Correlation [56], R- Squared [57], Root Mean Square Error [58] and Accuracy [59]. Table 5.1 illustrates about the results, where training ratio is kept to be 30% and the rest is used as testing data.

Table 5.1: Training data is 30% on dataset with 78 features.

Model No	**Model Name**	**r**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.6461	0.41745	0.2	85.99
2	Bagged MARS using gCV Pruning	0.6759	0.45684	0.21	87.95
3	Boosted Generalized Linear Model	0.687	0.473	0.23	85.34
4	Boosted Linear Model	0.3945	0.15563	0.21	83.06
5	CART	0.452	0.2043	0.16	85.34
6	Conditional Inference Random Forest	0.6859	0.47046	0.17	85.34
7	Conditional Inference Tree	0.5167	0.26698	0.24	85.34
8	Extreme Learning Machine	0.263	0.0693	0.44	77.52
9	Gaussian Process with Polynomial Kernel	0.6851	0.46936	0.18	87.62
10	Generalized Additive Model using Splines	0.4839	0.23416	0.33	75.9
11	Glmnet	0.6992	0.48888	0.23	87.3
12	K-Nearest Neighbour	0.5685	0.32319	0.18	85.34
13	Multilayer Perceptron	0.4313	0.18602	0.23	80.13
14	Multivariate Adaptive Regression Spline	0.463	0.214	0.26	82.74
15	Parallel Random Forest	0.7253	0.52606	0.19	85.99
16	part DSA	0.4224	0.20081	0.2	82.74
17	Partial Least Squares	0.7083	0.50169	0.23	**89.58**
18	Boosted Tree	0.647	0.419	0.2	86.32
19	Self Organizing Map	0.562	0.31584	0.13	87.3
20	Stochastic Gradient Boosting	0.5502	0.30272	0.12	87.95

Table 5.1 concludes that the model Partial Least Squares depicts the highest accuracy i.e. 89.58 and Generalized Additive Model using Splines is portraying the least i.e. 75.90. Table 5.2 represents about the results, where training ratio is kept to be 40% and the rest is used as testing data.

Table 5.2: Dataset with 78 features, training data be 40%

Model No	**Model Name**	**r**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.6368	0.40551	0.2	84.41
2	Bagged MARS using gCV Pruning	0.6582	0.43323	0.22	86.31
3	Boosted Generalized Linear Model	0.671	0.451	0.24	86.69
4	Boosted Linear Model	0.4544	0.20648	0.23	82.13
5	CART	0.4263	0.18173	0.17	84.41
6	Conditional Inference Random Forest	0.6435	0.41409	0.17	84.79
7	Conditional Inference Tree	0.4989	0.2489	0.2	84.41
8	Extreme Learning Machine	0.401	0.161	0.69	46.77
9	Gaussian Process with Polynomial Kernel	0.7129	0.50823	0.17	**88.59**
10	Generalized Additive Model using Splines	0.5632	0.31719	0.3	80.61
11	Glmnet	0.6777	0.45928	0.23	85.93
12	K-Nearest Neighbour	0.6047	0.36566	0.18	85.55
13	Multilayer Perceptron	0.35	0.1225	0.18	83.65
14	Multivariate Adaptive Regression Spline	0.426	0.182	0.17	84.41
15	Parallel Random Forest	0.6699	0.44877	0.2	85.55
16	part DSA	0.4775	0.22801	0.2	84.41
17	Partial Least Squares	0.7308	0.53407	0.23	87.83
18	Boosted Tree	0.625	0.391	0.21	85.55
19	Self Organizing Map	0.5415	0.29322	0.16	84.41
20	Stochastic Gradient Boosting	0.5387	0.2902	0.13	87.07

Table 5.2 concludes that Gaussian Process with Polynomial Kernel model depicts the highest accuracy i.e. 88.59 and the least is depicted by Extreme Learning Machine model i.e. 46.77. Table 5.3 depicts about the results, where training ratio is kept to be 50% and the rest is used as testing data.

Table 5.3: Dataset with 78 features, training data be 50%

Model No	Model Name	r	R	RMSE	Accuracy
1	Bagged CART	0.6425	0.41281	0.19	85
2	Bagged MARS using gCV Pruning	0.6462	0.41757	0.23	84.09
3	Boosted Generalized Linear Model	0.709	0.5030	0.22	84.09
4	Boosted Linear Model	0.5658	0.32013	0.22	81.36
5	CART	0.3906	0.15257	0.18	83.18
6	Conditional Inference Random Forest	0.6092	0.37112	0.17	84.55
7	Conditional Inference Tree	0.4676	0.21865	0.23	83.18
8	Extreme Learning Machine	0.135	0.0181	0.68	49.55
9	Gaussian Process with Polynomial Kernel	0.734	0.53876	0.2	**88.18**
10	Generalized Additive Model using Splines	0.6412	0.41114	0.27	84.09
11	Glmnet	0.6847	0.46881	0.22	82.73
12	K-Nearest Neighbour	0.6506	0.42328	0.19	84.09
13	Multilayer Perceptron	0.4727	0.22345	0.2	85
14	Multivariate Adaptive Regression Spline	0.391	0.153	0.18	83.18
15	Parallel Random Forest	0.71	0.5041	0.18	85.91
16	part DSA	0.3906	0.15257	0.18	83.18
17	Partial Least Squares	0.7402	0.5479	0.23	87.73
18	Boosted Tree	0.644	0.415	0.2	84.09
19	Self Organizing Map	0.6136	0.3765	0.12	87.73
20	Stochastic Gradient Boosting	0.529	0.27984	0.14	86.36

Table 5.3 concludes the highest accuracy is depicted by model Gaussian Process with Polynomial Kernel i.e. 88.18 and the lowest is depicted by Extreme Learning Machine model i.e. 49.55. The average accuracy shown by the models is between 80% to 85%. Table 5.4 represents about the results, where training ratio is kept to be 60% and the rest is used as testing data.

Table 5.4: Dataset with 78 features, training data be 60%

Model No	Model Name	r	R	RMSE	Accuracy
1	Bagged CART	0.6454	0.41654	0.19	82.95
2	Bagged MARS using gCV Pruning	0.6367	0.40539	0.24	82.39

3	Boosted Generalized Linear Model	0.716	0.512	0.23	82.95
4	Boosted Linear Model	0.568	0.32262	0.23	80.68
5	CART	0.3251	0.10569	0.19	81.25
6	Conditional Inference Random Forest	0.5056	0.25563	0.18	81.82
7	Conditional Inference Tree	0.169	0.4111	0.25	81.25
8	Extreme Learning Machine	0.171	0.0291	0.36	73.86
9	Gaussian Process with Polynomial Kernel	0.7575	0.57381	0.18	88.64
10	Generalized Additive Model using Splines	0.6935	0.48094	0.25	87.5
11	Glmnet	0.7356	0.54111	0.22	84.66
12	K-Nearest Neighbour	0.6683	0.44662	0.19	85.8
13	Multilayer Perceptron	0.4855	0.23571	0.19	84.09
14	Multivariate Adaptive Regression Spline	0.588	0.345	0.23	82.39
15	Parallel Random Forest	0.6974	0.48637	0.19	84.66
16	part DSA	0.3251	0.10569	0.19	81.25
17	Partial Least Squares	0.7544	0.56912	0.23	**89.2**
18	Boosted Tree	0.621	0.385	0.2	81.82
19	Self Organizing Map	0.6387	0.40794	0.12	87.5
20	Stochastic Gradient Boosting	0.5612	0.31495	0.14	85.8

Table 5.4 concludes that the highest accuracy is depicted by Partial Least Squares model i.e. 89.2 and the lowest accuracy is depicted by Extreme Learning Machine model i.e. 73.86. The average accuracy portrayed by models is in the range of 80 to 85. Table 5.5 concludes about the results, where training ratio is kept to be 70% and the rest is used as testing data.

Table 5.5: Dataset with 78 features, training data be 70%

Model No	**Model Name**	**r**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.689	0.47472	0.18	86.36
2	Bagged MARS using gCV Pruning	0.7016	0.49224	0.21	83.33
3	Boosted Generalized Linear Model	0.771	0.595	0.22	87.12
4	Boosted Linear Model	0.3706	0.13734	0.2	81.06
5	CART	0.5093	0.25939	0.22	82.58
6	Conditional Inference Random Forest	0.612	0.37454	0.17	84.09
7	Conditional Inference Tree	0.5076	0.25766	0.25	81.06

8	Extreme Learning Machine	0.141	0.0198	0.24	79.55
9	Gaussian Process with Polynomial Kernel	0.7516	0.5649	0.18	**88.64**
10	Generalized Additive Model using Splines	0.7201	0.51854	0.25	**88.64**
11	Glmnet	0.7668	0.58798	0.22	85.61
12	K-Nearest Neighbour	0.6563	0.43073	0.2	84.09
13	Multilayer Perceptron	0.5253	0.27594	0.2	80.3
14	Multivariate Adaptive Regression Spline	0.371	0.137	0.2	81.06
15	Parallel Random Forest	0.7438	0.55324	0.19	86.36
16	part DSA	0.3706	0.13734	0.2	81.06
17	Partial Least Squares	0.7627	0.58171	0.24	**88.64**
18	Boosted Tree	0.685	0.47	0.21	83.33
19	Self Organizing Map	0.6439	0.41461	0.16	84.07
20	Stochastic Gradient Boosting	0.5076	0.25786	0.16	84.09

Table 5.5 portrays that models Partial Least Squares, Gaussian Process with Polynomial Kernel and Generalized Additive Model using Splines, show the highest accuracy of 88.64. Extreme Learning Machine shows the least accuracy of 79.55. Fig 5.1 compares the correlation evaluation parameter for different training and testing ratios on dataset with 78 features.

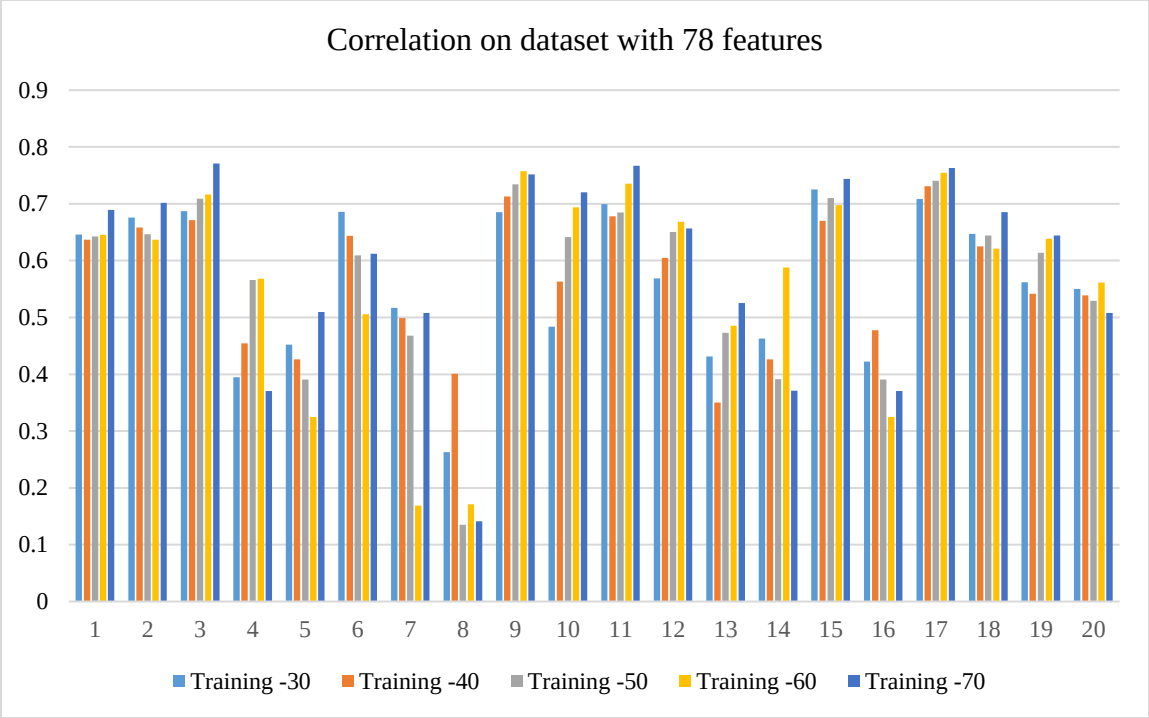


Fig.5.1: Comparison of correlation for different machine learning models.

As concluded in Fig. 5.1, model Gaussian Process with Polynomial Kernel (Model 9) and Partial Least Squares (Model 17) has shown a steady increase in the value of correlation. Extreme Learning Machine (Model 8) shows the least value of correlation and then there is a sudden spike in the value of correlation at training of 40%. partDSA has shown lower values of correlation as compared to other models. Fig 5.2 compares the R-Squared evaluation parameter for different training and testing ratios on dataset with 78 features.

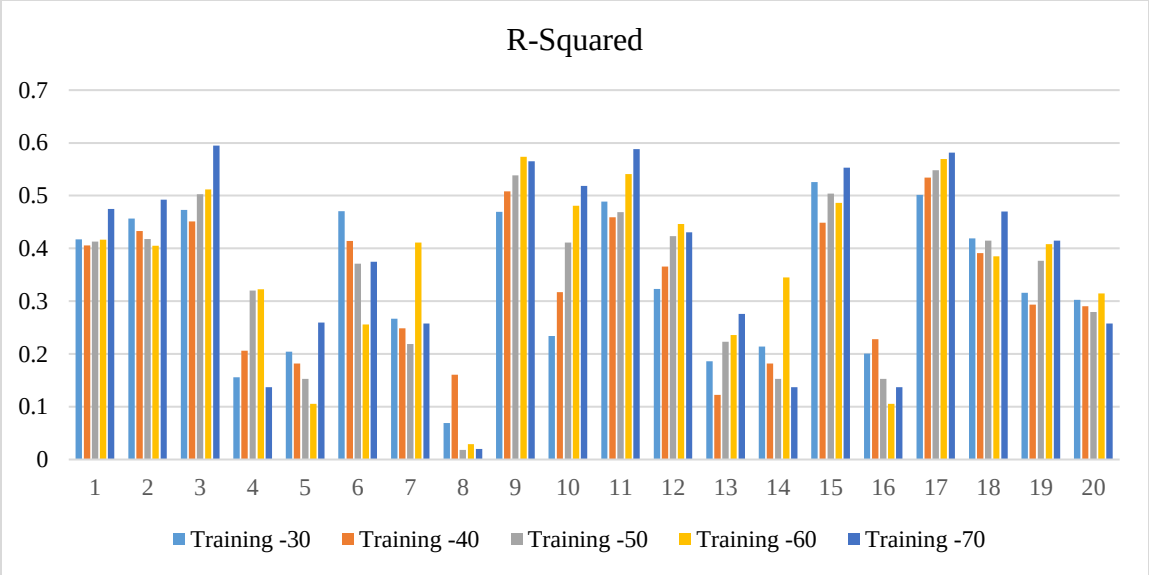


Fig.5.2: Comparison of R-Squared for different machine learning models.

Fig 5.2 depicts the results for the evaluation parameter R-Squared for different models at different training ratios. It shows that multiple models like Boosted Generalized Linear Model (Model 3), Gaussian Process with Polynomial Kernel (Model 9), Glmnet (Model 11), Parallel Random Forest (Model 15) and Partial Least Squares (Model 17) have shown high values of R-Squared. Boosted Linear Model (Model 4), CART (Model 5), Extreme Learning Machine (Model 8), Multilayer Perceptron (Model 13), Multivariate Adaptive Regression Spline (Model 14) and partDSA (Model 16) shows the least values of R-squared. Extreme Learning Machine shows the lowest values at all training ratios. Fig 5.3 compares the Root Mean Square Error evaluation parameter for different training and testing ratios on dataset with 78 features.

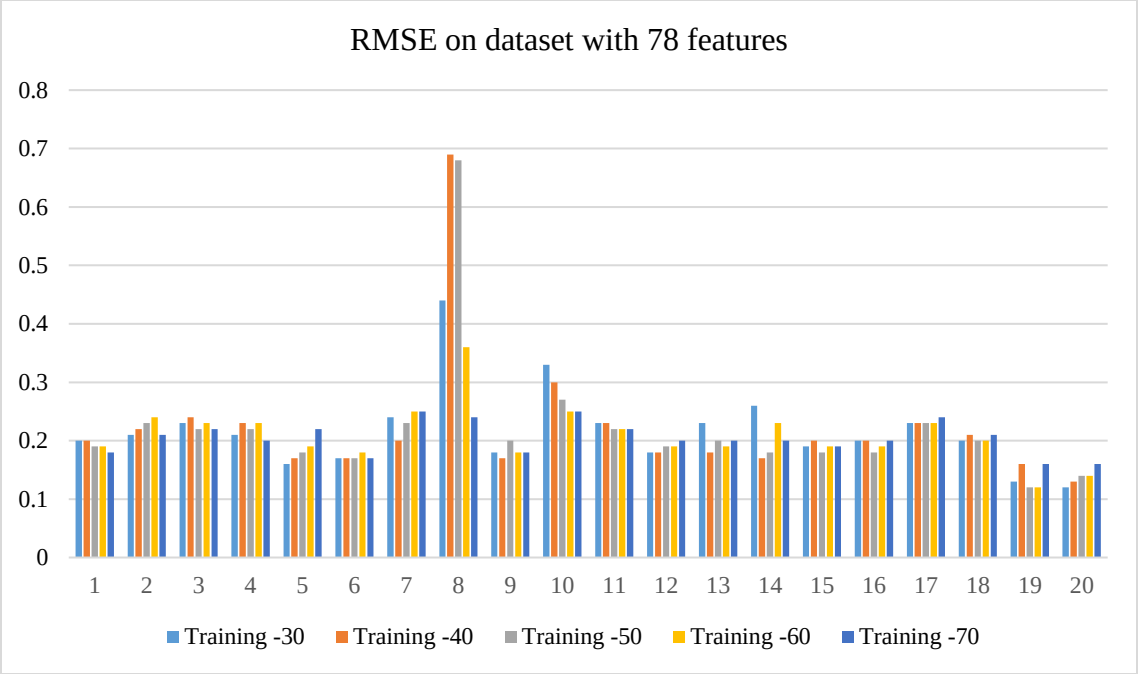


Fig.5.3: Comparison of RMSE for different machine learning models.

Extreme Learning Machine (Model 8) has shown higher values of Root Mean Square Error, which concludes the poor performance of the model at various training ratios as depicted in Fig 5.3. Fig 5.4 compares the Accuracy evaluation parameter for different training and testing ratios on dataset with 78 features.

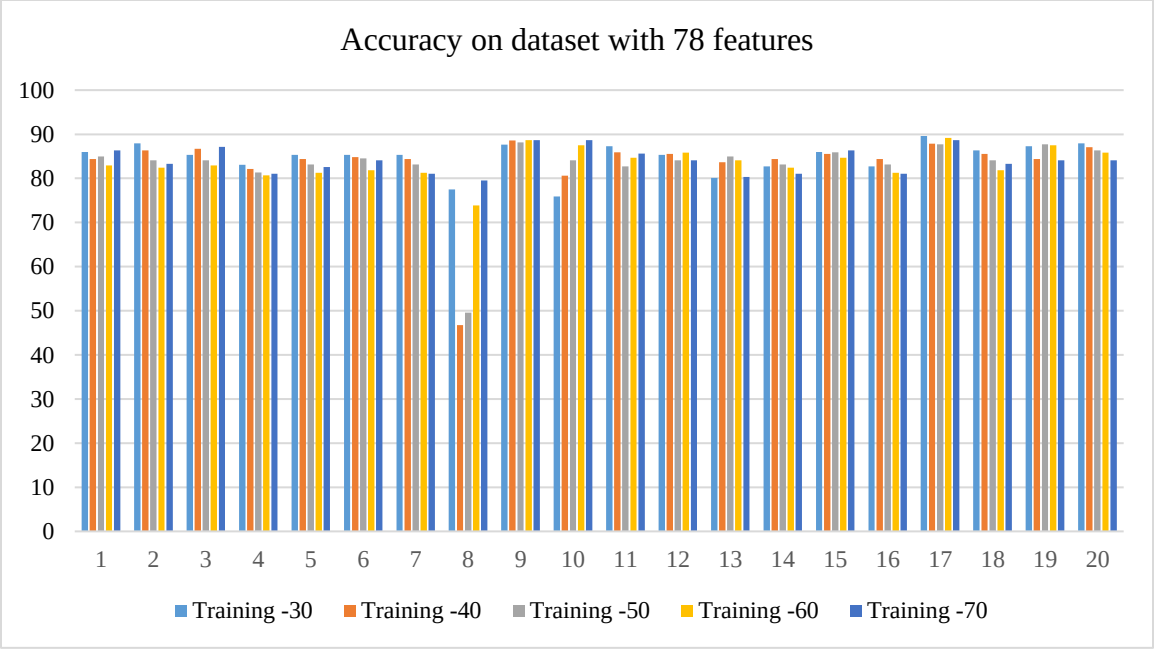


Fig.5.4: Comparison of accuracy for different machine learning models.

The results depicts that the model performances varies on changing the training and testing ratios while considering 78 parameters. At 30% training ratio, the highest accuracy of 87.75% is depicted by Bagged Mars using gCV Pruning and Stochastic Gradient Boosting. The lowest accuracy of 75.9% is shown by Generalized Additive Model using Splines. At 40%, 50%, 60% and 70% training ratio, the highest accuracy is being depicted by Gaussian Process with Polynomial Kernel and the lowest being depicted by Extreme Learning Machine.

Now, the original cleansed dataset is reduced to dataset with 10 features. The complete prediction methodology is again followed on this reduced dataset. Table 5.6 to Table 5.10 concludes about the results of various machine learning models over the reduced dataset with 10 features. Table 5.6 demonstrates about the results, where training ratio is kept to be 30% and the rest is used as testing data.

Table 5.6: Dataset with 10 features, training data be 30%

Model. No	**Model Name**	**R**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.7304	0.53348	0.14	88.69
2	Bagged MARS using gCV Pruning	0.7069	0.49971	0.17	89.25

3	Boosted Generalized Linear Model	0.733	0.537	0.22	89.25
4	Boosted Linear Model	0.3706	0.13781	0.22	81.43
5	CART	0.6622	0.43851	0.14	87.62
6	Conditional Inference Random Forest	0.6659	0.44342	0.13	87.3
7	Conditional Inference Tree	0.7197	0.51797	0.13	**92.51**
8	Extreme Learning Machine	0.326	0.106	0.2	82.74
9	Gaussian Process with Polynomial Kernel	0.705	0.49702	0.2	86.97
10	Generalized Additive Model using Splines	0.7567	0.57249	0.2	89.97
11	Glmnet	0.7361	0.54184	0.22	90.23
12	K-Nearest Neighbour	0.7197	0.51797	0.17	85.99
13	Multilayer Perceptron	0.6523	0.4255	0.23	87.3
14	Multivariate Adaptive Regression Spline	0.723	0.523	0.16	88.6
15	Parallel Random Forest	0.8251	0.68079	0.14	91.53
16	part DSA	0.6622	0.43851	0.14	87.62
17	Partial Least Squares	0.6891	0.47486	0.23	89.9
18	Boosted Tree	0.703	0.494	0.16	87.95
19	Self Organizing Map	0.6507	0.42341	0.13	87.3
20	Stochastic Gradient Boosting	0.6934	0.4808	0.09	90.55

Table 5.6 concludes that the highest accuracy of 92.51% is depicted by Conditional Inference Tree model and the lowest accuracy of 81.43% is shown by Boosted Linear Model. Table 5.7 illustrates about the results, where training ratio is kept to be 40% and the rest is used as testing data.

Table 5.7: Dataset with 10 features, training data be 40%

Model. No	**Model Name**	**R**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.7214	0.52042	0.14	86.31
2	Bagged MARS using gCV Pruning	0.71	0.5041	0.17	88.59
3	Boosted Generalized Linear Model	0.73	0.532	0.22	86.31
4	Boosted Linear Model	0.3708	0.13749	0.21	81.75
5	CART	0.6543	0.42811	0.15	89.73
6	Conditional Inference Random Forest	0.5954	0.3545	0.15	84.03

7	Conditional Inference Tree	0.7254	0.52621	0.15	87.83
8	Extreme Learning Machine	0.138	0.0189	0.21	82.51
9	Gaussian Process with Polynomial Kernel	0.7755	0.6014	0.2	88.97
10	Generalized Additive Model using Splines	0.7566	0.57244	0.17	90.11
11	Glmnet	0.7505	0.56325	0.22	87.07
12	K-Nearest Neighbour	0.7012	0.49168	0.17	85.55
13	Multilayer Perceptron	0.5809	0.33744	0.33	82.89
14	Multivariate Adaptive Regression Spline	0.682	0.466	0.18	88.59
15	Parallel Random Forest	0.7935	0.62964	0.12	89.35
16	part DSA	0.6543	0.42811	0.15	86.69
17	Partial Least Squares	0.7334	0.53788	0.22	88.97
18	Boosted Tree	0.699	0.488	0.17	87.83
19	Self Organizing Map	0.6672	0.44516	0.12	87.89
20	Stochastic Gradient Boosting	0.7303	0.53334	0.1	**90.49**

The highest accuracy of 90.49% at training ratio 40% on dataset with 10 features is depicted by Stochastic Gradient Boosting model. Boosted Linear Model records the lowest accuracy of 81.75%. Table 5.8 illustrates about the results, where training ratio is kept to be 50% and the rest is used as testing data.

Table 5.8: Dataset with 10 features, training data be 50%

Model. No	**Model Name**	**r**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.7781	0.60544	0.13	89.09
2	Bagged MARS using gCV Pruning	0.782	0.61152	0.16	**90.91**
3	Boosted Generalized Linear Model	0.739	0.546	0.22	86.36
4	Boosted Linear Model	0.4109	0.16884	0.21	81.36
5	CART	0.6338	0.4017	0.17	87.27
6	Conditional Inference Random Forest	0.6954	0.48358	0.14	86.82
7	Conditional Inference Tree	0.7685	0.59059	0.14	90
8	Extreme Learning Machine	0.4	0.16	0.21	81.82
9	Gaussian Process with Polynomial Kernel	0.7795	0.60762	0.2	90
10	Generalized Additive Model using Splines	0.7657	0.5863	0.19	90

11	Glmnet	0.7358	0.5414	0.22	84.55
12	K-Nearest Neighbour	0.7082	0.50155	0.17	83.64
13	Multilayer Perceptron	0.6614	0.43745	0.2	85
14	Multivariate Adaptive Regression Spline	0.814	0.662	0.17	90.45
15	Parallel Random Forest	0.8457	0.71521	0.11	**90.91**
16	part DSA	0.723	0.52273	0.13	87.27
17	Partial Least Squares	0.7368	0.54287	0.23	90.45
18	Boosted Tree	0.775	0.601	0.15	90
19	Self Organizing Map	0.5978	0.35736	0.13	86.82
20	Stochastic Gradient Boosting	0.7302	0.53319	0.1	90

While keeping the training and testing ratio to be the same, the highest accuracy of 90.91% is recorded for Parallel Random Forest Model and the lowest of 81.36% is recorded for Boosted Linear Model. Table 5.9 depicts about the results, where training ratio is kept to be 60% and the rest is used as testing data.

Table 5.9: Dataset with 10 features, training data be 60%

Model. No	**Model Name**	**r**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.7619	0.58049	0.14	87.5
2	Bagged MARS using gCV Pruning	0.8004	0.64064	0.16	89.77
3	Boosted Generalized Linear Model	0.768	0.59	0.22	91.48
4	Boosted Linear Model	0.3701	0.13764	0.2	81.79
5	CART	0.685	0.46923	0.15	90.34
6	Conditional Inference Random Forest	0.759	0.57608	0.12	90.34
7	Conditional Inference Tree	0.7857	0.61732	0.14	**94.32**
8	Extreme Learning Machine	0.332	0.11	0.23	81.25
9	Gaussian Process with Polynomial Kernel	0.7887	0.62205	0.21	90.91
10	Generalized Additive Model using Splines	0.7556	0.5934	0.18	90.29
11	Glmnet	0.7764	0.6028	0.22	92.05
12	K-Nearest Neighbour	0.6956	0.48346	0.18	84.09
13	Multilayer Perceptron	0.7336	0.53817	0.2	86.93
14	Multivariate Adaptive Regression Spline	0.803	0.645	0.18	89.77
15	Parallel Random Forest	0.8432	0.71099	0.11	92.05
16	part DSA	0.685	0.46923	0.15	86.36
17	Partial Least Squares	0.7721	0.59614	0.22	90.91
18	Boosted Tree	0.809	0.654	0.15	89.2
19	Self Organizing Map	0.6644	0.44143	0.13	86.93

20	Stochastic Gradient Boosting	0.7229	0.52258	0.11	88.64

While testing the models at 60% training data, it is found the highest accuracy of 94.32% is shown by Conditional Inference Tree model and the lowest of 81.25% is shown by Extreme Learning Machine model. Table 5.10 illustrates about the results, where training ratio is kept to be 70% and the rest is used as testing data.

Table 5.10: Dataset with 10 features, training data be 70%

Model. No	**Model Name**	**R**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.7858	0.61748	0.14	88.64
2	Bagged MARS using gCV Pruning	0.8081	0.65303	0.17	90.15
3	Boosted Generalized Linear Model	0.8	0.64	0.21	**91.67**
4	Boosted Linear Model	0.5853	0.34258	0.24	80.3
5	CART	0.6866	0.47142	0.17	90.15
6	Conditional Inference Random Forest	0.7417	0.55012	0.13	89.39
7	Conditional Inference Tree	0.7327	0.53685	0.13	89.39
8	Extreme Learning Machine	0.518	0.268	0.23	81.82
9	Gaussian Process with Polynomial Kernel	0.801	0.6416	0.2	89.39
10	Generalized Additive Model using Splines	0.7784	0.60591	0.2	90.51
11	Glmnet	0.8038	0.64609	0.21	90.51
12	K-Nearest Neighbour	0.7271	0.52867	0.16	85.61
13	Multilayer Perceptron	0.6424	0.41268	0.24	84.09
14	Multivariate Adaptive Regression Spline	0.809	0.654	0.18	89.39
15	Parallel Random Forest	0.8453	0.71453	0.11	90.91
16	part DSA	0.7595	0.57684	0.13	87.88
17	Partial Least Squares	0.7817	0.61105	0.22	89.39
18	Boosted Tree	0.852	0.726	0.14	**91.67**
19	Self Organizing Map	0.6626	0.43904	0.14	85.61
20	Stochastic Gradient Boosting	0.7765	0.60295	0.09	90.91

The highest accuracy of 91.67% is recorded for Boosted Generalized Linear Model and Boosted Tree model. The lowest accuracy of 81.82% is recorded for Extreme Learning Machine model. Fig. 5.5 depicts the comparison between machine learning models over the Correlation evaluation parameter on reduced dataset with 10 features.

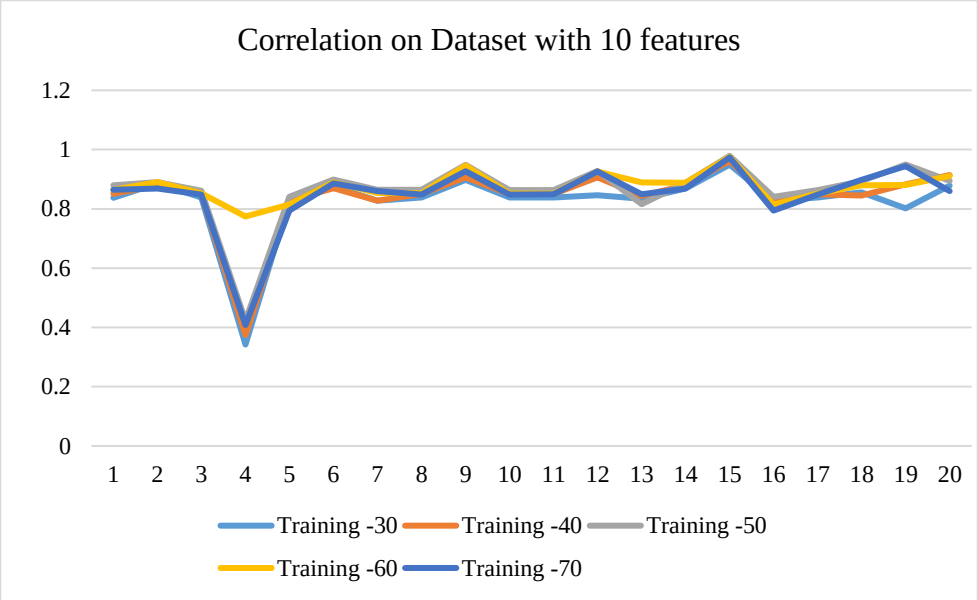


Fig.5.5: Comparison of correlation for different machine learning models.

Boosted Linear Model (Model 4) shows the lowest value of correlation and the highest values is shown by Parallel Random Forest (Model 15) at all training ratios. A sudden increase in the value of correlation is visible for model Self Organizing Map (Model 19) at 70% training ratio while a huge drop is seen at 30% training data for the same model. Fig. 5.6 depicts the comparison between machine learning models over the R-Squared evaluation parameter on reduced dataset with 10 features.

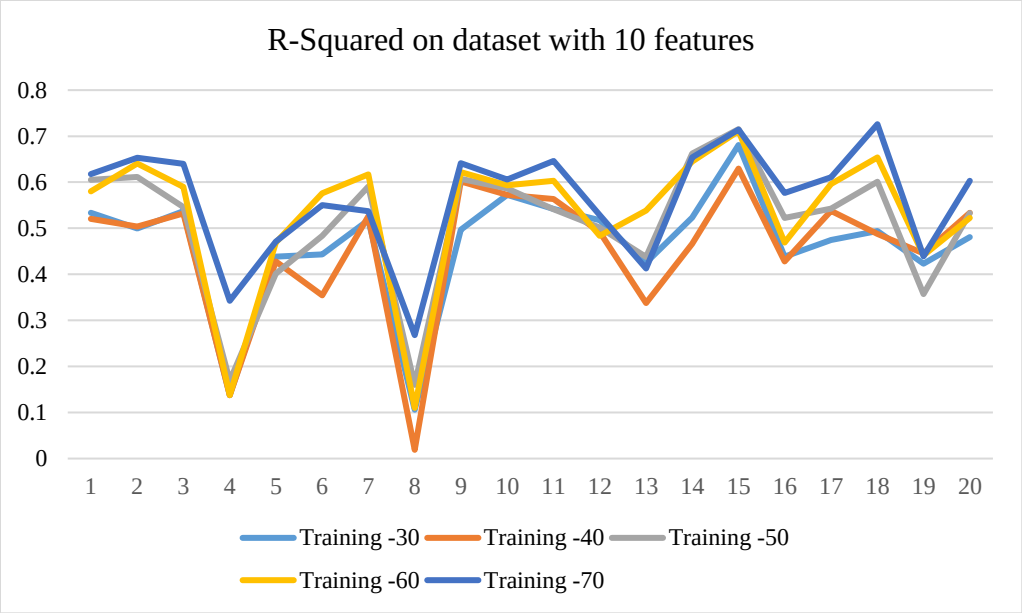


Fig.5.6: Comparison of R-Squared for different machine learning models.

At all training ratios, highest values of R-Squared are shown by Parallel Random Forest model (Model 15) and Boosted Tree model (Model 18). The lowest values for R-Squared are portrayed by Boosted Linear Model (Model 4) and Extreme Learning Machine model (Model 8). Fig. 5.7 depicts the comparison between machine learning models over the Root Mean Square Error evaluation parameter on reduced dataset with 10 features.

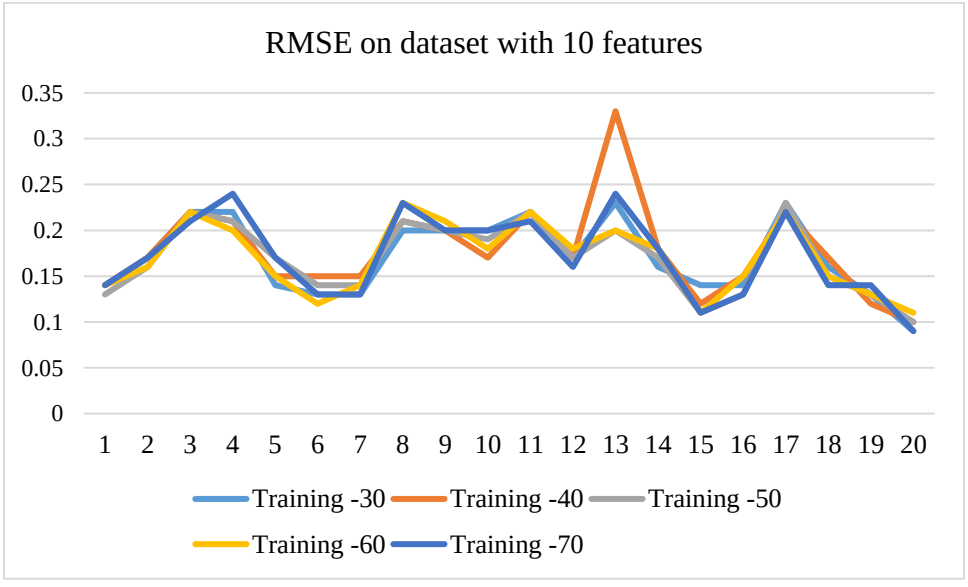


Fig.5.7: Comparison of RMSE for different machine learning models.

Multilayer Perceptron (Model 13) shows the highest value of RMSE at 40% training ratio which means lower accuracy. Fig. 5.8 depicts the comparison between machine learning models over the Accuracy evaluation parameter on reduced dataset with 10 features.

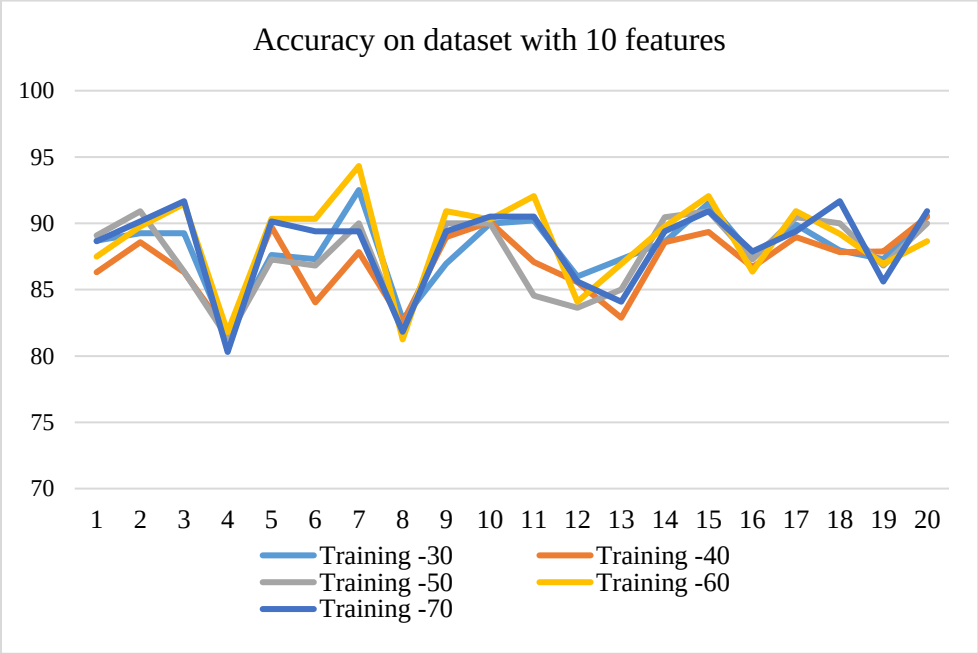


Fig.5.8: Comparison of accuracy for different machine learning models.

For the dataset with 10 features the results are being compared. At 30% training data, the highest accuracy of 92.51% is depicted by Conditional Inference Tree and the lowest of 82.74% is shown by Extreme Learning Machine. At 40% training data, the highest accuracy of 90.11% is depicted by Generalized Additive Model using splines and the lowest of 82.51% is shown by Extreme Learning Machine. At 50% training ratio, the highest accuracy of 90.91% is depicted by Bagged MARS using gCV Pruning and Parallel Random Forest while the lowest of 81.36% is shown by Boosted Linear Model. At 60% training ratio, the highest accuracy of 94.32% is depicted by Conditional Inference Tree and the lowest of 81.25% is shown by Extreme Learning Machine. At 70% training ratio, the highest accuracy of 91.67% is depicted by Boosted Generalized Linear Model and Boosted Tree while the lowest of 80.30% is shown by Boosted Linear Model.

Now, the reduced dataset is further reduced to 4 classes. The complete prediction methodology is again followed on this reduced dataset. Table 5.11 to Table 5.15 concludes

about the results of various machine learning models over the reduced dataset with 4 classes. Table 5.11 elucidates about the results, where training ratio is kept to be 30% and the rest is used as testing data.

Table 5.11: Dataset with 4 features, training data be 30%

Model. No	**Model Name**	**r**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.8373	0.70107	0.11	92.18
2	Bagged MARS using gCV Pruning	0.8864	0.7857	0.12	95.44
3	Boosted Generalized Linear Model	0.838	0.702	0.16	94.14
4	Boosted Linear Model	0.3422	0.1171	0.2	82.74
5	CART	0.8275	0.68476	0.1	92.18
6	Conditional Inference Random Forest	0.8757	0.76685	0.1	92.18
7	Conditional Inference Tree	0.8275	0.68476	0.1	92.18
8	Extreme Learning Machine	0.838	0.703	0.16	94.14
9	Gaussian Process with Polynomial Kernel	0.8976	0.80569	0.1	96.42
10	Generalized Additive Model using Splines	0.8382	0.70258	0.16	94.14
11	Glmnet	0.8383	0.70275	0.16	94.14
12	K-Nearest Neighbour	0.8458	0.71538	0.08	92.51
13	Multilayer Perceptron	0.8339	0.69539	0.13	92.83
14	Multivariate Adaptive Regression Spline	0.869	0.755	0.14	95.14
15	Parallel Random Forest	0.9491	0.90079	0.05	**98.05**
16	part DSA	0.8275	0.68476	0.1	92.18
17	Partial Least Squares	0.8389	0.70375	0.16	94.14
18	Boosted Tree	0.856	0.733	0.13	92.18
19	Self Organizing Map	0.802	0.6432	0.06	93.81
20	Stochastic Gradient Boosting	0.8785	0.77176	0.04	96.09

On dataset with 4 features at 30% training ratio, Parallel Random Forest (Model 15) portrays the highest accuracy of 98.05%. The average accuracy lies between 93% to 95%. Table 5.12 illustrates about the results, where training ratio is kept to be 40% and the rest is used as testing data.

Table 5.12: Dataset with 4 features, training data be 40%

Model. No	Model Name	r	R	RMSE	Accuracy
1	Bagged CART	0.8515	0.72505	0.1	91.25
2	Bagged MARS using gCV Pruning	0.8784	0.77159	0.13	94.68
3	Boosted Generalized Linear Model	0.85	0.723	0.15	93.54
4	Boosted Linear Model	0.3745	0.14025	0.21	81.75
5	CART	0.8269	0.68376	0.11	91.25
6	Conditional Inference Random Forest	0.8702	0.75725	0.09	91.25
7	Conditional Inference Tree	0.8277	0.68509	0.11	91.25
8	Extreme Learning Machine	0.85	0.723	0.15	93.54
9	Gaussian Process with Polynomial Kernel	0.9066	0.82192	0.09	95.82
10	Generalized Additive Model using Splines	0.8501	0.72267	0.15	93.54
11	Glmnet	0.8502	0.72284	0.15	93.54
12	K-Nearest Neighbour	0.906	0.82084	0.06	95.82
13	Multilayer Perceptron	0.8436	0.71166	0.17	92.02
14	Multivariate Adaptive Regression Spline	0.879	0.773	0.13	94.68
15	Parallel Random Forest	0.9619	0.92525	0.05	**98.48**
16	part DSA	0.8269	0.68376	0.11	91.25
17	Partial Least Squares	0.8493	0.72131	0.15	93.54
18	Boosted Tree	0.845	0.714	0.13	91.25
19	Self Organizing Map	0.8823	0.77845	0.05	95.44
20	Stochastic Gradient Boosting	0.914	0.8354	0.03	96.96

The highest accuracy of 98.48 is shown by Parallel Random Forest. Table 5.13 illustrates about the results, where training ratio is kept to be 50% and the rest is used as testing data.

Table 5.13: Dataset with 4 features, training data be 50%

Model. No	Model Name	r	R	RMSE	Accuracy
1	Bagged CART	0.879	0.77264	0.09	95.91
2	Bagged MARS using gCV Pruning	0.8906	0.79317	0.13	95.45
3	Boosted Generalized Linear Model	0.861	0.741	0.15	94.09
4	Boosted Linear Model	0.4228	0.17876	0.22	81.36

5	CART	0.8405	0.70644	0.1	91.36
6	Conditional Inference Random Forest	0.8993	0.80874	0.09	95.45
7	Conditional Inference Tree	0.8644	0.74719	0.09	94.55
8	Extreme Learning Machine	0.864	0.746	0.15	94.09
9	Gaussian Process with Polynomial Kernel	0.9481	0.89889	0.08	98.64
10	Generalized Additive Model using Splines	0.8635	0.74563	0.15	94.09
11	Glmnet	0.8637	0.74598	0.16	98.18
12	K-Nearest Neighbour	0.9286	0.8623	0.05	96.82
13	Multilayer Perceptron	0.8153	0.66471	0.16	89.09
14	Multivariate Adaptive Regression Spline	0.888	0.789	0.14	95.45
15	Parallel Random Forest	0.9788	0.95805	0.04	**100**
16	part DSA	0.8405	0.70644	0.1	91.36
17	Partial Least Squares	0.8629	0.7466	0.15	94.09
18	Boosted Tree	0.893	0.798	0.11	94.55
19	Self Organizing Map	0.9499	0.90231	0.02	98.18
20	Stochastic Gradient Boosting	0.8942	0.79959	0.04	95.91

At 50% training ratio, the highest accuracy of 100% is shown by Parallel Random Forest. Table 5.14 illustrates about the results, where training ratio is kept to be 60% and the rest is used as testing data.

Table 5.14: Dataset with 4 features, training data be 60%

Model. No	**Model Name**	**r**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.8643	0.74701	0.1	95.41
2	Bagged MARS using gCV Pruning	0.8902	0.79246	0.14	95.45
3	Boosted Generalized Linear Model	0.852	0.725	0.16	91.48
4	Boosted Linear Model	0.7743	0.6641	0.2	90.18
5	CART	0.8148	0.6639	0.12	89.77
6	Conditional Inference Random Forest	0.8887	0.78979	0.09	94.89
7	Conditional Inference Tree	0.8538	0.72897	0.09	93.75
8	Extreme Learning Machine	0.852	0.725	0.16	91.48
9	Gaussian Process with Polynomial Kernel	0.9467	0.89624	0.08	98.3
10	Generalized Additive Model using Splines	0.8517	0.72539	0.16	91.46
11	Glmnet	0.8519	0.72573	0.16	91.48
12	K-Nearest Neighbour	0.9258	0.85711	0.05	96.59
13	Multilayer Perceptron	0.8895	0.79121	0.11	92.05
14	Multivariate Adaptive Regression Spline	0.888	0.788	0.14	95.45

15	Parallel Random Forest	0.9756	0.9518	0.04	**99.43**
16	part DSA	0.8148	0.6639	0.12	89.77
17	Partial Least Squares	0.8512	0.72454	0.16	89.77
18	Boosted Tree	0.879	0.773	0.12	96.02
19	Self Organizing Map	0.8806	0.77546	0.05	94.89
20	Stochastic Gradient Boosting	0.9112	0.83029	0.03	96.59

Parallel Random Forest shows the highest accuracy of 99.43% at 60% training ratio. Table 5.15 illustrates about the results, where training ratio is kept to be 70% and the rest is used as testing data.

Table 5.15: Dataset with 4 features, training data be 70%

Model. No	**Model Name**	**r**	**R**	**RMSE**	**Accuracy**
1	Bagged CART	0.8641	0.74667	0.1	92.42
2	Bagged MARS using gCV Pruning	0.8687	0.75464	0.15	94.7
3	Boosted Generalized Linear Model	0.848	0.719	0.16	88.64
4	Boosted Linear Model	0.4079	0.16638	0.24	80.3
5	CART	0.794	0.60344	0.13	87.88
6	Conditional Inference Random Forest	0.8857	0.78446	0.09	93.94
7	Conditional Inference Tree	0.8599	0.73943	0.09	93.94
8	Extreme Learning Machine	0.848	0.72	0.16	88.64
9	Gaussian Process with Polynomial Kernel	0.9267	0.85877	0.08	95.45
10	Generalized Additive Model using Splines	0.8483	0.71961	0.16	88.64
11	Glmnet	0.8488	0.72046	0.16	90.91
12	K-Nearest Neighbour	0.9257	0.85692	0.06	96.21
13	Multilayer Perceptron	0.8496	0.71782	0.2	95.45
14	Multivariate Adaptive Regression Spline	0.869	0.755	0.15	94.7
15	Parallel Random Forest	0.9734	0.94751	0.04	**99.24**
16	part DSA	0.794	0.63044	0.13	87.88
17	Partial Least Squares	0.8478	0.71876	0.16	88.64
18	Boosted Tree	0.898	0.806	0.1	96.21
19	Self Organizing Map	0.9448	0.89265	0.02	97.73
20	Stochastic Gradient Boosting	0.8598	0.73926	0.06	93.94

Parallel Random Forest depicts the highest accuracy of 99.24%. The average accuracy lies between 88% to 93%. The lowest accuracy of 80.3% is shown by Boosted Linear Model. Fig. 5.9 depicts the comparison between machine learning models over the Correlation evalauation parameter on reduced dataset with 4 classes.

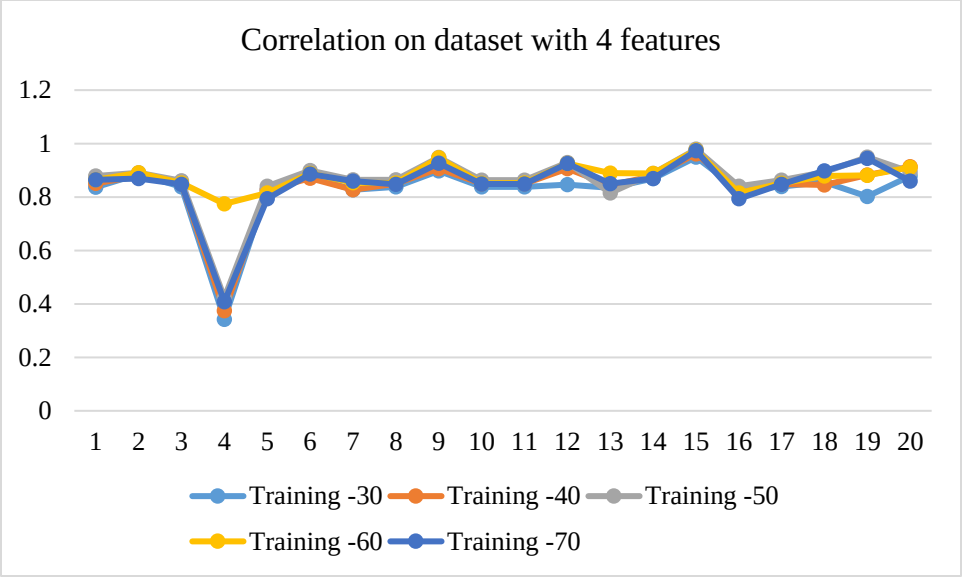


Fig.5.9: Comparison of correlation for different machine learning models.

Boosted Linear Model (Model 4) shows the lowest value of correlation. Parallel Random Forest (Model 15) shows the highest value of correlation. Fig. 5.10 depicts the comparison between machine learning models over the R- Squared evaluation parameter on reduced dataset with 4 classes.

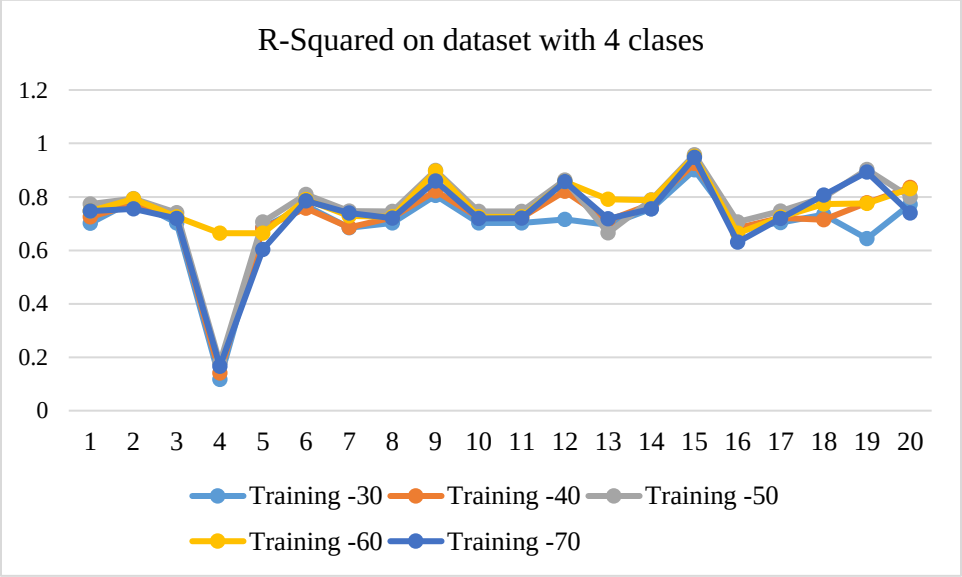


Fig.5.10: Comparison of R-Squared for different machine learning models.

Boosted Linear Model portrays the lowest value of R-Squared. The highest value of R-Squared is shown by Parallel Random Forest that means higher value of accuracy. Fig. 5.11 depicts the comparison between machine learning models over the Root Mean Square Error evaluation parameter on reduced dataset with 4 classes.

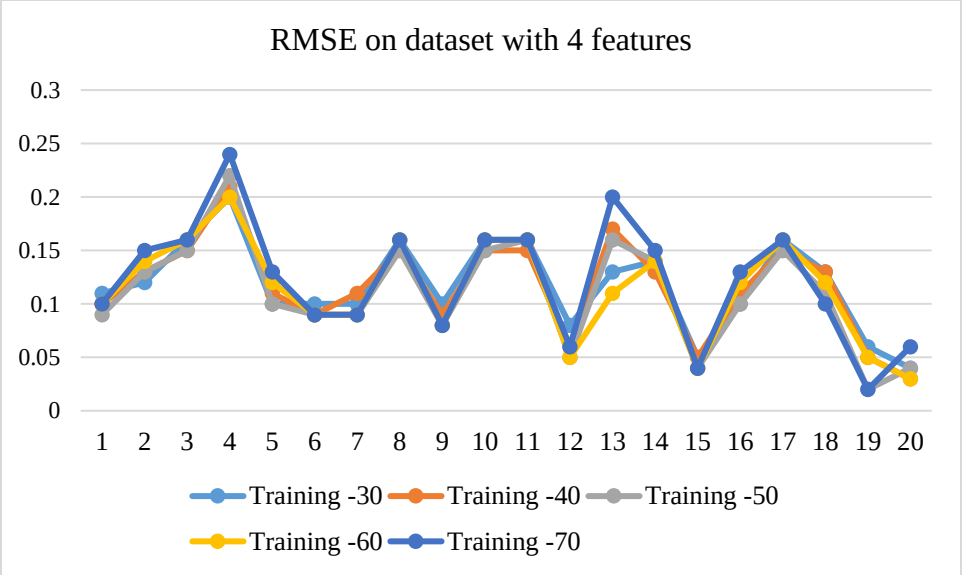


Fig.5.11: Comparison of RMSE for different machine learning models.

The greater values of RMSE is portrayed by Boosted Linear Model and Multilayer Perceptron that means the poor performance of models at this dataset. Fig. 5.12 depicts the comparison between machine learning models over the Accuracy evaluation parameter on reduced dataset with 4 classes.

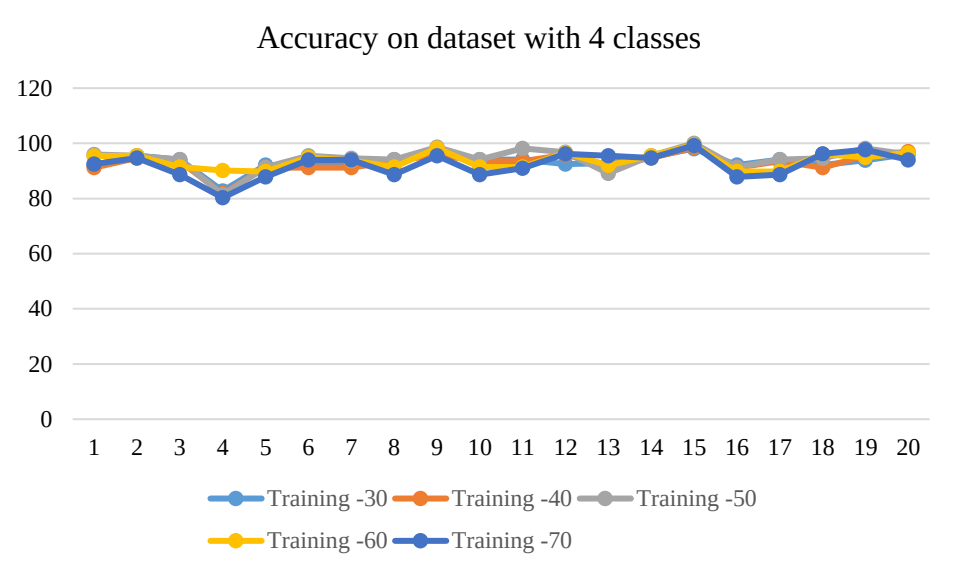


Fig.5.12: Comparison of accuracy for different machine learning models.

For a dataset with 4 classes, the results are being analyzed. At 30% training ratio, the highest accuracy of 98.05% is depicted by Parallel Random Forest and the lowest of 82.74% is shown by Boosted Linear Model. At 40% training ratio, the highest accuracy of 96.96% is depicted by Stochastic Gradient Boosting and the lowest of 81.75% is shown by Boosted Linear Model. At 50% training ratio, the highest accuracy of 100% is depicted by Parallel Random Forest and the lowest of 81.36% is shown by Boosted Linear Model. At 60% training ratio, the highest accuracy of 98.30% is depicted by Guassian Process with Polynomial Kernel and the lowest of 89.77% is shown by CART. At 70% training ratio, the highest accuracy of 99.24% is depicted by Parallel Random Forest and the lowest of 80.3% is shown by Boosted Linear Model.

After applying the prediction module, on these 20 machine learning models top ten models are chosen for cross validation. Table 5.16 depicts the top 10 chosen models based on their accuracy.

Table 5.16: Top 10 chosen models

Model. No	**Model Name**
1	Gaussian Process with Polynomial Kernel
2	Self Organizing Map
3	Partial Least Squares
4	Stochastic Gradient Boosting
5	Parallel Random Forest
6	Bagged CART

7	Multilayer Perceptron
8	Conditional Inference Random Forest
9	Boosted Generalized Linear Model
10	Bagged MARS using gCV Pruning

K- fold cross validation is being applied on top 10 models chosen. In current work, 10-fold cross validation has been chosen and the results are compiled in Table 5.17.

Table 5.17: K-fold cross validation on top 10 models.

Runs	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**	**Model 10**
1	86.82	91.36	90	92.27	89.55	82.27	85.45	84.09	89.09	90.45
2	88.64	88.64	90.91	88.18	90	88.18	87.27	85.91	84.55	91.82
3	84.09	95	92.73	88.64	84.09	90.45	84.55	86.82	90	90.91
4	90.91	90.91	90	87.73	86.82	85.91	86.36	85.45	88.64	88.64
5	88.18	93.18	85	87.27	87.73	88.18	90	90.91	92.27	90
6	94.09	91.82	88.64	88.18	90.91	87.27	87.27	87.73	88.18	89.55
7	89.09	91.36	89.55	90.45	80.45	88.64	85	85.91	92.27	88.18
8	90.91	92.73	88.18	88.18	86.36	88.18	84.55	85.45	91.82	89.09
9	88.64	93.64	90	89.55	85.45	86.36	85	86.82	89.09	91.82
10	82.73	89.09	88.64	87.73	84.55	90	86.36	87.73	87.27	88.64
Average	**88.41**	**91.77**	**89.36**	**88.81**	**86.51**	**87.54**	**86.18**	**86.68**	**89.31**	**89.91**

Based on the average accuracy, top models are selected for ensemble approach. Fig 5.13 represents the average accuracy of the models selected for cross validation. It helps to select top 5 models for Ensembling.

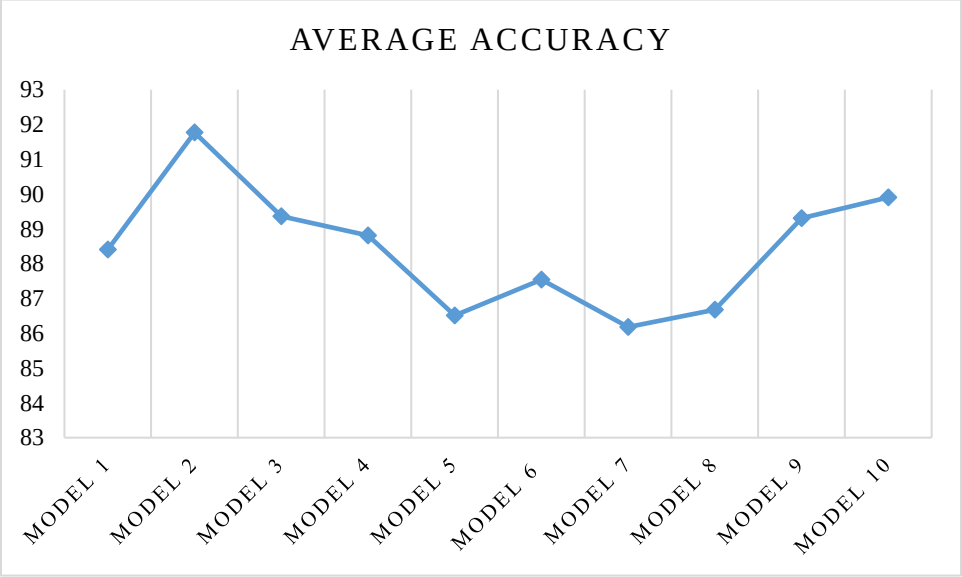


Fig 5.13: Average accuracy for top 10 chosen models.

The top 5 models are chosen based on the average accuracy obtained through k-fold cross validation technique. The top 5 models are depicted in Table 5.18.

Table 5.18: Top 5 chosen models chosen

S.no	**Model Name**
1	Self Organizing Map
2	Bagged MARS using gCV Pruning
3	Partial Least Squares
4	Boosted Generalized Linear Model
5	Gaussian Process with Polynomial Kernel

With these top five selected models, an ensemble model is generated with the help of these models. Multiple ensemble models are designed by using the selected models as depicted in Table 5.19.

Table 5.19: Different ensemble models with the top 5 models.

Sr. No	**Possible Variations**	**Accuracy**
1	Self Organizing Map + Bagged MARS using gCV Pruning	86.14
2	Self Organizing Map + Partial Least Squares	85.206
3	Self Organizing Map + Boosted Generalized Linear Model	86.441
4	Self Organizing Map + Gaussian Process with Polynomial Kernel	86.254
5	Bagged MARS using gCV Pruning + Partial Least Squares	85.3762
6	Bagged MARS using gCV Pruning + Boosted Generalized Linear Model	84.33
7	Bagged MARS using gCV Pruning + Gaussian Process with Polynomial Kernel	86.771
8	Partial Least Squares + Boosted Generalized Linear Model	85.991
9	Partial Least Squares + Gaussian Process with Polynomial Kernel	86.3054
10	Boosted Generalized Linear Model + Gaussian Process with Polynomial Kernel	87.2121
11	Self Organizing Map + Bagged MARS using gCV Pruning + Partial Least Squares	**88.10**
12	Self Organizing Map + Partial Least Squares + Boosted Generalized Linear Model	86.8182
13	Self Organizing Map + Boosted Generalized Linear Model + Gaussian Process with Polynomial Kernel	87.274

14	Bagged MARS using gCV Pruning + Partial Least Squares + Boosted Generalized Linear Model	86.8181
15	Bagged MARS using gCV Pruning + Boosted Generalized Linear Model + Gaussian Process with Polynomial Kernel	85.990
16	Partial Least Squares + Boosted Generalized Linear Model + Gaussian Process with Polynomial Kernel	86.185
17	Bagged MARS using gCV Pruning + Partial Least Squares + Gaussian Process with Polynomial Kernel	86.3124
18	Self Organizing Map + Partial Least Squares + Gaussian Process with Polynomial Kernel	85.31
19	Self Organizing Map + Bagged MARS using gCV Pruning + Boosted Generalized Linear Model	86.6163
20	Self Organizing Map + Bagged MARS using gCV Pruning + Gaussian Process with Polynomial Kernel	86.3124
21	Self Organizing Map + Bagged MARS using gCV Pruning + Partial Least Squares + Boosted Generalized Linear Model	86.82
22	Self Organizing Map + Bagged MARS using gCV Pruning + Partial Least Squares + Gaussian Process with Polynomial Kernel	86.3031
23	Self Organizing Map + Bagged MARS using gCV Pruning + Boosted Generalized Linear Model + Gaussian Process with Polynomial Kernel	85.498
24	Bagged MARS using gCV Pruning + Partial Least Squares + Boosted Generalized Linear Model + Gaussian Process with Polynomial Kernel	86.7492
25	Self Organizing Map + Partial Least Squares + Boosted Generalized Linear Model + Gaussian Process with Polynomial Kernel	86.543
26	Self Organizing Map + Bagged MARS using gCV Pruning + Partial Least Squares + Boosted Generalized Linear Model + Gaussian Process with Polynomial Kernel	87.72

Fig. 5.14 depicts the accuracy for different ensemble models as defined in Table 5.18.

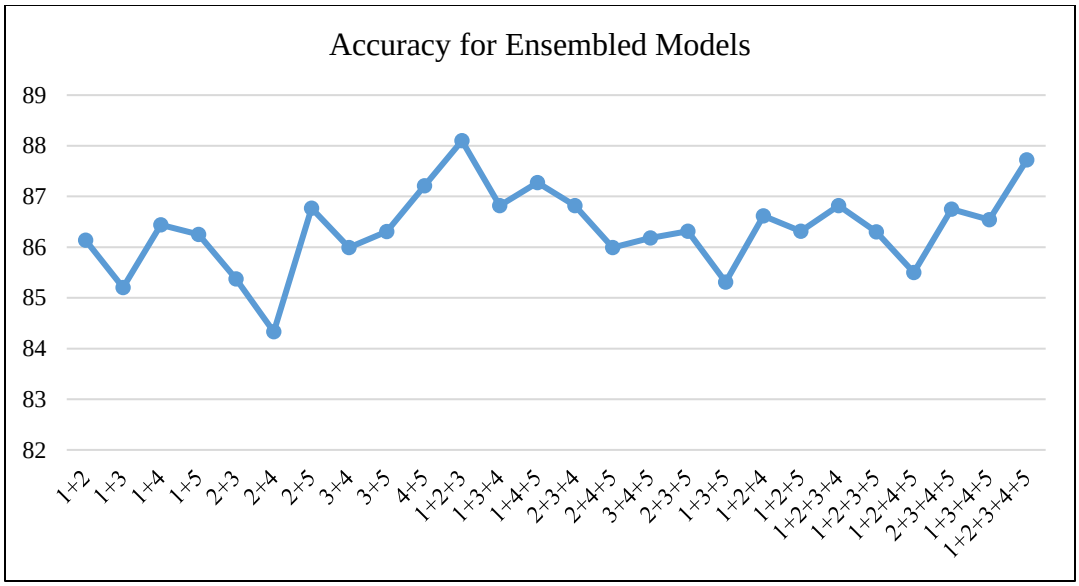


Fig.5.14: Accuracy for different ensembled models.

By applying this ensemble approach, it can be concluded that the highest accuracy of 88.10% is being depicted by the combination of Self Organizing Map, Bagged Mars using gCV Pruning and Partial Least Squares. Bagged MARS using gCV Pruning and Boosted Generalized Linear Model ensemble gives the least accuracy of 84.33%. The overall satisfaction rate attained in module 1 is 3.91.

5.1.3. Sentimental Analysis

The data is collected through feedback during the survey and the results are analyzed. Table 5.20 depicts the classification of reviews into positive, neutral and negative classes.

Table 5.20: Classification of reviews received through feedback.

Reviews	**Percentage**
Positive	54%
Neutral	30%
Negative	16%

Fig 5.15 portrays the classification of reviews obtained through feedback during survey. Table 5.21 shows the accuracy of the classifiers applied on feedback data.

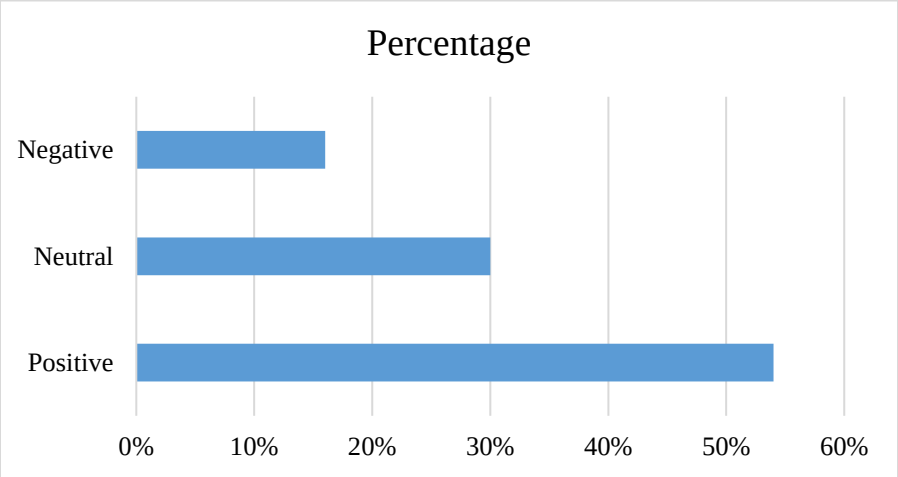


Fig.5.15: Classification of reviews received through during feedback.

Table 5.21: Model Accuracy for feedback data

Model Name	**Accuracy**
Naïve Bayes	68.51
Support Vector Machine	84.32
Decision Tree	74.56

The data is also gathered from various social networking sites and the reviews are classified into three classes i.e. positive, neutral and negative. Table 5.22 depicts the classification of reviews collected through social networking sites. Fig 5.16 depicts the classification of reviews gathered through social media. Table 5.23 concludes the accuracy percentage for different classifiers [60, 61, 62] on reviews collected.

Table 5.22: Classification of reviews received through social media.

Reviews	**Percentage**
Positive	66%
Neutral	24%
Negative	10%

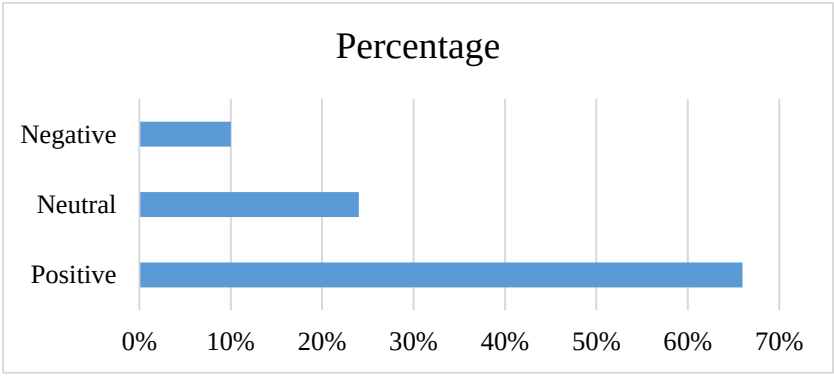


Fig 5.16: Classification of reviews received through social media.

Table 5.23: Model Accuracy for reviews

Model Name	**Accuracy**
Naïve Bayes	66.34
Support Vector Machine	82.79
Decision Tree	71.23

The results conclude that the students have shown a more positive response at social networking sites and Support Vector Machine has shown the highest accuracy among the chosen classifiers. The satisfaction score attained through sentiment analysis of feedback and reviews gathered is 4.35.

Chapter 6
Conclusion and Future Scope

In this specific chapter, the conclusion and future scope of this research work is discussed.

6.1 Conclusion

In this research work, student satisfaction rates are computed within the University. The proposed approach is divided into two modules. The first module deals with the reviews obtained through the questionnaire. It analyzes multiple parameters and approaches for prediction of data is applied. Weighted approach is used for computing the overall satisfaction rate within the University. An ensemble approach using multiple machine learning models is applied for prediction of data. The second module deals with the feedback and the reviews obtained through social media. Various models have been applied for sentiment analysis of the data. These models include Naïve Bayes, Support Vector Machine and Decision Tree.

The overall satisfaction score of the University is computed to be 3.91. Ensemble approach is applied for the prediction of data. The highest accuracy of the ensemble model is obtained to be 84.33% . The positive feedback obtained through questionnaire is 54% and through reviews is 66%. The satisfaction score attained through sentiment analysis is 4.35.

6.2 Future Scope

In future, this research work can be extended by refining the questionnaire and considering all those parameters that affect the satisfaction levels in other universities as well. This research can be extended on a zonal and a national level. Various techniques can be applied to know the importance of parameters that affect the satisfaction levels. Several other sentiment approaches can be applied that includes dealing with unsupervised data. The reviews classified distinctively by analyzing various emotions viz. fear, anger, sarcasm, anxiety, surprise etc. Reviews from anonymous participants can be obtained from student counselling committee that will help to obtain to a fairer picture of the satisfaction level.

The University has its collaboration with other foreign universities and reviews in other languages can also be analyzed using sentiment analysis.

References

[1] “Key Factors for Determining Student Satisfaction in,” no. March, pp. 61–67, 2004.

[2] Customer Satisfaction Survey, https://www.b2binternational.com/publications/customer-satisfaction-survey/ , viewed on November 20.

[3] Student Satisfaction Survey, https://blink.ucsd.edu/sponsor/OSI/opa/sss/index.html

[4] Thrun, Sebastian, and Lorien Pratt, eds. *Learning to learn*. Springer Science & Business Media, 2012.

[5] Sutton, Richard S., and Andrew G. Barto. *Introduction to reinforcement learning*. Vol. 135. Cambridge: MIT Press, 1998.

[6] L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: a survey, Journal of Artificial Intelligence Research 4 (1996) 237–285 ISSN 1076-9757.

[7] P. Kulkarni, “Introduction to reinforcement and systemic ma-chine learning,” *Reinforcement and Systemic Machine Learning for Decision Making*, pp. 1–21, 2012.

[8] Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." *International Journal* 2.6 (2012): 282-292.

[9] Castillo, Luis. "A virtual laboratory for multiagent systems: Joining efficacy, learning analytics and student satisfaction." *Computers in Education (SIIE), 2016 International Symposium on*. IEEE, 2016.

[10] P. Long, G. Siemens. Penetrating the fog: analytics in learning and education. EDUCAUSE. Sept-Oct 2011, p. 31-40.

[11] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, J.L. Santos. Learning Analytics Dashboard Applications. American Behavioral Scientist (2013), p. 1-10.

[12] T. Kamis, Z. Othman, and N. H. M. Saad, “Satisfaction level among Diploma in Medical Electronic students on the sustainability of physical facilities in Politeknik Sultan Salahuddin Abdul Aziz Shah,” *Proc. - 2015 Innov. Commer. Med. Electron. Technol. Conf. ICMET 2015*, no. November, pp. 114–120, 2016.

[13] Z. Zamakhsari, “An Investigation on Students Participation and Satisfaction Towards Online Learning,” pp. 143–147, 2015.

[14] Mendonca, Marcio, et al. "Fuzzy Cognitive Maps Applied to Student Satisfaction Level in an University." *IEEE Latin America Transactions* 13.12 (2015): 3922-3927.

[15] M. Alshammari, R. Anane, and R. J. Hendley, “Students ’ Satisfaction in Learning Style-Based Adaptation,” pp. 55–57, 2015.

[16] S. Nikolic, C. Ritz, P. J. Vial, M. Ros, and D. Stirling, “Decoding Student Satisfaction: How to Manage and Improve the Laboratory Experience,” *IEEE Trans. Educ.*, vol. 58, no. 3, pp. 151–158, 2015.

[17] T. Wolf, “Assessing student learning in a virtual laboratory environment,” *IEEE Trans. Educ.*, vol. 53, no. 2, pp. 216–222, May 2010.

[18] P. Cunningham, “Are Facebook ‘ likes ’ Enough to Assess Student Satisfaction in Open Distance Learning (ODL)? An Incursion into Students ’ Experience of ODL through Online Social Networks (OSNs),” pp. 1–8, 2014.

[19] G. Armenski, M. Kostoska, S. Ristov, and M. Gusev, “Student satisfaction of e-Learning tools for Computer Architecture and Organization course,” *IEEE Glob. Eng. Educ. Conf. EDUCON*, no. April, pp. 630–637, 2014.

[20] Zhang, Heping, Yang Zhan, and Hu Ding. "Teaching Quality Evaluation Based on Student Satisfaction." *Management and Service Science (MASS), 2011 International Conference on*. IEEE, 2011.

[21] Choudhary, Muhammad Abbas. "Factors influencing engineering students' performance and their relationship with the student satisfaction with the teaching, learning as well as overall university experiences." *Information Technology Based Higher Education and Training (ITHET), 2012 International Conference on*. IEEE, 2012.

[22] A. Carbone and J. Ceddia, “Common areas for improvement in physical science units that have critically low student satisfaction,” *Proc. - 2013 Learn. Teach. Comput. Eng. LaTiCE 2013*, pp. 17–24, 2013.

[23] U. Udwhg *et al.*, “Factors Influencing Engineering Students ’ Performance and their Relationship with the Student Satisfaction with the Teaching , Learning as well as Overall University Experiences,” pp. 5–9, 2012.

[24] Mullen, Tony, and Nigel Collier. "Sentiment Analysis using Support Vector Machines with Diverse Information Sources." *EMNLP*. Vol. 4. 2004.

[25] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." *Icwsm* 11.538-541 (2011): 164.

[26] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford* 1.2009 (2009): 12.

[27] Lin, Chenghua, and Yulan He. "Joint sentiment/topic model for sentiment analysis." *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009.

[28] Prabowo, Rudy, and Mike Thelwall. "Sentiment analysis: A combined approach." *Journal of Informetrics* 3.2 (2009): 143-157.

[29] Martineau, Justin, and Tim Finin. "Delta TFIDF: An Improved Feature Space for Sentiment Analysis." *Icwsm* 9 (2009): 106.

[30] Boiy, Erik, and Marie-Francine Moens. "A machine learning approach to sentiment analysis in multilingual Web texts." *Information retrieval* 12.5 (2009): 526-558.

[31] Gamon, Michael. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis." *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

[32] Mostaghel, Rana. "Customer Satisfaction: service quality in online purchasing in Iran." (2006).

[33] Creswell, John W., et al. "Advanced mixed methods research designs." *Handbook of mixed methods in social and behavioral research* 209 (2003): 240.

[34] Customer Satisfaction Survey, http://libweb.surrey.ac.uk/library/skills/Introduction%20to%20Research%20and%20Managing%20Information%20Leicester/page_51.htm-%20Ref

[35] https://surveyanyplace.com/questionnaire-pros-and-cons/

[36] Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. "Improving propensity score weighting using machine learning." *Statistics in medicine* 29.3 (2010): 337-346.

[37] Sabzevari, Hassan, Mehdi Soleymani, and Eaman Noorbakhsh. "A comparison between statistical and data mining methods for credit scoring in case of limited

available data." *Proceedings of the 3rd CRC Credit Scoring Conference, Edinburgh, UK*. 2007.

[38] Elith, Jane, John R. Leathwick, and Trevor Hastie. "A working guide to boosted regression trees." *Journal of Animal Ecology* 77.4 (2008): 802-813.

[39] Strobl, Carolin, et al. "Conditional variable importance for random forests." *BMC bioinformatics* 9.1 (2008): 307.

[40] Strobl, Carolin, et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution." *BMC bioinformatics* 8.1 (2007): 25.

[41] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." *Neurocomputing* 70.1 (2006): 489-501.

[42] Seeger, Matthias. "Gaussian processes for machine learning." *International journal of neural systems* 14.02 (2004): 69-106.

[43] Wahba, Grace. *Spline models for observational data*. Society for industrial and applied mathematics, 1990.

[44] Yuan, Guo-Xun, Chia-Hua Ho, and Chih-Jen Lin. "An improved glmnet for l1-regularized logistic regression." *Journal of Machine Learning Research*13.Jun (2012): 1999-2030.

[45] Keller, James M., Michael R. Gray, and James A. Givens. "A fuzzy k-nearest neighbor algorithm." *IEEE transactions on systems, man, and cybernetics* 4 (1985): 580-585.

[46] Gardner, Matt W., and S. R. Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." *Atmospheric environment* 32.14 (1998): 2627-2636.

[47] Friedman, Jerome H. "Multivariate adaptive regression splines." *The annals of statistics* (1991): 1-67.

[48] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.

[49] Wold, Herman. "Partial least squares." *Encyclopedia of statistical sciences*(1985).

[50] Wu, Bo, and Ram Nevatia. "Cluster boosted tree classifier for multi-view, multi-pose object detection." *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007.

[51] Tamayo, Pablo, et al. "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." *Proceedings of the National Academy of Sciences* 96.6 (1999): 2907-2912.

[52] Friedman, Jerome H. "Stochastic gradient boosting." *Computational Statistics & Data Analysis* 38.4 (2002): 367-378.

[53] Cleveland, William S., and Susan J. Devlin. "Locally weighted regression: an approach to regression analysis by local fitting." *Journal of the American statistical association* 83.403 (1988): 596-610.

[54] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Ijcai*. Vol. 14. No. 2. 1995.

[55] Christensen, N., and Dennis P. Lettenmaier. "A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado River Basin." *Hydrology and Earth System Sciences Discussions* 3.6 (2006): 3727-3770.

[56] Lee Rodgers, Joseph, and W. Alan Nicewander. "Thirteen ways to look at the correlation coefficient." *The American Statistician* 42.1 (1988): 59-66.

[57] Cameron, A. Colin, and Frank AG Windmeijer. "An R-squared measure of goodness of fit for some common nonlinear regression models." *Journal of Econometrics* 77.2 (1997): 329-342.

[58] Willmott, Cort J., and Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." *Climate research* 30.1 (2005): 79-82.

[59] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Ijcai*. Vol. 14. No. 2. 1995.

[60] Rish, Irina. "An empirical study of the naive Bayes classifier." *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. No. 22. IBM, 2001.

[61] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." *Neural processing letters* 9.3 (1999): 293-300.

[62] Du, Wenliang, and Zhijun Zhan. "Building decision tree classifier on private data." *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*. Australian Computer Society, Inc., 2002

List of Publications

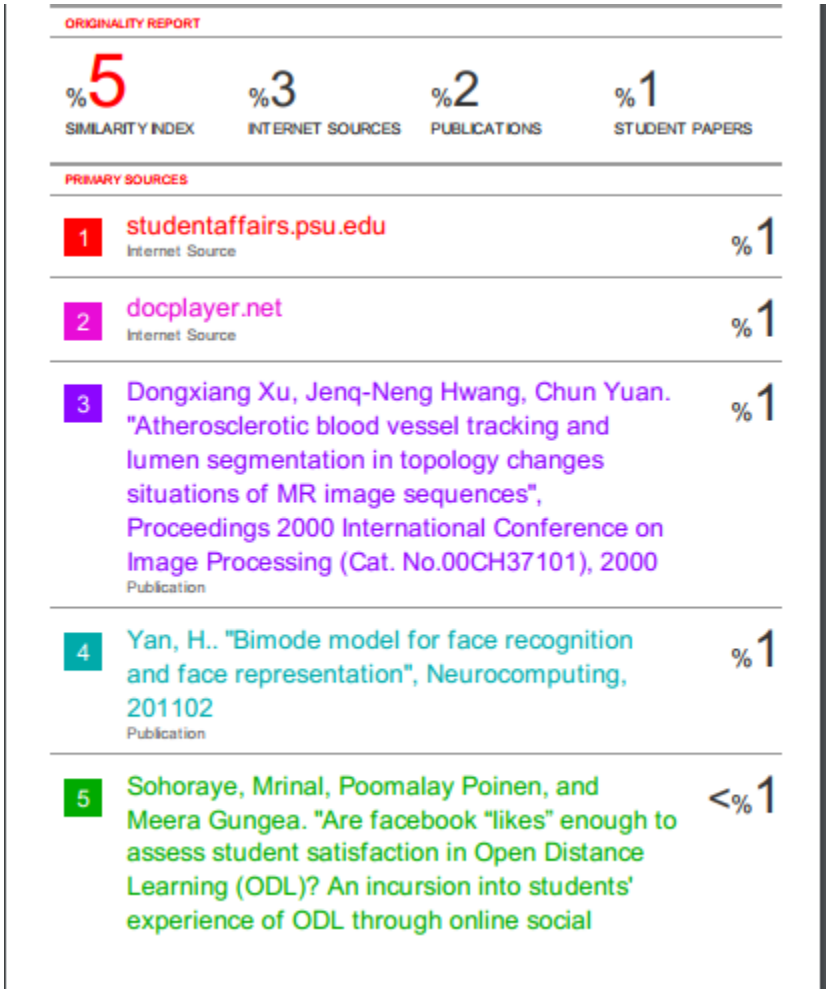
1. Himika, Sukhchandan Randhawa and Maninder Kaur, "*Student satisfaction: A survey at an Indian University" in IEEE* International Conference on Intelligent Computing and Control (I2C2 – 2017) [Accepted]
2. Himika, Sukhchandan Randhawa and Maninder Kaur, "*Student Satisfaction Prediction in an Indian University through Ensembling"* in Sadhana Journal of Indian Academy of Sciences, Springer. [Communicated]

Video Link

The video of this research work can be accessed from:

https://youtu.be/roRfRsyrW44

Plagiarism Report



ORIGINALITY REPORT

%5	%3	%2	%1
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	studentaffairs.psu.edu Internet Source	%1
2	docplayer.net Internet Source	%1
3	Dongxiang Xu, Jenq-Neng Hwang, Chun Yuan. "Atherosclerotic blood vessel tracking and lumen segmentation in topology changes situations of MR image sequences", Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101), 2000 Publication	%1
4	Yan, H.. "Bimode model for face recognition and face representation", Neurocomputing, 201102 Publication	%1
5	Sohoraye, Mrinal, Poomalay Poinen, and Meera Gungea. "Are facebook "likes" enough to assess student satisfaction in Open Distance Learning (ODL)? An incursion into students' experience of ODL through online social	<%1