OPTIMIZED DECISION MAKING ON REAL ESTATE DATA USING DATA ANALYTICS

Thesis Report

Submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

In

Software Engineering

Submitted By

GURSIMRAN KAUR

(801631007)

Under the supervision of

Ms. HARKIRAN KAUR

(Lecturer)

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT,

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY) PATIALA, INDIA, 147004

August, 2018

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, ***"Optimized decision making on Real Estate data using Data Analytics"***, in partial fulfillment of the requirements for the award of degree of Master of Engineering in Software Engineering submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, (Deemed to be University) Patiala, is an authentic record of my own work carried out under the supervision of **Ms. Harkiran Kaur** and refers other researcher's work which are duly listed in the reference section.

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

Gursimran Kaur

Signature

Gursimran Kaur

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

11/09/18

Ms Harkiran Kaur

Lecturer

Computer Science & Engineering Department

Thapar Institute of Engineering and Technology

Patiala

Countersigned by:

(Dr. Maninder Singh)

Head,

Computer Science & Engineering Department

Thapar Institute of Engineering and Technology,

Patiala

(Dr. S.S Bhatia)

Dean (Academic Affairs)

Thapar Institute of Engineering and Technology, Patiala

ACKNOWLEDGMENT

This research work would be incomplete without acknowledging the people who supported and guided me for the successful completion of this work.

First of all I wish to acknowledge the benevolence of God who gave me courage and strength to face the challenges and to overcome the obstacles that occurred while working on this task.

It gives me immense pleasure in expressing thanks to **Ms Harkiran Kaur**, Lecturer, Computer Science & Engineering Department, Thapar Institute of Engineering and Technology, Patiala for her valuable guidance and continual encouragement throughout this research work. The appreciation and continual support she has imparted has been a great motivation to me in reaching a higher goal. Her guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

I am also heartily thankful to **Dr. Maninder Singh**, Honorable Head of Computer Science & Engineering Department, Thapar Institute of Engineering and Technology, (Deemed to be University), Patiala for his kind support and providing basic infrastructure and healthy research environment.

I would also thank the Institution, all the faculty and staff members of Computer Science & Engineering Department, Thapar Institute of Engineering and Technology, (Deemed to be University), Patiala for their direct-indirect help, cooperation and suggestions towards this work.

Last but not the least, I would like to thank my family for their wonderful support and encouragement without which none of this would have been possible.

Gursimran Kaur

Gursimran Kaur

(801631007)

ABSTRACT

Buying or selling a property is a financial as well as an emotional undertaking. In this advanced era, these processes can be addressed differently than earlier, with more accuracy and optimization into them. In traditional times, real estate domain has been slow to embrace the recently emerged techniques. So, it's a high time to start. Technologies such as machine learning can bring tangible benefits to all the parties involved. These may involve sellers, renters, buyers and tenants as well as brokers and agents.

In this work, to make a CUBE without OLAP is clearly to form SQL queries that thinks the result sets (i.e., needed) and that contains comparable data (i.e., would come to execution in view of fuzzy OLAP exercises). There are some noteworthy burdens with this approach in this case. As an issue of first significance, the execution would be unacceptable when the database is broad with various relations required. Despite the fact that, the tests were performed with real time estimations and the audit reaction time was progressed. At the other hand, the queries were executed against the STAR Schema design which is stacked with abundance data to restrain the amount of joins required.

To do comparable queries against a consistent data source would break down execution. Regardless, since the OLAP gadgets are especially created for these kind of queries, they are clearly enhanced for short query response times. A bit of these progressions abuse the read-generally nature of OLAP models and can hardly be found in a by and large valuable source database engine. Second, the reporting would be obliged. A great favored instance of OLAP gadgets is that the customer perceives is multidimensional and the documenting is achieved to a great degree of versatility.

OLAP is outstandingly versatile with both segments, and paying little mind to whether it isn't so ordinary, uncovering more than two estimations is totally possible. Adding to this the roll-up and roll-down assignments improves these kind of gadgets than databases concerning separation of data. Clearly, the necessity for multidimensional data examination for a modest relationship with an obliged database may not require all the

expansive furthest reaches of OLAP gadgets, which consistently are excessive, regardless of the fact that there are open source choices for Business Intelligence game plans too.

This work proposes a real-estate mining process that is performed with the aid of J48 and Support Vector Machine (SVM) classification technique. Here, input dataset is high dimensional real-estate data which is a great barrier for classification. Therefore, initially feature dimension reduction using KPIs have been applied to reduce features space without losing the accuracy of classification. Here, unitary method has been used for selecting basic features from primary (self-created) dataset and secondary (taken from Kaggle website) datasets. Once the feature reduction is performed, the classification is applied based on J48 Decision tree and Support Vector Machine (SVM) classifier. From there on, the achieved information is changed into an arrangement issue that states whether the property has been acquired or not. To prepare the order information, J48 and SVM has been executed. In spite of the fact that, these models perform altogether to order land acquiring, at the same time, experience the ill effects of the parameter tuning issue.

This issue has been settled by considering the outstanding meta-heuristic improvement strategy i.e., NSGA-III. It iteratively upgrades the meta-J48 model to enhance the classification rate by thinking about change using mutation and crossover operations. The acquired arrangements are non-dominant in nature, consequently, proposed model can give better accuracy as well as different parameters simultaneously. Broad experiments have been performed. It has been discovered that the proposed method beats as far as Accuracy, True Positive Rate, True Negative Rate, Precision and F_Measure. Consequently, the proposed strategy is relevant for ongoing land clients.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Data mining is a method used by organizations to change essential data into important information. By using programming to look for plans in broad groups of data, associations can take in additional about their customers and develop all the more intense publicizing frameworks and moreover increase arrangements and decay costs. Data mining depends upon great data assembling and warehousing. This section portrays the idea of Data Mining, Machine Learning (ML), Data Warehousing, Cube Technology, Data Analysis and Types of data analysis focusing on a primary dataset. The connection between data cubes, data distribution centers, and social databases is additionally inspected.

1.1 Data Mining

Data Mining is an analytical method that is utilized to extract hidden patterns from available sources of data. On the contrary, big data problem can be better addressed by an enhanced data analysis process [1]. Data mining comprises of a set of specific methods and techniques aimed specifically at extracting patterns from raw data. Data mining is a favorable borderline in data and information systems. It is skilled in extracting meaningful hidden patterns from voluminous data streams. The strength of data mining lie in the fact that it can deal with varied data types such as text data, image data, relational data, Data Warehouse (DW)s, files, temporal data, demographic data, seismic data, financial data, spatial data and many more possible sources. Though it is not that young now but yet very promising and challenging field which is resulted from the natural evolution of database technology as per needs and requirements and because of its versatility and applicability nature, it is gaining popularity in all possible domains [2]. Data mining is defined as:

"Data mining is an innovation, planned with a goal to empower information investigation, information examination and in addition information perception of huge databases at an abnormal state of deliberation with no particular speculation at the top of the priority list [3]".

Data Mining Systems involve various different technologies such as database technology, neural networks, image processing, Machine Learning (ML), pattern recognition, fuzzy logic, statistics and pattern recognition as shown in Figure 1.1.

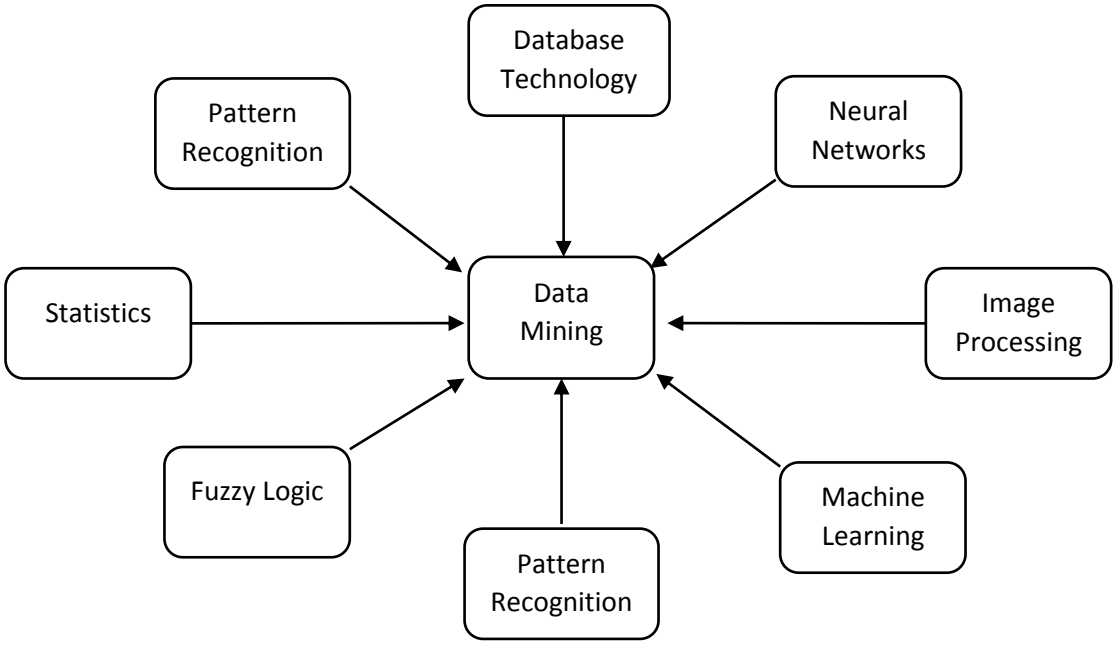


Figure 1.1: The Different technologies available in Data mining system [1]

The abundant data requests for a systematic progress of data mining tools that are focused to transform huge bulky data lumps into of knowledge [4]. Another important issue that gives rise to data mining is the Big Data problem. This problem has been intelligently solved by data mining systems. Moreover, today's ever growing demands cannot be entertained by any single discipline.

Thus, there was a great ultimatum of such technology which must be embedded with the salient features of many expert disciplines. This need is also satisfied by data mining systems, having glimpses from all possible domains. In the data world, data mining is a promising and well-known field, attracting a abundant arrangement of consideration owing to the extensive availability of information in diverse forms. Moreover, there is an urgent need of turning this data into meaningful and useful information to finally gain knowledge that too at a fast pace [5].

In simple words, Data mining is a practice where techniques are applied to preprocessed data (clean, complete, transformed and reduced) for extracting the desired data patterns. It always results in extracting patterns for decision making. Data mining comprises many intelligent techniques to analyze the given data. Data but extract hid useful patterns for decision making.

1.1.1 Advantages of Data Mining

The advantages of data mining are following as:

i. **Publicizing/Retail:** Data mining enables publicizing relationship to accumulate models in light of chronicled information to predict who will react to the new demonstrating undertakings, for example, standard mail, electronic driving effort and so forth. Through the outcomes, promoters will have a fitting framework to oversee pitching priceless things to focused clients. Information mining passes on an unfathomable level of central fixations to retail affiliations in like path as appearing. Through market canister examination, a store can have an authentic age plan in a way that clients can purchase visit anchoring things together with surprising. Besides, it is like way empowers the retail relationship to offer certain discounts for specific things that will pull in more clients.

ii. **Store/Banking:** Information mining gives money related establishments information about push information and credit deciding. By building a model from chronicled customer's data, the bank, and cash related establishment can pick excellent and unsavory credits. Also, data mining draws in banks to see counterfeit charge card trades to remain ace card's proprietor.

iii. **Amassing:** By applying information mining in operational building information, makers can see the inadequate device and pick extol control parameters. For instance, semiconductor makers have a test that even the states of get-together conditions at various wafer age plants are equal, the nature of wafer is a stunning strategy the same and some for cloud reasons even has surrendered. Data mining has been applyied to pick the degrees of control parameters that prompt the making of the astonishing wafer. By then those impeccable control parameters are utilized to make wafers with required quality.

iv. **Governments:** Data mining helps organization relationship by burrowing and breaking down records of the budgetary exchange to cover up away takes after that can see unlawful examination avoidance or criminal exercises.

1.1.2 Disadvantages of Data Mining

The drawbacks of data mining are following as:

i. **Protection Issues:** The pressures over the distinct safety have been ending up immensely starting late particularly after the web is influencing with agreeable frameworks, electronic business, social gatherings, online diaries. In light of security issues, people fear their own particular information is gathered and used misleadingly that conceivably causing them an awesome degree of aggravates. Affiliations assemble information about their customers from different perspectives for understanding their getting hones plots. At any rate, affiliations don't prop up endlessly, nearly days they may be gotten by other or gone. Starting at now, the individual information they have likely is sold to other or spill.

ii. **Security issues:** Security is a fundamental issue. Affiliations have information about their specialists and customers including government deficiency number, birthday and store. In any case, how true this information is taken care is as yet being alluded to. There have been a goliath proportion of cases that architects got to and stole gigantic data of customers from the tremendous undertaking, for instance, Ford Motor Credit Company, Sony… with so much individual and money related information available, the Visa stolen and discount intimidation change into an essential issue.

iii. **Abuse of information/off-kilter information:** Data is accumulated through data burrowing expected for the ethical purposes can be manhandled. This information may be abused by overwhelming people or relationship to take focal reasons for powerless people or mislead a social gathering of people.

1.1.3 Uses of Data mining

Data Mining is a very basic level utilized today my relationship with a solid purchaser center — retail, money related, correspondence, and advancing relationship, to "dug in" into their regard based information and pick evaluating, client inclinations and thing

organizing, effect on deals, buyer unwavering quality and corporate focal points. With information mining, a retailer can utilize inspiration driving offer records of client buys to make things and types of progress to address particular client packages. There are various fundamental areas where information mining is widely utilized:

i. **Future Healthcare:** Information mining holds the mind-blowing potential to update flourishing structures. It consumes evidence and examination to see finest observation that improve awareness and reduce charges. Mining can be applied to predict the size of patients in individual course. Framework is made to achieve certification which the patients acquire through fitting consideration at the perfect place as well as the ideal time. Data mining comparatively helps human organizations prosperity net providers to see duplicity and abuse [8].

ii. **Market Basket Analysis:** Market compartment examination is an indicating strategy in light of a theory that on the off chance that you purchase a specific social gathering of things you will in all probability purchase another party of things. This strategy may engage the retailer to value the buy direct of a purchaser. This data may assist the retailer with knowing the purchaser's needs and change the store's layout in like way. Utilizing differential examination association of results between various stores, between clients in various estimation social events should be possible [9].

iii. **Amassing Engineering:** Data mining gadgets can be to a great degree importance to discover outlines in complex gathering process. Data mining systems can be used in system level sketching to isolate the associations between plan, portfolio, and customer needs data. It can be used to anticipate the change length time, cost, and conditions among various assignments [10].

iv. **Client Relationship Management (CRM):** CRM is tied in with obtaining and holding customers, moreover improving customers' devotion and executing customer-focused frameworks. To keep up an authentic relationship with a customer a business need to assemble data and examine the information. This is the place data mining has its impact. With data mining developments the assembled data can be used for examination. As opposed to being puzzled where

to focus to hold customer, the searchers for the course of action get isolated results [11].

v. **Pressure Detection:** Traditional methods for deception disclosure are monotonous and complex. Data mining helps in giving huge illustrations and changing data into information. Any information that is considerable and profitable is learning. A flawless coercion area system should guarantee information of the significant number of customers. A controlled procedure consolidates the collection of test records. A model is developed using this data and the figuring is made to perceive whether the record is false or not.

vi. **Intrusion Detection:** Any action that will deal with trustworthiness and protection of an advantage is interference. The mindful measures to avoid an intrusion fuses customer approval, refuse programming bungles, and information protection. Data mining can help improve intrusion area by adding a level of focus to quirk revelation. It makes an analyst perceive an activity from essential standard framework development. Data mining in like manner helps evacuate data which is more imperative to the issue [12].

vii. **Lie Detection:** Catching a criminal is basic but drawing out reality from criminal is troublesome. Law approval can use mining strategies to look at bad behaviors, screen correspondence of suspected dread based oppressors. This technique helped to find noteworthy cases in data which are large unstructured substance. The data test assembled from past examinations are broke down and a model for lie distinguishing proof is made. With this model, methods can be made by the need [13].

viii. **Cash related Banking:** With modernized putting aside additional money wherever tremendous extent of information should be made with new exchanges. Information mining can add to placing everything all together issues in putting aside some money and back by discovering cases, causalities, and associations in business data and market costs that are not rapidly obvious to officials in light of the way that the volume information is exorbitantly colossal or is made too rapidly to screen by specialists. The administrators may discover these data for

better dividing, concentrating on, getting, holding and keeping up a favorable client.

1.2 Machine Learning (ML)

Machine Learning (ML) encourages PCs to do what falls into place without a hitch for people and creatures: the gain for a fact. Machine Learning (ML) procedures operate computational approaches to learn data straightforwardly as of information without depending on a foreordained condition as a model. The procedures enhance their implementation as the no. of tests accessible for knowledge increments [14]. Machine Learning (ML) uses two sorts of frameworks: supervised learning, which readies a model on known data and yield data with the goal that it can anticipate future yields, and unsupervised learning, which finds shrouded designs in input data as shown in Figure 1.2.

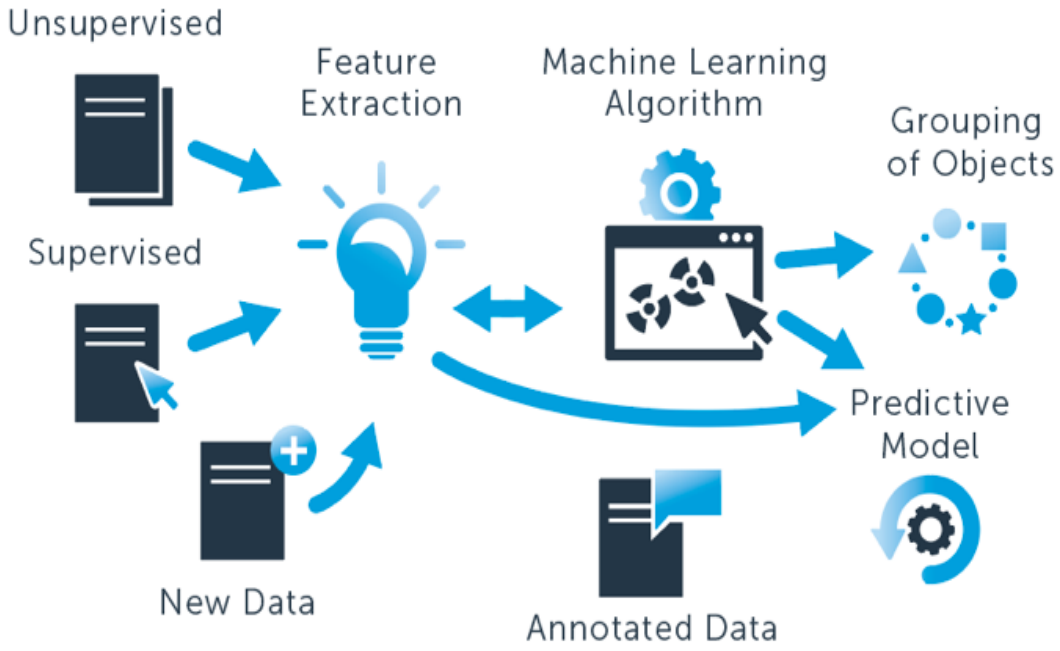


Figure 1.2: The basic illustration of Machine Learning (ML) [14]

1.2.1 Supervised Learning

The point of supervised Machine Learning (ML) is to fabricate a model that makes expectations in view of proof within the sight of vulnerability. A supervised learning calculation takes a known arrangement of input information and known reactions to the

information (yield) and prepares a model to create sensible expectations for the reaction to new information. Supervised learning utilizes classification and regression methods to create prescient models [15].

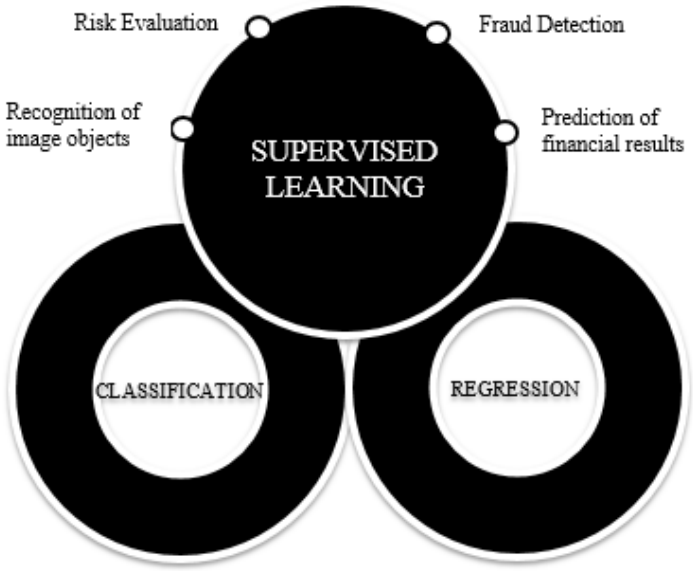


Figure 1.3: The basic illustration of supervised learning [15]

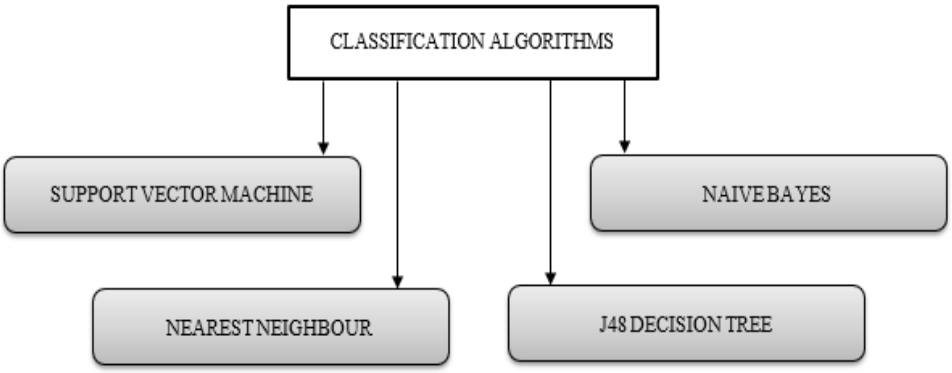


Figure 1.4: Types of classification algorithms [15]

Classification methods foresee discrete reactions, for instance, regardless of whether an email is spam, or whether a tumor is cancerous. Classification models order input information into classifications. Common applications incorporate restorative imaging, discourse acknowledgment, and credit scoring. Classification is the foremost energetic

phases of the data mining process. It is a data mining utilized to divide the data into subcategories [15].

Classification is a supervised leaning method, which employs the class information of the data, for classifying the new incoming records, whose class information is not available. To perform an effective classification, two parameters are required- the number of classes and another is the criteria for deciding the class members. Classification works on categorical data that is of unordered and discrete nature. Before the classification process is applied, the given data is pre-processed which may include data cleaning, attribute relevance analysis and transformation [18]. After that, information set is decomposed into two phases - learning and testing phase because classification is carried out in two phases-:

- Learning Phase (Training)
- Testing Phase.

Finally, the maximum harmonized class is taken as the data class. Popular classification approaches are Decision trees, Bayesian classification, Backpropagation method, Support Vector Machine, k Nearest Neighbor, Neural Networks, Rule-based, Genetic algorithms, and Fuzzy logic.

Regression methods foresee nonstop reactions, for instance, changes in temperature or variances in control request, run of the mill applications etc. Regression methods include algorithms likewise linear regression, neural networks, ensemble methods, support vector regression etc., as shown in Figure 1.5.

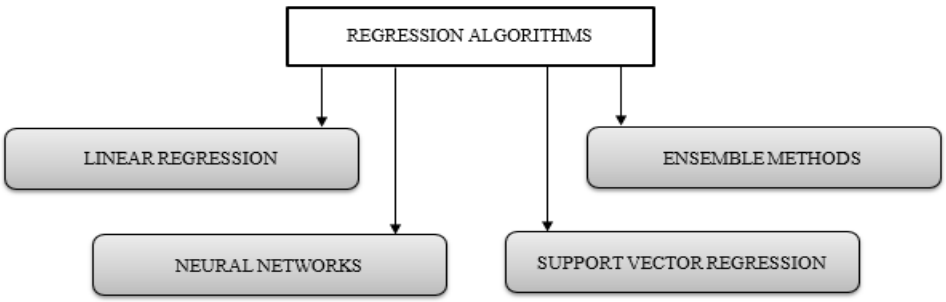


Figure 1.5: Types of regression algorithms [14]

1.2.2 Unsupervised Learning

Unsupervised learning finds hidden patterns in information [15]. It is utilized to induce conclusions after data-sets comprising of original data deprived of categorized reactions as shown in Figure 1.6.

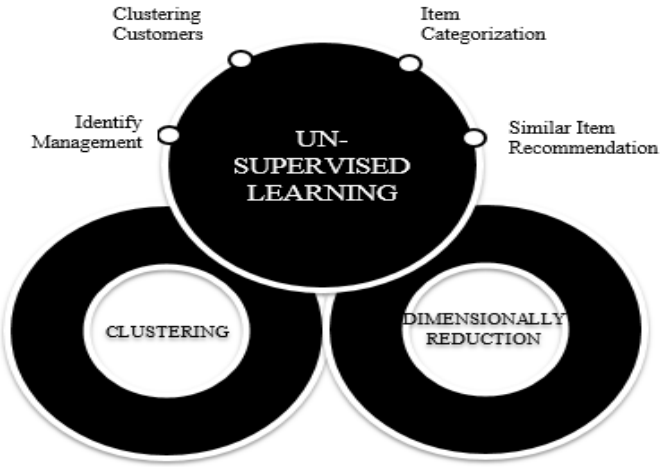


Figure 1.6: The basic illustration of unsupervised learning [15]

Clustering is the most widely recognized unsupervised learning system. It is utilized for exploratory information investigation to discover groupings in information. Applications for clustering incorporate quality grouping examination, statistical surveying, and protest acknowledgment [15]. Clustering algorithms include k means, hierarchical, Gaussian mixture, hidden Markova model etc., as shown in Figure 1.7.

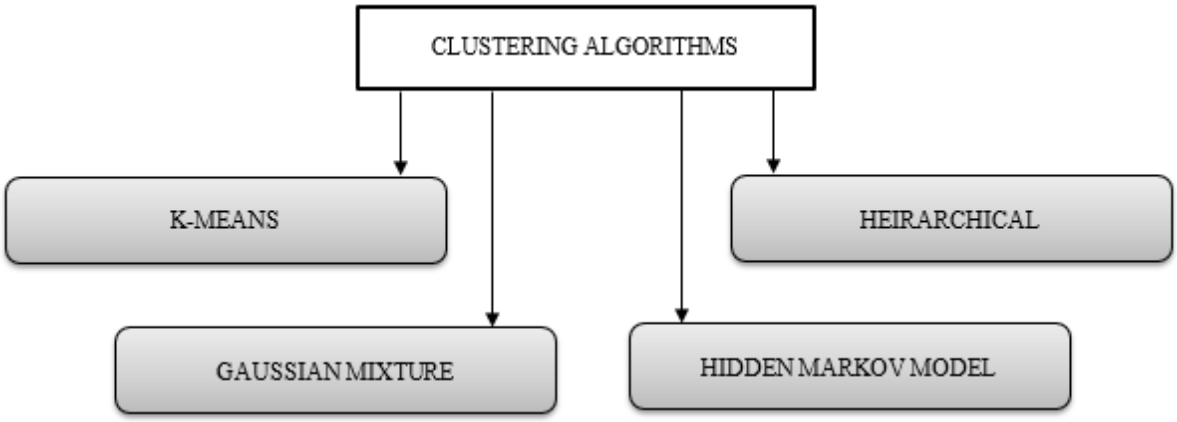


Figure 1.7: Types of clustering algorithms [15]

1.2.3 Uses of Machine Learning (ML)

One of the mainstream uses of AI is Machine Learning (ML), in which PCs, programming, and gadgets perform through discernment (fundamentally the same as human cerebrum) [16,17]. Variety of applications of Machine Learning (ML) are:

i. **Virtual Personal Assistants:** Machine Learning (ML) is a vigorous piece of individual associates, as they gather and refine the data based on past attachment with them. Afterward, this arrangement of information is used to render results that are customized to inclinations. Virtual Assistants are incorporated into an assortment of stages.

ii. **Activity Predictions:** GPS route administrations have been utilized. While using that, present areas and speeds are being spared at a focal server for overseeing activity. This information is then used to construct a guide of the current movement. This helps in keeping the action and does choke examination. The fundamental issue is that there is less number of autos that are furnished with GPS. Machine Learning (ML) in such situations evaluates the districts where blockage can be found based on day by day encounters.

iii. **Online Transportation Networks:** When booking a taxi, the application devices cost the ride. When sharing these administrations booking can or cannot limit the alternate routes. The appropriate response to such problems of booking a taxi on desired route is Machine Learning (ML). Jeff Schneider, the building lead at Uber ATC uncovers in a meeting that they utilize Machine Learning (ML) to characterize value flood hours by foreseeing the rider request. In the whole cycle of the administrations, Machine Learning (ML) is assuming a noteworthy part.

iv. **Recordings Surveillance:** Visualizing a self-contained individual observing different camcorders is surely a troublesome activity to be performed. This is the reason for preparing PCs to carry out this activity bodes well. The video reconnaissance framework these days are fueled by AI that makes it conceivable to identify wrongdoing before they happen. They track surprising conduct of individuals like standing unmoving for quite a while, faltering, or snoozing on seats and so on. The framework would thus be able to give a caution to human

orderlies, which can eventually maintain a strategic distance from accidents. This occurs with machine getting the hang of doing its activity at the backend.

v. **Online Customer Support:** Various sites these days offer the choice to talk with client bolster agent while they are exploring inside the site. Be that as it may, few out of every odd site has a live official to answer your inquiries. In the majority of the cases, you converse with a chatbot. These bots tend to extricate data from the site and present it to the clients. Then, the chatbots propel with time. They have a tendency to comprehend the client questions better and serve them with better answers, which is conceivable because of its Machine Learning (ML) calculations.

1.3 Data Analysis Technique: Classification

Machine Learning (ML) is one of the quickest developing regions of software engineering, with extensive applications. It alludes to the mechanized discovery of significant examples in information. Machine Learning (ML) devices are worried about supplying programs with the capacity to learn and adjust.

Machine Learning (ML) has turned out to be one of the backbones of Information Technology and with that, a somewhat focal, but normally concealed, some portion of our life. With the consistently expanding measures of information getting to be accessible, there is a valid justification to trust that understanding information examination will turn out to be considerably more unavoidable as an important element for innovative advancement [18]. Figure 1.8 explains how the input dataset is cleaned and further is cleaned data is reduced to training and testing set to apply classification methods to check whether test results are accepted or rejected.

Individuals are regularly inclined to committing errors amid examinations or, perhaps, when endeavoring to build up connections between numerous highlights. Data Mining and Machine Learning (ML) are twins from which a few bits of knowledge can be determined through appropriate learning calculations. There has been a gigantic advancement in data mining and Machine Learning (ML) because of the development of small innovation which realized the interest in finding shrouded designs in information to determine appreciation. The combination of measurements, Machine Learning (ML), data

hypothesis, and figuring has made a strong science, with a firm numerical base, and with great instruments. Machine Learning (ML) calculations are sorted out into a scientific categorization in view of the coveted result of the calculation. Administered learning produces a capacity that maps contributions to wanted yields. The exceptional information age has made machine taking in methods wind up advanced every once in a while [18].

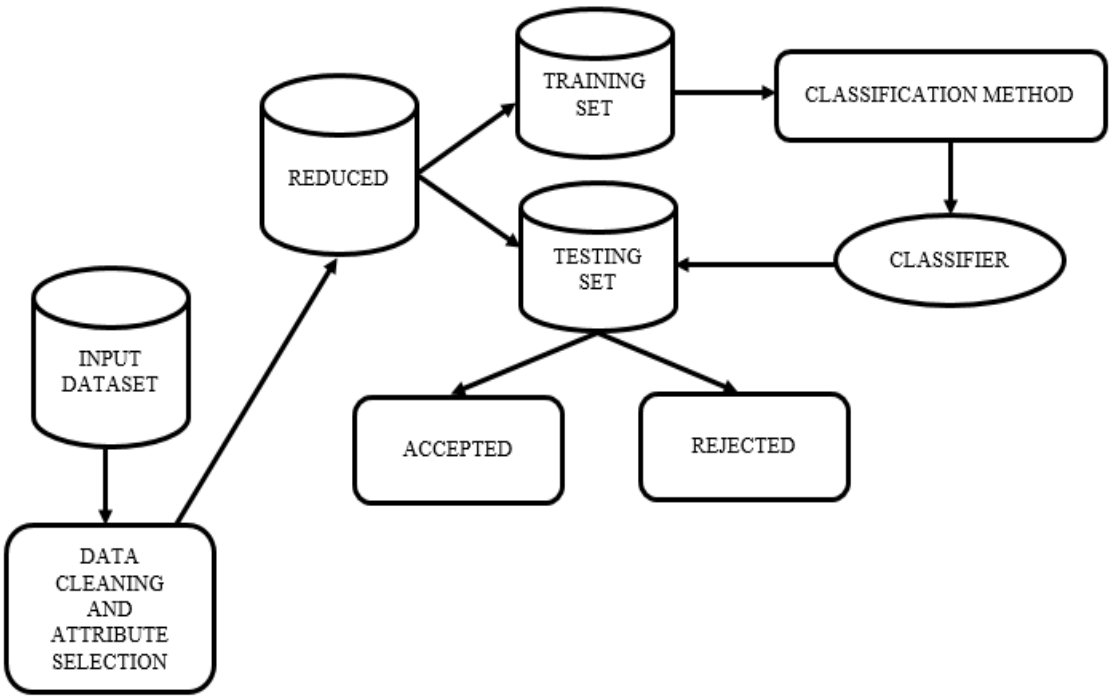


Figure 1.8: Classification Model [18]

Machine Learning (ML) is perfectly intended for accomplishing the accessibility hidden within Big Data. Machine Learning (ML) hand over's on the guarantee of extracting from big and distinct data sources through outlying less dependence scheduled on individual track as it is data determined and spurts at machine scale. Machine Learning (ML) is fine suitable towards the intricacy of handling through dissimilar data origin and the vast range of variables as well as the amount of data concerned where Machine Learning (ML) prospers on increasing datasets [18]. The procedure of utilizing Machine Learning (ML) to a practical problems using a classification model is described in Figure 1.9.

At the liberty from the confines of individual-level thought and study, Machine Learning (ML) is clever to find out and show the patterns hidden in the information. Machine Learning (ML) is the way toward taking in an arrangement of principles from occasions (cases in a preparation set), or all the more, as a rule, making a classifier that can be utilized, to sum up from new occurrences. -

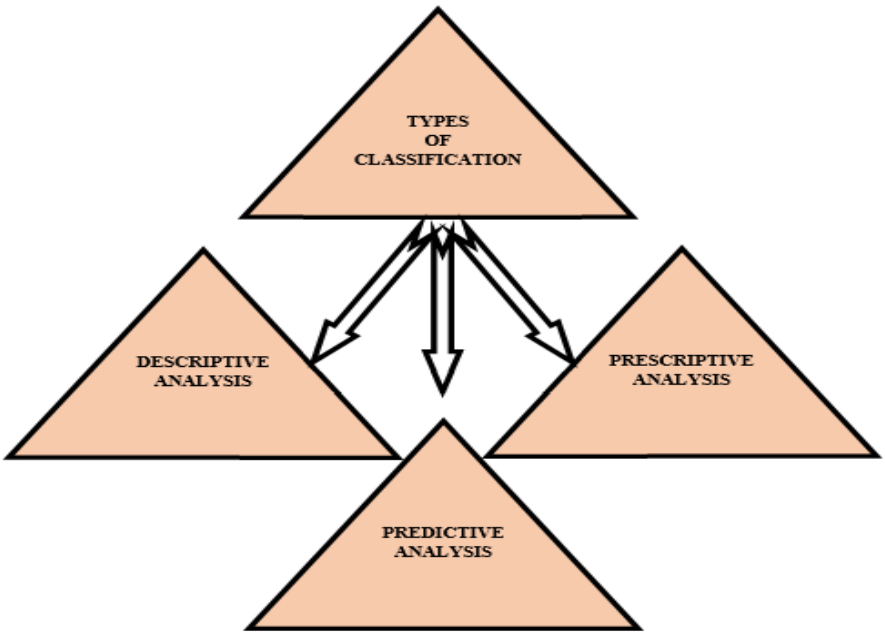


Figure 1.9: Types of Classification [18]

This work centers around the order of Machine Learning (ML) calculations and deciding the most proficient calculation with most elevated exactness and accuracy. And also setting up the execution of various calculations on vast and littler informational collections with a view group them accurately and give knowledge on the best way to construct regulated Machine Learning (ML) models. There are predominantly three kinds of analysis is performed as shown in figure 1.9.

1.3.1 Descriptive Analysis

Descriptive Analysis is a procedure of assessing, purging, changing and demonstrating information with the objective of finding helpful data, advising conclusions, and supporting basic leadership. Descriptive Analysis has numerous features and methodologies, incorporating various procedures under an assortment of names while

being utilized in various business, science, and sociology spaces. So, the whole of this descriptive analysis process is carried out using Extract, Transform and Load (ETL) Process [19] on data and moving towards Data Warehouse (DW).

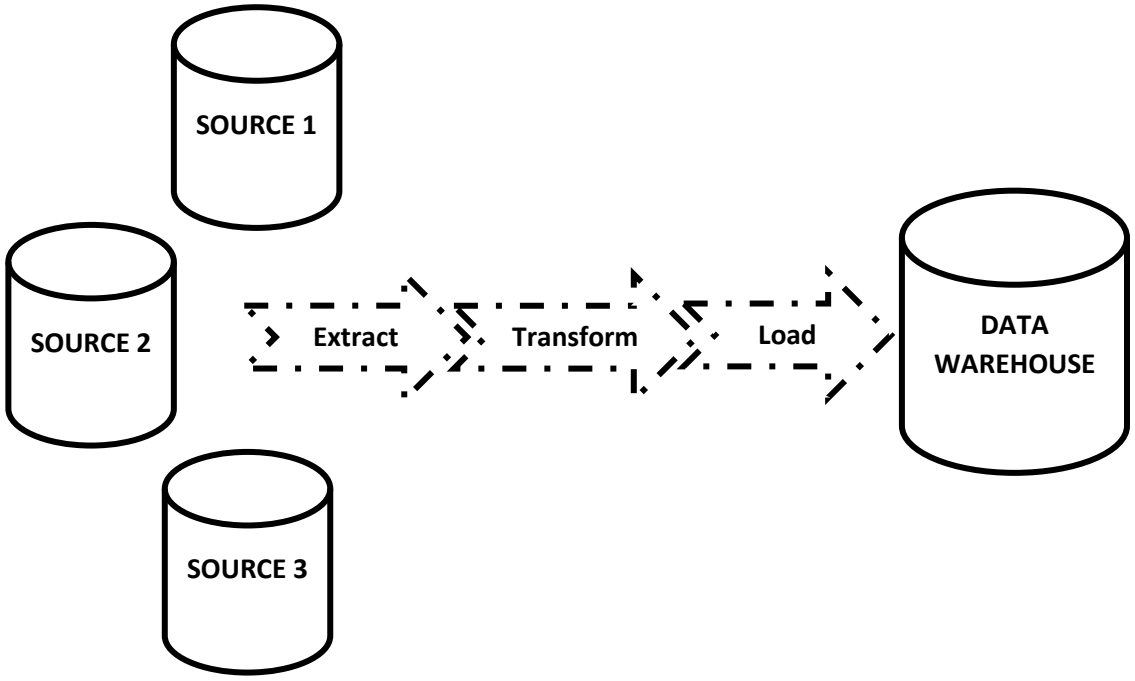


Figure 1.10: The process of converting OLTP to OLAP [19]

Data Warehouse (DW): Data warehousing is a process that was developed from the immense amount of historic electronic data and from the need to utilize that data to achieve objectives that are connected to daily processes. The corporate database stores definite information on the tasks performed by branches. To meet the daily needs queries can be issued to retrieve data. All together for this procedure to work, database directors should first define the desired query after nearly examining database inventories. This can take a couple of hours in light because the amount of data stored is gigantic, so the query complexity increases.

At last, a report is produced. The present Data Warehousing processes isolate Online Analytical Processing (OLAP) systems from Online Transactional Processing (OLTP) systems by creating another data repository that incorporates essential data from different sources, legitimately organizes data formats, and after that makes data accessible for analysis and assessment aimed for arranging and basic decision making forms.

Extract, Transform and Load (ETL) Process: The way toward extracting data from source frameworks and carrying it into the information distribution center or so commonly called Data Warehouse (DW) (DW) is known as Extract, Transform and Load (ETL) Process, which remains for extraction, transformation, and loading of useful data. The acronym ETL is maybe excessively oversimplified, in light of the fact that it overlooks the transportation stage and infers that each of alternate phases of the process is distinct. The procedure and tasks of ETL have been notable for a long time, and are not really exceptional to DW environment.

Extract: During extraction, the desired data is distinguished and extracted from a wide range of sources, including database frameworks and applications. Regularly, it isn't conceivable to distinguish the particular subset of interest; consequently more data than would normally be appropriate must be removed, so the recognizable proof of the important information will be done at a later point in time. Contingent upon the source framework's capacities, a few changes may happen during this extraction procedure. The size of the extracted data varies from many kilobytes up to gigabytes, contingent upon the source framework and the business circumstance. Basically, extraction is done from a variety of heterogeneous sources and also from On-Line Transactional Processing (OLTP) systems. The extraction step is designed in a way that it doesn't contrarily influence the source system. Each different source utilizes an alternate format. Data source formats include RDBMS, XML (like CSV, JSON) files. Thus, the transformation process converts data into a suitable format.

Transform: After information is extracted, it must be physically transported to the objective framework for additional preparation. Contingent upon the picked method for transportation, a few changes should be possible during this procedure, as well. Before transferring data to Data Warehouse (DW), it must be transformed to a homogeneous form. So, the homogeneous form is achieved using SQL queries that is Data Definition Language (DDL) and Data Manipulation Language (DML) commands. The primary aim of transformation is to load data to the Data Warehouse (DW) in a cleaned and general configuration. This is on an account, that when the information is gathered from various sources, each source will have their own benchmarks. Alternate things that are done in this progression are:

1. Cleaning (e.g. "Male" to "M" and "Female" to "F" and so on.)
2. Separating (e.g. choosing just certain sections to load)
3. Improving (e.g. Full name to First Name, Middle Name, Last Name)
4. Splitting one column into multiple columns
5. Combining information from various sources

Hence, the data is transformed from a heterogeneous form to homogeneous form. Sometimes information does not require any changes. So, here such data is said to be "rich data" or "direct move" or "go through" data.

Load: This is the last advancement in the ETL process. In this progression, the extracted and transformed data is loaded to the Data Warehouse (DW). With a specific end goal of influencing data to load proficiently, it is important to file the database and impair requirements before loading the data. All the three stages in the ETL process can run in parallel. Information extraction requires significant investment thus the second step of the change process is executed all the while. This gets the final data for the third step of loading. When the data is prepared it is loaded instantly. So, finally, the homogeneous data is loaded in Data Warehouse (DW).

Hence, OLTP systems are converted to OLAP systems, because input in the form of OLTP systems is given into source 1, source 2 and source 3 and output is achieved in Data Warehouse (DW) in the form of multidimensional cube structure after performing three basic steps that is Extract, Transform and Load as shown in Figure 1.10. Since traditional On-Line Transactional Processing (OLTP) systems are outdated, so they are replaced with On-Line Analytical Processing (OLAP) (OLAP) systems because of their purposeful and routine necessities.

OLAP technology is the need of an hour for every business domain to apply aggregations on large datasets wherein these datasets are residing in large repositories called Data Warehouse (DW). OLAP systems are used to organize business domains and for the decision-making process using aggregations that describes descriptive analytics. OLAP databases can be broken down into one or more cubes which are organized by cube administrator according to the way it may retrieve the data. These cube structures are all

about aggregating measures based on dimensions and hierarchies. Finally, OLAP makes decisions based on these pre-aggregated values of cube structures [20].

Business Intelligence systems are the solutions for gathering data from numerous sources and transforming that collected data so that it is consistent and stored in a single zone and presents us the information for decision making. Business Intelligence systems can have up to 5 layers. That is explained below:

i. **Data Source Layer:** Data source layer is composed of data that the organizations use in the daily routine that is data in text files, Microsoft office, and external data also inform or manual diaries. So, this data is difficult to use to create reports and therefore is converted to a homogeneous form using the transformation layer.
ii. **Data Transformation Layer:** Data stored in systems is extracted by this layer from multiple sources and is modified so that it is internally consistent and loaded onto a data storage system.
iii. **Data stockpiling and recovery layer:** this layer is an information distribution center that has been made in a social database administration framework. The information stockroom is essentially a capacity framework. Develop information stockroom contains finish information that is noteworthy information and present information. To decrease the weight on Data Warehouse (DW) and to rearrange clients issue to get to information, information about individual subjects territories are separated from Data Warehouse (DW), brief and stacked into information bazaars. The shops can be social DBMS or multidimensional OLAP.
iv. **A systematic layer:** This layer is utilized to transform information into data and give fast and simple access to that data for chiefs. Multidimensional OLAP databases frame this systematic layer of business insight frameworks.
v. **Presentation layer:** Visualization devices shape the introduction layer of business insight frameworks. At that point, devices help to picture information in any arrangement. These are useful to convey forward routine business. Dashboard reports with KPI's empower administrators to rapidly decide if the undertaking is meeting every one of the targets of present need or no.

1.3.2 Predictive Analysis

Prediction is another popular analysis process which works on ordered continuous-valued data. It is unsupervised learning as no class label attribute is available. Prediction is applied to achieve numeric prediction from given data. Classification can only derive that whether to give the loan or not but prediction will predict the loan amount also. Prediction basically works with statistical methods like Regression techniques. All prediction techniques revolve around two types of variables: predictor and response. Even few of the classifiers can be transformed for a prediction like SVM, back propagation, and k-nearest-neighbor.

1.3.2.1 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally portrayed by a confining hyperplane. By the day's end, given named getting ready data (coordinated taking in), the estimation yields a perfect hyperplane which characterizes new cases. In two dimensional space, this hyperplane is a line disengaging a plane in two areas wherein each class lay on either side [21].

1.3.2.2 Naive Bayes

A Naive Bayes classifier is a calculation that uses Bayes' hypothesis to arrange objects. Innocent Bayes classifiers accept solid, or guileless, autonomy between qualities of information focuses. Prevalent employments of gullible Bayes classifiers incorporate spam channels, content examination, and restorative determination. These classifiers are generally utilized for Machine Learning (ML) since they are easy to execute. Guileless Bayes is otherwise called basic Bayes or autonomy Bayes [22].

1.3.2.3 Nearest Neighbor

The Nearest Neighbor algorithm is easy to execute but difficult to actualize because it executes rapidly, yet it can once in a while miss shorter routes which are effortlessly seen with human understanding, because of its "avaricious" nature. In the most pessimistic scenario, the calculation results in a visit that is any longer than the ideal visit.

To be exact, for each steady r there is an example of the traveling salesperson problem with the end goal that the length of the visit processed by the closest neighbor calculation

is more noteworthy than r times the length of the ideal visit. In addition, for each number of urban areas, there is a task of separations between the urban areas for which the closest neighbor heuristic delivers the novel most noticeably awful conceivable visit.

1.3.2.4 J48 Decision Tree

Decision Tree algorithm is to find the way in which the traits vector carries on for different events. Also on the bases of the planning illustrations, the classes for them as of late made events are being found. This calculation creates the norms for the desire for the target variable [23]. With the help of tree arrange calculation the essential apportionment of the data is easily sensible. J48 is an extension of ID3.

The additional features of J48 are speaking to missing characteristics, choice trees pruning, tenacious quality regard ranges, induction of standards, et cetera. J48 is an open source Java use of the C4.5 figuring. In the example of potential overfitting, pruning can be used as a gadget for précising. In various calculations the gathering is performed recursively until every single leaf is unadulterated, that is the request of the data should be as impeccable as would be reasonable. This calculation delivers the benchmarks from which particular character of that data is created. The objective is consistently hypothesis of a choice tree until the point that it grabs amicability of flexibility and precision.

1.3.3 Prescriptive Analysis

The moderately new field of prescriptive examination enables clients to recommend various distinctive conceivable activities to and control them towards an answer. More or less, these investigations are tied in with giving counsel. Prescriptive investigation endeavor to measure the impact of future choices so as to prompt on conceivable results previously the choices are really made. Getting it done, prescriptive investigation predicts what will happen, as well as why it will happen giving proposals with respect to moves that will make preferred standpoint of the forecasts.

These examinations go past graphics and prescient investigation by prescribing at least one conceivable blueprints. Basically, they foresee numerous prospects and enable organizations to evaluate various conceivable results in light of their activities. Prescriptive examination utilizes a blend of methods and devices, for example, business

rules, calculations, Machine Learning (ML) and computational demonstrating systems. These strategies are connected against contribution from various informational collections including chronicled and value-based information, continuous information encourages, and huge information.

The prescriptive investigation is generally mind-boggling to direct, and most organizations are not yet utilizing them in their day by day course of business. At the point when actualized effectively, they can affect how organizations decide, and on the organization's main concern. Bigger organizations are effectively utilizing prescriptive examination to streamline generation; planning and stock in the store network to ensure that are conveying the correct items at the opportune time and advancing the client encounter.

1.3.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning model that is defined as the finite dimensional vector spaces where each dimension characterizes a feature of a particular object. In this way, SVM has been proved as an effective method in high-dimensional space problems. Due to its computational competence on huge datasets SVM is typically used in document classification, sentiment analysis, and prediction-based tasks. Support Vector Machine (SVM) is noticed as the first choice for classification problems. Support Vector Machines (SVMs) are nothing but machines constructed for classifying positive and negative classes. These are mainly dependent on support vectors, which are decisive points for classifying data. The attractiveness of SVMs lies in its mathematical equations, pretty pictorial representations, excellent generalization abilities. They give optimal and global solutions, with low overfitting and overcomes the curse of dimensionality problem.

SVMs are designed based on optimization methods and are the most significant tools for solving the problems of Machine Learning (ML) with finite training data. These are dependent on the Structural Risk Minimization (SRM) principle. They exploit large mathematical foundations to avoid overfitting and better empirical results. Mostly SVMs adjust Machine Learning (ML) problems to optimization problems; specifically, the

convex optimization problems are used in the early era of SVMs in the 1990s. Without Statistical Learning Theory (SLT) the definition of SVMs is incomplete.

There are three aspects which make SVM's more successful namely maximal margin hyperplane construction using support vectors, dual theory, and kernel trick. SVMs are straightaway used in various applications like handwriting recognition, intrusion detection, Speech recognition, Bioinformatics, information extraction, face detection and many more. And also they are treated as a remarkable technique because of its scope to handle data with larger dimensions. Unlike neural networks, it will not sustain the local minima problem, and consider very less modeling parameters, producing stable results. But, they have slower training times particularly with non-linear data and huge input data. SVMs are generally used as binary classifiers.

1.3.3.2 NSGA-III

In this section, the proposed NSGA-III-tree-based based Machine Learning (ML) technique to classify real estate data. NSGA-III is a well-known metaheuristic technique which can find optimal solutions. Initially, a collection of reference points is created. Intended for an ideal point, therefore it is really estimated with the bare minimum cost discovered to date to get intent objective and is kept up to date in the search [24].

Following normalization, the particular clustering user is used to split the particular users inside into a set of clusters where the cluster is definitely depicted by the reference point. Then, a non-dominated organizing dependent on the importance (not Pareto-dominance) works to help classify in unique non-domination levels Dominance, which is a key principle in NSGA-III, will be presented later.

1.3.3.3 Genetic Algorithm (GA)

Genetic Algorithm (GA) is a meta-heuristic that impersonates the procedure of normal assessment. The heuristic is routinely used to create valuable answers for improvement and pursuit issues. The hereditary calculation has a place with the bigger class of transformative calculations, which create answers for enhancement issues utilizing procedures motivated by normal advancement.

For example, legacy, change determination and hybrid. The hereditary calculations are essential while finding affiliation rules since they work with the worldwide pursuit to find the arrangement of things recurrence and they are less mind-boggling than different calculations regularly utilized in information mining [25]. The hereditary calculations for the revelation of affiliation rules have been incorporated in genuine issues, for example, business database, science and misrepresentation identification occasion consecutive examination. Genetic operators are of three types as explained below:

i. **Selection:** Selection deals when using the probabilistic natural selection, in that, more fit chromosomes are chosen to survive. Where fitness is actually a comparable way of measuring how well a chromosome solves the problem at hand. There differ strategies to implement selection in genetic algorithms. They are really tournament selection, roulette wheel selection, proportionate selection, rank selection, and steady-state selection etc.

ii. **Crossover:** Crossover is applied on those solutions which are still surviving and has not been discarded. This operation is completed by selecting a random gene along the size of the chromosomes and swapping most of the genes following that point. The preferred crossover selects any two solutions strings randomly from your mating pool many portion belonging to the strings is exchanged concerned with the strings. Also these are the best solutions. The choice point is selected randomly. A probability of crossover is usually introduced to be able to give freedom for an individual solution string to figure out or possibly a solution would go with crossover or otherwise not [26]. Out of all best solutions, first two solutions are taken over and their fitness is checked again to find their Accuracy, True Positive Rate, True Negative Rate, Precision and F_measure. If fitness of child is greater than fitness of parent then parent is parent is equal to child. Crossover generally checks the tuning of different parameters such as number of trees, depth of tree, height of tree, epsilon, seed, etc. The values of dataset that are used for training and testing is not changed rather the parameter tuning is changed each time. So, it is concluded that crossover is used to find the combinations of parameters. Crossover rate decides as of how many times crossover occurs during the time period algorithm performs. Crossover rate is computed because if it is not

computed and let it go as it is, it may result in decrease in convergence speed of algorithm. Hence, checking crossover rate do not allow the delay to be introduced

iii. **Mutation:** Mutation would be the occasional introduction of recent features in the remedy string belonging to the population pool to maintain diversity inside the population. Though crossover offers the main responsibility looking for the optimal solution, mutation is usually used for this purpose. Mutation takes only one solution and interchanges two parameters. Mutation operator changes a 1 to 0 or vice versa. If these is change in sequence, result obviously changes. This change in result may be positive or negative. If the result is positive, result is carried over otherwise remained same as earlier result. Because positive result means performance is high. Mutation rate decides as of how many times mutation occurs during the time period algorithm performs. Mutation rate is computed because if it is not computed and let it go as it is, it may result in decrease in computational speed of algorithm. Hence, checking mutation rate do not allow delay to be introduced.

1.4 Summary

Data mining is a wonderful mix of various smart methods. In this way, it is effectively connected to the conceivable area for creating enormous outcomes. Data mining is a gift to the associations which are coming up short on information and requirements, fulfilling their needs successfully. The fundamental quality of data mining lies in the hands of the ETL Process. Data mining has its underlying foundations from measurements, Machine Learning (ML), and numerous other potential regions. Subsequently, it is a reason that data mining is the one step answer for every one of the data issues regardless of their areas. For additional explore, there is a more extensive degree in the data mining applications while remembering distinctive difficulties looked amid the mining process. In real estate service that is a combination of three types of data analysis that ends up being a gift and as a fundamental piece of its data discovery as talked about in this section.

CHAPTER 2
LITERATURE SURVEY

This section portrays the detailed survey on descriptive, performance, predictive and prescriptive analysis. Literature reviews are written for the overview did for the following implementation work. These references speak to differed parts of the extensive variety of the data analytics area. From a most recent couple of decades, with an appearance of innovation in computers, numerous prestigious researchers had contributed fundamentally in the information mining area. To comprehend the issue space as far as its relevance, an assortment of procedures, and nature of data, a broad overview had been headed. The overview was begun with a more extensive vision by including the exploration papers from every unique area.

2.1 Literature review based on descriptive analysis

In this section, the survey of the papers from the prominent areas is incorporated. Each paper is overviewed with respect to abstract, objectives, problem definition, methodology, and conclusions. The primary goal of directing this review is to have an understanding of the well-known application areas of the data mining especially Machine Learning (ML). Here, the overview identified with Data Warehouse (DW), Extract Transform Load (ETL), On-Line Analytical Processing (OLAP), Cube Technology and Data Mining.

Ping A Y. et al. (2017) [28] described the OLAP bid in Enterprise Marketing Management (EMM) system. According to the author in [28], with the growth of budgetary globalization and extraordinary mechanism, marketing experienced insightful variations. The marketing process is an administration verdict creation process. The crucial aim of Enterprise Marketing Management (EMM) system is to figure out the data storeroom to deliver confirmation provision for the judge to make a verdict. Use of OLAP and MDX technology in the marketing management systems, apprehend the erection of multidimensional data and fact analysis.

Wang P. et al. (2014) [29] proposed a descriptive model that explained reinforcement costs shows assorted attributes with respect to power-driven vitality costs. While the power-driven energy costs had generally been considered in the writing, such investigation on supplementary facilities costs is incomplete. The utilization of trustworthy stochastic methodologies for representing the conduct of operating reinforcement costs in the power market. With de-control in power frameworks, helper facilities arcades have emerged to get these facilities through sensible sell-offs. In general, the cost of providing assistant facilities are lower than fabricating vitality. However, the arcade capacity for operational stores are generally inferior to energy market, the incomes as of retailing these facilities can be equivalent to energy. Hourly save costs are more eccentric than the day by day reinforcement costs. Despite the fact that just a single backup cost is arranged, the results can be drawn out to other backup costs that have indicated alike highlights, extraordinary erraticism, and rehashed substantial spikes.

He Z. et al. (2011) [30] proposed the application of OLAP and Data Warehouse (DW) in Merchandising System. In this paper, Data Warehouse (DW) of commodities sales was built on the basis of commodities sales system data resources, which further establishes commodities sales multidimensional datasets on which data analysis was conducted, which helped in significant decision making. On the basis of this system and using the analysis services, Data Warehouse (DW) was built and proposed On-Line Analytical Processing (OLAP) system for judgments, giving a suitable method for an enterprise to route gigantic volumes of information and to launch an operative verdict sustenance structure.

Haiyan H. et al. (2010) [31] discussed the idea of a multi-dimensional methodical approach based on On-Line Analytical Processing (OLAP). Also, it blends efficiently the development platform of Microsoft-Business Intelligence and trained the application of On-Line Analytical Processing (OLAP) technology in Auto Works Company. On-Line Analytical Processing (OLAP) technology does not just merely examine queries but also counted omnidirectional research manufacture and gathered these data to guide everyday difficulties and even estimate future undertakings based on predictable data. In this paper, unnaturally presentation of MS's BI development platform determines knowledge from a

Data Warehouse (DW) of AW's DW. With the On-Line Analytical Processing (OLAP) technology became progressively corporate, the range of its application has been discovered and prolonged progressively.

Wang Y. et al. (2010) [32] defined that the assemblage of Data Warehouse (DW) technology and On-Line Analytical Processing (OLAP) technology opens Data Warehouse (DW). This is the innovative way for Data Storage Systems (DSS). This paper examined the substantial knowledge of Data Warehouse (DW) and On-Line Analytical Processing (OLAP) and detected the application of On-Line Analytical Processing (OLAP) in apprentice enactment investigation using Microsoft SQL Server and Analysis Services. The authors determined the appropriate evidence by examining exhibited teaching value and enhancing teaching possessions for the decision maker. OLAP emphasized on the verdict building provision of verdict building executive. The Multi-Dimensional (MD) analysis model is suitable for analysis demand for Data Warehouse (DW) and is the most essential methodical factor to make the DW successful. This article applied Multi-Dimensional (MD) analysis model of Data Warehouse (DW) to apprentice enactment investigation of colleges and universities, which had appropriate manifestation ability and quicker analysis speed that fulfilled numerous analysis application demand of the apprentice enactment information.

Mitsujoshi S. et al. (2010) [33] proposed an Emotion Concept Model (ECM) which combines the desire and annoyance dimension and numerous feelings, to obtain a descriptive analysis on sentiment and mood in voice. For descriptive analysis, the main objective of this paper was to attain ample quantity of vocal sound data and to analyze the correlation between sentiments and moods and improving the accuracy of voice emotion recognition. By using prior filtering of desire and annoyance sentiments, moods will be categorized more accurately. It’s not always the case that desire and annoyance tally to each mood. There can be a case where basic negative moods are generated from desire sentiments and basic positive moods are generated from annoyance sentiments. Shunji Mitsujoshi et al. (2010) [34] explained that Emotion Concept Model recommends a conceivable link between desire and annoyance sentiments and numerous moods by defining the effect of sentiments on moods.

Zhao H. (2008) [34] explained that with the expansion of the scale of advanced teaching, surplus data about a set of courses chosen by learners had been produced. This paper analysed the application of Data Warehouse (DW) and On-Line Analytical Processing (OLAP) for the analysis of a set of courses selected by learners. This undertakes the design of DW about universities set of courses chosen by learners and applies ETL process upon a set of courses chosen by learner's data. This paper built multi-dimensional cube analysis data model by use of OLAP technology on the set of courses chosen by learner's data and captured the query and showed analysis of set of courses chosen by learner's multi-dimensional data, so that it could analyze the set of courses formation situation from various angles and assist the University teaching Data Storage Systems (DSS). This paper applied the STAR model of Data Warehouse (DW) to the analysis of a set of courses chosen by learners and determines that this model can mollify to various analytical application demands for a set of courses chosen by learner's information.

Quafafauetal M. et al. (2005) [35] finds OLAP application in knowledge Data Warehouse (DW). A profound analysis of Data Warehouse (DW), knowledge discovery and the necessities basic for profound assimilation has been done. The authors suggested a Data Warehouse (DW) model. Also, they proposed a STAR schema and established a consideration using operatives. OLAP analysis depends upon the observed data and a set of OLAP operators for restructuration and granularity. The main aim was to discover concealed outlines in data. The combination of knowledge into DW conduce to supplemented analysis circumstances where objects and their relations were unambiguously handled and visualized.

2.2. A literature review based on performance analysis

In this segment, the papers firm to Key Performance Indicators (KPI's) is exhibited. In the Real Estate domain, there is quite great probability to work with various attributes of data. In the extent of the real estate mining, generally, some important features are basically considered. However, data mining is focusing on all the features except feature selection that standout amongst the most encouraging and nesting territory. It requires a most extreme level of exactness and accuracy. In addition, the real estate data is a rich source of hidden designs, whose extraction could be a standout amongst the most

fascinating features of performance mining. It could be considered as extraordinary if feature selection is done to the accurate level of mining.

He Y. et al. (2018) [36] discovered the sanctioned relationships/correlation amongst topical and topological data in archive systems. Record organized was a sort of captivating data-set which can give both relevant and topological data. A crucial point in displaying such data-sets was to find legitimate denominators underneath the content and connection. Generally, past work presented that records were firmly connected with each other that offer normal inactive themes. In any case, the propensity to connect to various hubs was ignored, which was unavoidable in informal communities. Consolidating network recognition and theme displaying in a unified structure and advance to Canonical Correlation Analysis (CCA), to catch the idle semantic relationships between the two heterogeneous elements, network, and subject. Regardless of the homophily or heterophily, Canonical Correlation Analysis (CCA) can appropriately catch the innate relationships which the dataset itself without any earlier speculation. This paper proposed a novel model Canonical Topical-Topological Analysis (CTTA) which simultaneously performed subject displaying and network identification on record systems. The presentation of CCA makes the model fit to investigate relationships between the heterogeneous points and networks. Complete assessments on three diverse datasets demonstrated that CTTA beats condition of-the-workmanship baselines with significant enhancements. The use of CCA likewise offers motivations for thinks about which might want to find relationships between two arrangements of factors in different fields.

Kadar J. et al. (2017) [37] analyzed the factors that enhanced the superiority of intranet website grounded on WebQual methodology, a case study of an organization by applying ANOVA test. Enhancing the nature of sites, particularly intranet ought to be performed ceaselessly, on the grounds that intranet is an intermediate of announcement between workers in an association. Client's observations on the nature of intranet in an organization should be estimated towards deciding the level of client fulfillment. The strategy utilized as a part of this examination was an overview technique with the WebQual approach. Examination was performed purposively in some work units in an organization as testing. WebQual approach utilized comprised of 23 estimations and was isolated into 3 measurements. To be specific ease of use, data quality, and administration

cooperation quality. The outcomes appeared by ANOVA test effectively affect the nature of intranet site in an organization. In singular constraint test, usability and service interaction factors had a huge effect on the nature of intranet, yet data quality constraint had no huge impact (negative). The measurable computation utilizing ANOVA was closed for three factors that are ease of use, data quality and administration collaboration quality, which have noteworthy constructive outcomes at the same time on site quality. This exploration contributed towards upgrading site in an organization as well as recommendations at the interval of redesigning on behalf of changes in specific chunks (quality data), as a consequence of examination.

Qiao F. et al. (2016) [38] analyzed the correlation and visualization of a large-scale data set that is the Global Data on Events, Location, and Tone (GDELT). GDELT is a constant huge scale database of worldwide human culture for open research which screens the universes communicate, print, and web news since 1979, making a free open stage for figuring on the whole world. In this paper, in the 1st place, an information crawler was composed and actualized, which gathers metadata of the GDELT database progressively and stores them in the Hadoop Distributed File System (HDFS). A hash-based technique was proposed after this to obtain in touch the "Occasion" table, "Notices" table and "GKG" table in GDELT, to process each data of each occasion. The GKG table associates every person, association, area, tally, subject, news source, and occasion in the world into a solitary monstrous system that catches what's occurring throughout the world, what its setting is and who's included, and how a planet feels about it, each and every day. Till date the information size of GKG table is around 277 million records.

Tan Z. et al. (2011) [39] hypothetically examined the relationship between Chinese land costs and bank advances, at that point in view of quarterly information of the Chinese date-book, exact test is conducted on the cooperation through the Vector Auto Regression (VAR) demonstrate, and gave the quantifiable portrayal around the correlation through co-integration test, as well as the commitment proportion of one gathering to additional through drive reaction capacity and fluctuation deterioration. Finally concurring to the aftereffects of experimental examination. High costs bought the tracking download to customers and furthermore bought gigantic financing challenges to the land undertakings.

Land business in generally was exceptionally restricted in its private capital, as there was essential rise in the huge capital required through bank advances, securitization and land trusts for the land improvement. As an imperative resource frame, the uncertainties of land value influence the aggregate sum of bank credit extension by influencing the request and supply for bank credit. From the viewpoint of land costs influencing the bank credit request, land value vacillations will have an effect on the credit request of reduced scale budgetary elements. Most importantly, land value changes will influence the occupants on the interest for bank credits.

Qixun F. et al. (2010) [40] abbreviated and inspected the current essential assessment methods for land extends, and dissected its assessment list framework. In light of these, utilizing dark connection demonstrate, assessment systems of the land ventures were planned and broke down the procedure of the case. Lately, as macro-monetary resistor of genuine bequest industry and the finish of tremendous benefit, quantitatively assessing the land ventures winds up real worry for numerous experts. The current routine with regards to land venture assessment process is, for the most part, accomplished through possibility considers. However, in hypothetical viewpoints, it is scarcely to discover numerous approaches aside from the AHP-based assessment strategy, fuzzy hypothesis based assessment technique, and multi-criteria examination assessment technique. In light of the present assessment record framework, present outside markers of land undertakings to enhance the assessment arrangement of the land venture and to make it more logical, judicious, not just for the assessment of the task itself has down to earth hugeness, yet in addition, the planning of the venture has awesome noteworthiness. As land venture is a high speculation, exceptional yield, high-hazard venture, to build up a complete and levelheaded assessment record framework for the assessment of land ventures is the main need.

2.3 Literature Review based on predictive analysis

In this area, the papers identified with predictive analysis are studied. As per literature predictions are more important. From these, J48 and Support Vector Machine had been chosen as models for predictions. Related literature is as follows:

Singh A. et al. (2016) [41] explained a contrast between classification procedures on different dataset approaches using Rapid Miner. This paper talked about various characterization systems with little and expansive datasets. The two datasets were illustration datasets utilized from vault destinations relying on the number of instances. These datasets were connected in various classifier like, Decision Tree, Random Forest and Naive Bayes to recognize the best classifier for a little dataset and substantial dataset. This paper gave the examination and investigation of different strategies utilized for expectation. In view of the investigation, Naïve Bayes is most appropriate for little datasets and Decision Tree is reasonable for vast datasets in light of the assessment done in this paper utilizing different systems driven by Rapid Miner while likening exactness, review, and precision.

Kaur G. et al. (2014) [42] enhanced J48 procedure for predicting the diabetes. This examination work manages a proficient data mining method for predicting diabetes from medicinal records of patients. The diabetes data-set was utilized which gathered the data of patients with and without having diabetes. The altered J48 classifier was utilized to build the precision rate of the data mining system. The data mining apparatus WEKA had been utilized as an Application Program Interface (API) of MATLAB so as to create the J-48 classifiers.

Rahman R. et al. (2013) [43] examined numerous classification procedures using diverse data mining tools for diagnosis of diabetes. Without medicinal finding confirmations, it was troublesome for the specialists to speak out about the review of illness with confirmation. Numerous examinations were done that include arrangement of expansive scale information. Because numerous tests confuse the primary conclusion process and prompt the trouble in getting the final products. This sort of trouble could be settled with the guide of Machine Learning (ML) strategies. In this examination, Rashedur M. Rahman in [41] introduced the investigation of various arrangement procedures utilizing three different data mining tools. This paper was to investigate the execution of various grouping systems so as to arrange expansive information.

Vaithiyanathan V. et al. (2013) [44] gave a correlation of different classification techniques utilizing different datasets. In this paper distinctive grouping strategies of Data

Mining were analyzed utilizing different datasets from the College of California, Irvine (UCI). Precision and time required for execution by every method were noticed. This work had been done to make an execution assessment of different classification algorithms. Naive Bayes calculation depends on the view of likelihood and j48 calculation depends on the decision tree. The paper embarks to make near assessment of classifiers with regards to Work, Soybean and Weather datasets. The tests were done utilizing weka 3.6. The outcomes in the paper show that the proficiency of j48 and Naive Bayes is great.

Weis M. et al. (2009) in [45] explained the correlation of various grouping calculations for weed location from pictures in light of shape parameters. Fluctuation of weed pervasion had been evaluated for particular site weed administration. Since testing weed manually is excessively tedious for practical applications, a framework for programmed weed testing was produced. In this paper, Martin Weis et al. in [45] assess distinctive characterization calculations with the principle center around k-closest neighbors, choice tree learning and Support Vector Machine classifiers. Execution measures for grouping precision were assessed by utilizing cross-validation procedures and by contrasting the outcomes and physically surveyed weed invasion.

Tezel S. et al. (2009) [46] enhanced SVM on imbalanced datasets in distance spaces. Imbalanced instructive accumulations acquaint a particular test with the data mining system. Generally, it acts as the remarkable event of interest and the cost of misclassifying the extraordinary occurrence is higher than misclassifying the normal occurrence. When the data is outstandingly twisted towards the standards, it could be to a great degree troublesome for learning arrangement to accurately recognize the extraordinary event. There have been various strategies to deal with imbalanced educational lists, from under-inspecting the larger part class to adding designed concentrations to the minority class in incorporate space. Divisions between the courses of action are known to be non-Euclidean and nonmetric since differentiating time game plan required mutilating in time. This reality made it hard to apply standard procedures like SMOTE for installing built data centers in incorporate spaces. This paper showed a creative approach that grew the minority class by incorporating made concentrations in expel spaces. Exploratory results on standard time plan exhibit that built concentrations

upgrade the portrayal rate of the phenomenal events and a great part of the time also improve the general precision of SVM.

Tang Y. et al. (2009) [47] gave correspondence Support Vector Machine (SVM) modeling for extremely unfair classification. Conventional arrangement calculations can be constrained in their execution on very lopsided informational collections. A well-known stream of work for countering the issue of class awkwardness has been the application of a sundry of testing procedures. In this correspondence, Yuchun Tang et al. (2009) [48] centered for planning changes in Support Vector Machines (SVMs) to suitably handle the issue of class awkwardness. The authors joined extraordinary rebalance heuristics in SVM demonstration, including cost-touchy learning, finished and under sampling. These SVM based procedures were looked at with different cutting edge approaches on an assortment of informational indexes by utilizing different measurements, including G-mean, zone under the beneficiary working trademark bend, F-measure, and zone under the exactness/review bend. Specifically, of the four SVM varieties considered in this correspondence, the novel granular SVMs under sampling calculation was the best as far as both adequacy what's more, proficiency.

Wei-wu Y. et al. (2009) [48] uses Support Vector Machines and Least Squares Support Vector Machines to analyze coronary illness. Nonlinear classifiers calculations of standard SVM and LS- SVM are examined. At that point, standard SVM nonlinear classifiers and LS SVM nonlinear classifiers were connected to analyze coronary illness in light of UCI benchmark informational collection. Contrasting another outcome, the high precision rate was gotten in the forecast. Use of SVM and LS SVM to analyze sickness demonstrates that SVM and LS SVM had potential application.

Zhao C. et al. (2006) [49] utilized SVM for predicting the destructive actions of various datasets. SVM was connected to the expectation of lethality of various informational collections contrasted and other two basic techniques, Multiple Linear Regression and RBFNN. Quantitative Structure-Activity Relationships (QSAR) models in view of computed atomic descriptors have been settled. Among them, SVM demonstration gave the most noteworthy relationship coefficient R. It demonstrates that the SVM performed preferable speculation capacity over the MLR and RBFNN strategies, particularly in the

test set and the entire informational index. Thus, in the long run, prompts preferred speculation over neural systems, which actualize the exact hazard minimization guideline and may not join to worldwide arrangements. The authors in [49] expect SVM technique as an intense instrument for the expectation of sub-atomic properties.

Rokacha L. et al. (2005) [50] worked on Decision trees (DTs). DTs were thought to be standout amongst the utmost prevalent methodologies for rephrasing classifiers. Specialists commencing different areas such as Statistics, ML, design acknowledgment, and DM had managed the issue of growing a decision tree from accessible information. This paper introduced a refreshed survey of current techniques for building decision tree classifiers in the best way. This paper recommended a unified algorithmic structure for showing these calculations and depicted different fragment criteria and pruning methodologies.

2.4 Literature review based on Prescriptive Analysis

In this area, the papers identified with prescriptive analysis on different domains are studied. From this literature, the variety of models were found for performing optimizations. As per guidance from papers Genetic Algorithm (GA) and their variants NSGA-I, NSGA-II and NSGA-III were chosen to carry forward for research. Following papers cover the related part.

Li K. et al. (2018) [51] Flip chip innovation had been generally utilized as a part of IC bundling, and the blend of this innovation and weld joint interconnection innovation had been used in the assembling of electronic gadgets all around. With the advancement of flip chip towards high thickness and ultra-fine pitch, the examination of flip chips is stood up with incredible difficulties. This paper built up an insightful framework towards utilizing the discovery of flip chips in light of vibration. Thirty-four highlights including 18-time-space highlights and 16 recurrence area highlights were extricated from the crude vibration information. The Support Vector Machines were utilized to execute the acknowledgment and arrangement of flip chips. With a specific end goal of enhancing the grouping exactness of SVM, cross approval and hereditary calculation were utilized to enhance the parameters of SVM individually. SVM, CV-SVM, and GA-SVM were connected to order independently and the outcomes were acquired. By correlation, GA-

SVM could perceive and group the flip chips quickly with high exactness. In this way, GA-SVM act as a powerful mechanism for the imperfection investigation of flip chips.

Hong J. et al. (2018) [52] one of the real difficulties of taking care of Big Data enhancement issues by means of conventional multi-objective transformative calculations (MOEAs) is their high computational expenses. This issue has been proficiently handled by non-ruled arranging hereditary calculation, the third form, (NSGA-III). Then again, a worry about the NSGA-III calculation is that it utilizes a settled rate for change administrator. To adapt to this issue, this investigation acquaints a versatile transformation administrator to improve the execution of the standard NSGA-III calculation. The proposed versatile transformation administrator system is assessed utilizing three hybrid administrators of NSGA-III including reenacted paired hybrid (SBX), uniform hybrid (UC) and single point hybrid (SI). In this manner, three enhanced NSGA-III calculations (NSGA-III SBXAM, NSGA-III SIAM, and NSGA-III UCAM) are created. These upgraded calculations are then actualized to settle various Big Data streamlining issues. The trial comes about to show that NSGA-III with UC and versatile transformation administrator outflanks the other NSGA-III calculations.

Hu C. et al. (2018) [53] expressed that contaminant occasions in drinkable Water Dissemination Systems (WDS) had happened as often as possible lately, causing serious harms, financial misfortune, and dependable societal effects. A basic and viable technique to screen WDS continuously remained as conveying a water quality sensor. The position of such sensors in a water appropriation organization had turned into a principal worry all over. Initially examining sensor position numerically and demonstrating it is NP-hard problem. Consequently, single-and multi-target advancement were recognized, and proposed a changed NSGA-III to illuminate many-target streamlining for the sensor situation issue. WDS of two sizes were utilized and recreation came about to show the legitimacy and adequacy of the proposed model. The future research works were likewise recognized and examined.

Elarbi M. et al. (2018) [54] examined that, as of late, decay has picked up a wide enthusiasm for taking care of multi-target improvement issues including in excess of three goals otherwise called Many-target Optimization Problems (MaOPs). Over the most

recent couple of years, there had been numerous recommendations to utilize deterioration to tackle unconstrained issues. In any case, the measure of works that had given to propose new disintegration based calculations to take care of obliged many-target issues. This paper proposed the Isolated Solution-based Constrained Pareto Power ISC-PC connection that could: (1) handle obliged many-target issues portrayed by various sorts of troubles and (2) supported the determination of not just infeasible arrangements related to disconnected sub-districts yet, in addition, infeasible arrangements with littler Constraint Violation (CV) values. The imperative taking care of system had been incorporated into the structure of the Constrained Non-Dominated Sorting Genetic Algorithm-III (C-NSGA-III) to deliver another calculation called Isolated Solution-based Constrained Non-Dominated Sorting Genetic Algorithm-III (ISC-NSGA-III). The observational outcomes had exhibited that requirement taking care of methodology could give better and aggressive outcomes when thought about against three proposed compelled deterioration based many-objective transformative calculations. Also, the adequacy of Isolated Solution-based Constrained Non-Dominated Sorting Genetic Algorithm-III (ISC-NSGA-III) on a true water administration issue was exhibited.

Chahardoli S. et al. (2018) [55] examined that a punctured topped end funnel-shaped steel safeguard was explored to upgrade its divider thickness and opening stature. The openings were worked out on the edge of the safeguard to bring down pinnacle compel at fall. For this reason, once completed with reenacting the safeguard using LS-Dyna programming. Further, checking the reproduced display utilizing test information, opening stature and divider thickness of the safeguard were upgraded to accomplish most extreme vitality assimilation alongside least pinnacle compel. A sum of 96 unique cases was mimicked, of which 7 cases were endangered as trial tests. The streamlining was executed utilizing Non-Dominated Sorting Genetic Algorithm – III (NSGA-III) calculations which were executed in MATLAB programming. Reaction surface approach were utilized to decide input capacities for these calculations. At last, ideal position for the gaps in cone like safeguards was observed to be the closest point to the upper base of the truncated cone. A generally decent assertion was seen between the aftereffects of Non-Dominated Sorting Genetic Algorithm-III (NSGA-III) calculations, and the

calculations could foresee ideal divider thickness and opening position at a worthy exactness at times.

Zhu Y. et al. (2017) [56] highlighted that determination could enhance order precision and diminishing the computational intricacy of arrangement. Information includes an interruption location frameworks constantly to introduce an issue of the imbalanced order in which a few characterizations just have a couple of cases while others have numerous occurrences. This irregularity could clearly restrain order effectiveness, yet a couple of endeavors had been made to address it. In this paper, a plan for the many-target issue was proposed for highlighting choice in IDS, which utilizes two systems, in particular, an exceptional mastery technique and a predefined technique focused on seek, for populace development. It could separate movement amongst ordinary and strange as well as by variation from the norm write. In view of the plan, Non-dominated Sorting Genetic Algorithm-III (NSGA-III) had been utilized to acquire a sufficient element subset with great execution. An enhanced many-target streamlining calculation I-NSGA-III was additionally proposed utilizing a novel specialty protection technique. It comprised of an inclination choice process that chose the person with the least chose highlights and a fit-choice process that chooses the person with the greatest entirety weight of its targets. The trial comes out to demonstrate that I-NSGA-III could ease the unevenness issue with higher grouping precision for classes having fewer occasions. Additionally, it could accomplish both higher characterization precision and lower computational unpredictability.

Tavana M. et al. (2016) [57] proposed X-bar control diagrams which were broadly used to screen and control business and assembling forms. This investigation considered an X-bar control graph outline issue with different and regularly clashing destinations, including the normal time the procedure stays in measurable control status, the sort I blunder, and the discovery control. A coordinated multi-target calculation was proposed for improving the prudent control outline plan. Further connected multi-target enhancement strategies were established on the reference-focus based on Non-dominated Sorting Genetic Algorithm-III (NSGA-III) calculation to productively tackle the improvement issue. At that point, Data Envelopment Analysis (DEA) was utilized to decrease the quantity of Pareto ideal answers for a reasonable size. Four DEA techniques

analyzed the ideal arrangements in light of relative proficiency. A few measurements were utilized to think about the execution of the Non-dominated Sorting Genetic Algorithm – III (NSGA-III). Moreover, the Data Envelopment Analysis (DEA) technique was utilized to think about the execution of Non-dominated Sorting Genetic Algorithm – III (NSGA-III). A notable contextual investigation was figured and settled to show the pertinence and display the viability of the proposed enhancement calculation. Furthermore, a few numerical cases were created to think about the Non-dominated Sorting Genetic Algorithm – III (NSGA-III). Results demonstrate that NSGA-III performs better in creating effective ideal arrangements.

2.5 Summary

This Chapter has talked about how data analysis is done. The writing has been inspected for a variety of systems. To work upon the system thoroughly, the review was divided into five parts. The first part explains how to extract data, clean data, store data in the warehouse and finally cube creation using a variety of tools. The second part explains how to extract features from the bulk of data and finding a relationship between the features. The third part explains how to forecast. Finally, the fourth part explains how to optimize.

CHAPTER 3

PROBLEM FORMULATION

This section portrays the gaps found after surveying the literature and helps in framing the objectives

3.1 Gaps in the literature

The extensive review has shown that the existing techniques suffer from various issues.

1. Presently OLTP projects are available that support flat files but Descriptive, Predictive or Prescriptive analytics have been ignored upon Real Estate Data.
2. The issue of analyzing and querying large dataset of Real Estate in a multidimensional structure or Cube view is not yet resolved.
3. Visualizations or dashboards of the same are yet not created.
4. Both the algorithms, J48 and SVM are not applied on Real Estate dataset, especially for write decision making.
5. Poor convergence speed is neglected by the most of the existing research in the field of Genetic Algorithm.
6. Due to poor convergence speed, most of the existing literature suffered from local optima.
7. Chances of suffering from premature convergence were seen in most of the works of literature.

3.2 Problem Definition

This research work focuses on the prediction of real-estate data. As known in prior, many persons purchased and sale property day by day. Some persons purchased it for living purpose whereas other purchased it for investment purpose. However, purchasing a given property depends upon various factors such as length, breadth, facing, locality, type of property etc. Therefore, it becomes difficult for the buyer whether to buy it or not. Therefore, in this work, we have taken real-estate as a classification problem of Machine Learning (ML). To achieve the designed objectives, the initial data set has been collected and analyzed using OLAP. Thereafter, J48 and SVM are applied to the collected data.

Although, these models perform significantly to classify real estate purchasing, but, suffer from the parameter tuning issue. Therefore, in this work, this issue has been handled by considering the well-known meta-heuristic optimization technique i.e., NSGA-III. It iteratively optimizes the meta-J48 model to improve the classification rate by considering mutation and crossover operator. The obtained solutions are non-dominated in nature, therefore, the proposed model can provide better accuracy as well as other parameters concurrently.

3.3 Objectives

After conducting the review of existing techniques, the following objectives have been formulated:

1. To apply the ETL process on real estate data, creating a multidimensional cube structure out of drawn homogeneous data and finally creating dashboards.
2. To evaluate the classification rate of real estate data using J48 and Support Vector Machine (SVM)
3. To propose Non-dominated Sorting Genetic Algorithm – III (NSGA-III) based Meta-Decision tree for real estate data using various parameters such as Accuracy, True Positive Rate, True Negative Rate, Precision and F_Measure.

3.4 Research Methodology

This work focuses on analyzing and classifies the real-estate dataset for classifying whether the property is purchased or not, using the cube technology and visualizing these progressions in the form of dashboards. After the creation of cubes, classification has been done by using J48 and SVM based classification models. Further, NSGA-III based technique is also applied to optimize the meta-decision tree to improve the accuracy rate. To accomplish the proposed work, the below steps have been followed:

i) Conducting literature survey of descriptive, predictive and prescriptive analytics from which various ideas such as cube technology, classification, predictive and optimization models in the field of machine learning is studied. Based upon the review, descriptive, predictive and prescriptive techniques are used to analyze the real-estate dataset.

ii) Feature selection is also performed before applying analytics on dataset. It is achieved by analyzing the Key Performance Indicators (KPIs). For extracting KPIs, the two tests are used including correlation test and ANOVA test by using MATLAB 2013a tool with statistics and machine learning toolbox. These tests are performed on candidate KPIs and condensed list of KPIs are obtained.

iii) The cube technology is applied by using Business Intelligence Development Studio which is an inbuilt tool in Visual Studio 2012, on real-estate dataset and cubes are generated for visualization.

iv) Further the classification analysis are also used to classify the real-estate dataset by using SVM and J48 techniques.

v) Additionally, to tune the parameters of the meta-decision trees, NSGA-III based optimization technique is also used. The main objective of NSGA-III is to tune the initial parameters required by meta-decision tree.

vi) The results of various models are also compared on the basis of their accuracy and the model with highest accuracy is used for making predictions of real-estate data.

CHAPTER 4

IMPLEMENTATING DESCRIPTIVE DATA ANALYSIS

This section portrays the idea of describing and mining data to create useful patterns out of On-Line Transactional Processing (OLTP) Systems.

Real estate systems hold a lot of information that can be mined for useful patterns. Various features are collected in whole to analyze the Real Estate domain. So, in order to define all the features of real estate, the dataset primary dataset is created and the secondary dataset is taken from Kaggle website [59]. Statistical Techniques are applied to evaluate the best feature for classification of this data. Analysis Of Variance (ANOVA) and Correlation Coefficients are used to evaluate the key features in the database. The main objective of Key Performance Indicators (KPIs) is to select the most relevant quantitative factors and finding the various key features such as Length, Breadth, Area, Collector rate and Commission per sale with House price index attribute. Extensive analysis has been done to evaluate a set of metrics that enable the performance management of a real estate system.

Descriptive Analytics normally convey reports and dashboards that give operational knowledge into the business and specially appointed inquiries and investigation give further responses to inquiries regarding it. The advancements supporting these requirements are called Business Intelligence or BI and they center on what has occurred in the business. Graphic Analytics offer some benefit by recognizing issues requiring consideration, yet simply after they have happened. The most imperative specialized part of the spellbinding examination is the recognizable proof of Key Performance Indicators (KPIs) and measurements used to assess business execution.

KPI is a kind of execution measurement. KPIs assess the achievement of an association or of a specific movement in which it locks in. Frequently achievement is basically the rehashed, intermittent accomplishment of a few levels of operational objective and some of the time achievement is characterized as far as gaining ground toward vital goals. Accordingly, picking the privilege KPIs depends upon a decent comprehension of what is essential to the organization. What is regarded as critical frequently relies upon the

division estimating the execution? Organizations frequently misperceive metrics and KPIs. A metric is just a measure while at the same time a KPI like productivity is on a very basic level basic to a business in light of the fact that they address them in a neighborhood setting instead of a venture wide one.

Customer_	Customer_	Customer_	Customer_	Contact_Numb	Customer_	Property_i d	Property_t ype	Property_ descriptio n	House_N umber	Property_ address	Property_c	Property_f acing	Property_ neighbour s	Property_n	No_of_ow ners	Gender
1	Harjit Singl	suman aror	Jalandhar	9814537071	Seller	1	constructe	shop	1	adarsh nag	Jalandhar	north	RLB	commercial	1	F
2	Nikita Anar	Kansal Kar	Patiala	9356125444	Buyer	2	constructe	showroom	233	bachittar na	Patiala	east	RL	commercial	2	F
3	Bhaavna A	Konica Res	Phagwara	8558823502	Rental	3	constructe	office	5	adarsh nag	Jalandhar	north	RB	commercial	1	F
4	Upendrana	Pindi Paint	Ludhiana	9818056405	Seller	4	constructe	kothi	688	shaheed bl	Jalandhar	west	LB	residential	1	M
5	Charanjit S	Sai Dhaba	Amritsar	9878687136	Buyer	5	constructe	shop	55	kairon mark	Amritsar	north	RLB	commercial	1	M
6	Diljit Singh	kanchan cr	Jalandhar	9872186026	Rental	6	constructe	flat	57	adarsh nag	Jalandhar	north	RL	residential	2	M
7	Jai Lal Ubh	Bhagwan K	Patiala	9835511101	Seller	7	constructe	kothi	95	bachittar na	Patiala	south	RB	residential	2	M
8	Hoon Balal	Krishna Ca	Phagwara	8195910007	Buyer	8	constructe	showroom	65	central tow	Jalandhar	north	LB	commercial	1	M
9	Chuni Lal E	United Che	Ludhiana	9814000003	Rental	9	constructe	office	565	shaheed bl	Ludhiana	east	RL	commercial	1	M
10	K. D. Bhan	Hotel C J Ir	Amritsar	9781333962	Seller	10	constructe	kothi	535	kairon mark	Amritsar	north	RB	residential	2	F
11	Jagbir Sing	paras silk s	Jalandhar	9464213275	Buyer	11	constructe	kothi	6	adarsh nag	Jalandhar	west	LB	residential	1	F
12	Banjinder S	Ashwani K	Patiala	9915200049	Rental	12	constructe	flat	8	aman vihar	Patiala	north	RLB	residential	3	F
13	Hans Raj H	Delhi Paint	Ludhiana	9877900786	Seller	13	constructe	factory	98	industrial a	Ludhiana	north	RL	industrial	5	M
14	Sheikh Am	Khurana St	Phagwara	9463841121	Buyer	14	constructe	shop	59	guru gobin	Jalandhar	south	RB	commercial	2	M
15	Rajinder Jo	Hotel Paga	Amritsar	9914023121	Rental	15	constructe	showroom	15	ajit nagar	Amritsar	north	LB	commercial	1	M
16	Karan Jotw	fancy empo	Jalandhar	7087872818	Seller	16	non-constr	plot	56	central tow	Jalandhar	east	RLB	commercial	4	M
17	Surbhi Cho	Bhandari C	Patiala	9152700001	Buyer	17	non-constr	plot	17	bank colon	Patiala	north	RL	residential	2	M
18	Manmohar	Hotel Base	Amritsar	9876096155	Rental	18	non-constr	plot	29	ajit nagar	Amritsar	west	RB	residential	1	M
19	Manoranja	Mohan Sw	Phagwara	9915137755	Seller	19	constructe	shop	19	green park	Jalandhar	north	LB	commercial	1	F
20	Madhur Ka	Star Copier	Ludhiana	8427712105	Buyer	20	constructe	kothi	89	kabir colon	Ludhiana	north	RLB	residential	2	F
21	Ram Kapoo	savita silk s	Jalandhar	9915789336	Rental	21	non-constr	plot	21	defence co	Jalandhar	south	RL	commercial	2	F
22	Avinash K	Happy Bro	Patiala	9915400039	Seller	22	constructe	kothi	96	bharpur ga	Patiala	north	RB	residential	1	F
23	Avneet Ka	Sharma Tra	Ludhiana	9815100595	Buyer	23	constructe	flat	8	inder nagar	Ludhiana	north	LB	residential	1	M
24	Gurkanwal	Handa Jew	Phagwara	9501200366	Rental	24	constructe	shop	69	central tow	Jalandhar	north	RL	commercial	2	M
25	Fateh Ali K	crazy bouti	Jalandhar	9872116570	Seller	25	constructe	showroom	25	guru gobin	Jalandhar	east	RB	commercial	1	M

Descripti on_id	Length (ft)	Breadth (ft)	Area_in_ sqft	Area_in_ marla	Area_in_ guz	Area_in_ sqinches	Area_in_ sarsahi	Area_in_ acre	Major_d escriptio n_id	No_of_r ooms	No_of_w ashroom s	No_of_ki tchen	No_of_s toreroom s	Dining_r oom	Drawing _room	Garage	Garage_c apacity	Garden	Basemen t	Basemen t_type	Usage	Electrici y_supp y
1	24	75	1800	8	200	259200	72	0.0396	1	2	2	1	2	1	1	lentered	2	Yes	Yes	Refurbish	Yes	Yes
2	20	63	1260	5.6	140	181440	50.4	0.02772	2	1	1	1	1	1	1	not-lenter	1	No	No	Nil	No	No
3	30	62	1860	8.266667	206.6667	267840	74.4	0.04092	3	2	2	1	2	1	1	lentered	2	Yes	Yes	New base	Yes	Yes
4	34	74	2516	11.18222	279.5556	362304	100.64	0.055352	4	1	1	1	1	1	1	not-lenter	1	No	No	Nil	No	No
5	40	78	3120	13.86667	346.6667	449280	124.8	0.06864	5	2	2	1	2	1	1	lentered	2	Yes	Yes	Deep bas	Yes	Yes
6	50	83	4150	18.44444	461.1111	597600	166	0.0913	6	1	1	1	1	1	1	not-lenter	1	No	No	Nil	No	No
7	50	96	4800	21.33333	533.3333	691200	192	0.1056	7	2	2	1	2	1	1	lentered	2	Yes	Yes	New gard	Yes	Yes
8	24	75	1800	8	200	259200	72	0.0396	8	1	1	1	1	1	1	not-lenter	1	No	No	Nil	No	No
9	20	63	1260	5.6	140	181440	50.4	0.02772	9	1	1	1	1	1	1	lentered	1	Yes	Yes	Retrofit	Yes	Yes
10	30	62	1860	8.266667	206.6667	267840	74.4	0.04092	10	3	3	1	3	1	1	nil	0	No	No	Nil	No	No
11	34	74	2516	11.18222	279.5556	362304	100.64	0.055352	11	1	1	1	1	1	1	nil	0	No	No	Nil	No	No
12	40	78	3120	13.86667	346.6667	449280	124.8	0.06864	12	2	2	1	2	1	1	nil	0	No	No	Nil	No	No
13	50	83	4150	18.44444	461.1111	597600	166	0.0913	13	2	2	1	2	1	1	lentered	1	No	No	Nil	No	No
14	50	96	4800	21.33333	533.3333	691200	192	0.1056	14	1	1	1	1	1	1	nil	0	No	No	Nil	No	No
15	24	75	1800	8	200	259200	72	0.0396	15	2	2	1	2	1	1	nil	0	No	No	Nil	No	No
16	20	63	1260	5.6	140	181440	50.4	0.02772	16	1	1	1	1	1	1	nil	0	No	No	Nil	No	No
17	30	62	1860	8.266667	206.6667	267840	74.4	0.04092	17	2	2	1	2	1	1	nil	0	No	No	Nil	No	No
18	34	74	2516	11.18222	279.5556	362304	100.64	0.055352	18	1	1	1	1	1	1	lentered	1	Yes	Yes	New base	Yes	Yes
19	40	78	3120	13.86667	346.6667	449280	124.8	0.06864	19	2	2	1	2	1	1	lentered	2	Yes	Yes	Retrofit	Yes	Yes
20	50	83	4150	18.44444	461.1111	597600	166	0.0913	20	1	1	1	1	1	1	not-lenter	1	Yes	Yes	Deep bas	Yes	Yes
21	50	96	4800	21.33333	533.3333	691200	192	0.1056	21	2	2	1	2	1	1	lentered	2	Yes	Yes	Refurbish	Yes	Yes
22	24	75	1800	8	200	259200	72	0.0396	22	2	2	1	2	1	1	not-lenter	2	Yes	Yes	New gard	Yes	Yes
23	20	63	1260	5.6	140	181440	50.4	0.02772	23	1	1	1	1	1	1	lentered	1	Yes	Yes	New base	Yes	Yes
24	30	62	1860	8.266667	206.6667	267840	74.4	0.04092	24	2	2	1	2	1	1	lentered	2	Yes	Yes	Retrofit	Yes	Yes
25	34	74	2516	11.18222	279.5556	362304	100.64	0.055352	25	1	1	1	1	1	1	lentered	1	Yes	No	Nil	Yes	Yes

Measures _id	Length (ft)	Breadth (ft)	Area_in_ sqft	Area_in_ marla	Area_in_ guz	Area_in_ sqinches	Area_in_ sarsahi	Area_in_ acre	Collector _rate_per _marla	Price_per_	Price_clos	Price_of_p	Earnest_m	Registry_a	StampDuty	Time_take	Commissio	Commissio	Registratio	HousePric
1	24	75	1800	8	200	259200	72	0.0396	595000	5712000	565250	5426400	542640	4883760	217056	2	1	108528	108528	29.6724
2	20	63	1260	5.6	140	181440	50.4	0.02772	45900	308448	43605	293025.6	29302.56	263723.04	11721.024	2	2	11721.024	5860.512	1.6023
3	30	62	1860	8.2666667	206.66667	267840	74.4	0.04092	595000	5902400	565250	5607280	560728	5046552	224291.2	6	1	112145.6	112145.6	30.6615
4	34	74	2516	11.182222	279.55556	362304	100.64	0.055352	12600	169075.2	11970	160621.44	16062.144	144559.3	9637.2864	2	2	6424.8576	3212.4288	0.8783
5	40	78	3120	13.866667	346.66667	449280	124.8	0.06864	86400	1437696	82080	1365811.2	136581.12	1229230.1	81948.672	3	1	27316.224	27316.224	7.4685
6	50	83	4150	18.444444	461.11111	597600	166	0.0913	297500	6584666.7	282625	6255433.3	625543.33	5629890	375326	2	2	250217.33	125108.67	34.2057
7	50	96	4800	21.333333	533.33333	691200	192	0.1056	9647	246963.2	9164.65	234615.04	23461.504	211153.54	14076.902	2	1	4692.3008	4692.3008	1.2829
8	24	75	1800	8	200	259200	72	0.0396	255000	2448000	242250	2325600	232560	2093040	139536	6	1	46512	46512	12.7167
9	20	63	1260	5.6	140	181440	50.4	0.02772	37600	252672	35720	240038.4	24003.84	216034.56	14402.304	2	0	0	4800.768	1.3126
10	30	62	1860	8.2666667	206.66667	267840	74.4	0.04092	23550	233616	22372.5	221935.2	22193.52	199741.68	8877.408	3	1	4438.704	4438.704	1.2136
11	34	74	2516	11.182222	279.55556	362304	100.64	0.055352	297500	3992053.3	282625	3792450.7	379245.07	3413205.6	151698.03	2	2	151698.03	75849.013	20.7377
12	40	78	3120	13.866667	346.66667	449280	124.8	0.06864	10625	176800	10093.75	167960	16796	151164	6718.4	2	1	3359.2	3359.2	0.9184
13	50	83	4150	18.444444	461.11111	597600	166	0.0913	10230	226424	9718.5	215102.8	21510.28	193592.52	12906.168	6	2	8604.112	4302.056	1.1762
14	50	96	4800	21.333333	533.33333	691200	192	0.1056	510000	13056000	484500	12403200	1240320	11162880	744192	2	1	248064	248064	67.8226
15	24	75	1800	8	200	259200	72	0.0396	52560	504576	49932	479347.2	47934.72	431412.48	28760.832	3	2	19173.888	9586.944	2.6211
16	20	63	1260	5.6	140	181440	50.4	0.02772	552500	3712800	524875	3527160	352716	3174444	211629.6	2	1	70543.2	70543.2	19.287
17	30	62	1860	8.2666667	206.66667	267840	74.4	0.04092	9410	93347.2	8939.5	88679.84	8867.984	79811.856	5320.7904	2	2	3547.1936	1773.5968	0.4849
18	34	74	2516	11.182222	279.55556	362304	100.64	0.055352	12750	171088	12112.5	162533.6	16253.36	146280.24	9752.016	6	1	3250.672	3250.672	0.8888
19	40	78	3120	13.866667	346.66667	449280	124.8	0.06864	765000	12729600	726750	12093120	1209312	10883808	483724.8	2	2	483724.8	241862.4	66.127
20	50	83	4150	18.444444	461.11111	597600	166	0.0913	12600	278880	11970	264936	26493.6	238442.4	10597.44	3	1	5298.72	5298.72	1.4487
21	50	96	4800	21.333333	533.33333	691200	192	0.1056	595000	15232000	565250	14470400	1447040	13023360	578816	2	2	578816	289408	79.1263
22	24	75	1800	8	200	259200	72	0.0396	7650	73440	7267.5	69768	6976.8	62791.2	2790.72	2	1	1395.36	1395.36	0.3815
23	20	63	1260	5.6	140	181440	50.4	0.02772	9500	63840	9025	60648	6064.8	54583.2	3638.88	6	1	1212.96	1212.96	0.3316
24	30	62	1860	8.2666667	206.66667	267840	74.4	0.04092	255000	2529600	242250	2403120	240312	2162808	144187.2	2	0	0	48062.4	13.1406
25	34	74	2516	11.182222	279.55556	362304	100.64	0.055352	510000	6843520	484500	6501344	650134.4	5851209.6	390080.64	3	1	130026.88	130026.88	35.5503

Figure 4.1: Primary Dataset

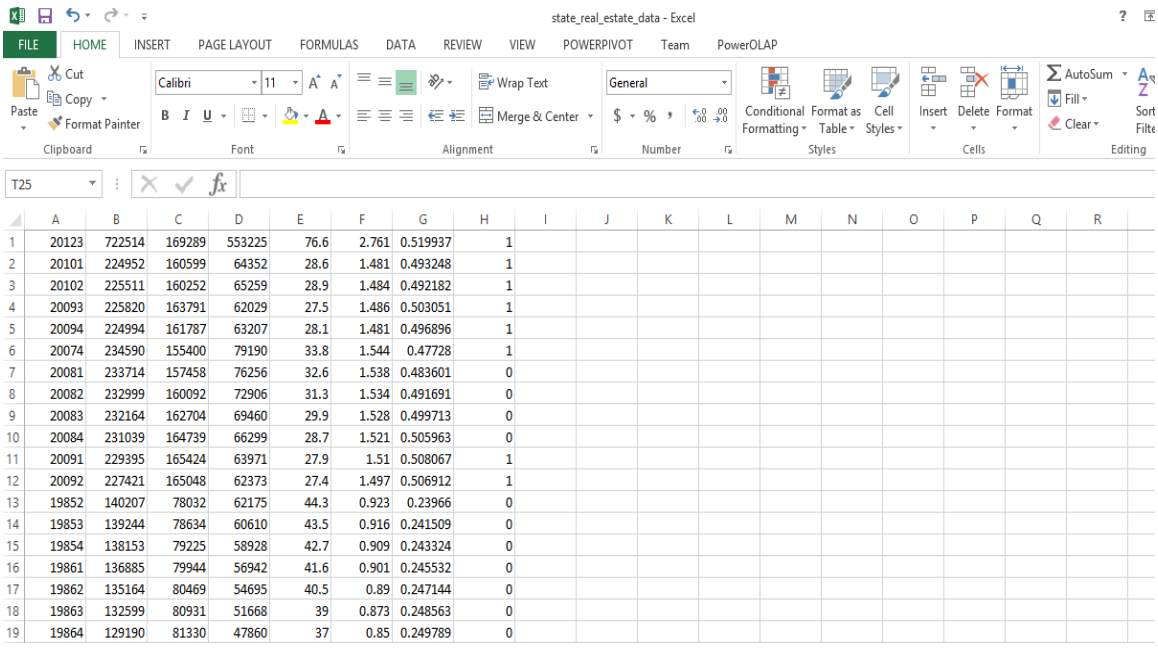
Workable Descriptive Analytics require:

i) A KPI design connecting KPIs and the measurements that influence them

ii) Adjusting KPIs and measurements to Data Delivery and Data Use Scenarios

iii) A Technical Architecture that helps the catch and transmission of KPIs and measurements

iv) Checking KPIs and measurements to guarantee they keep on representing powerful business measures.



state_real_estate_data - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW POWERPIVOT Team PowerOLAP

Cut Copy Format Painter Paste Clipboard | Calibri 11 Font | Wrap Text Merge & Center Alignment | General Number | Conditional Formatting Format as Table Cell Styles Styles | Insert Delete Format Cells | AutoSum Fill Clear Sort Filte Editing

T25

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	20123	722514	169289	553225	76.6	2.761	0.519937	1										
2	20101	224952	160599	64352	28.6	1.481	0.493248	1										
3	20102	225511	160252	65259	28.9	1.484	0.492182	1										
4	20093	225820	163791	62029	27.5	1.486	0.503051	1										
5	20094	224994	161787	63207	28.1	1.481	0.496896	1										
6	20074	234590	155400	79190	33.8	1.544	0.47728	1										
7	20081	233714	157458	76256	32.6	1.538	0.483601	0										
8	20082	232999	160092	72906	31.3	1.534	0.491691	0										
9	20083	232164	162704	69460	29.9	1.528	0.499713	0										
10	20084	231039	164739	66299	28.7	1.521	0.505963	0										
11	20091	229395	165424	63971	27.9	1.51	0.508067	1										
12	20092	227421	165048	62373	27.4	1.497	0.506912	1										
13	19852	140207	78032	62175	44.3	0.923	0.23966	0										
14	19853	139244	78634	60610	43.5	0.916	0.241509	0										
15	19854	138153	79225	58928	42.7	0.909	0.243324	0										
16	19861	136885	79944	56942	41.6	0.901	0.245532	0										
17	19862	135164	80469	54695	40.5	0.89	0.247144	0										
18	19863	132599	80931	51668	39	0.873	0.248563	0										
19	19864	129190	81330	47860	37	0.85	0.249789	0										

Figure 4.2: Secondary Dataset

This is how KPI's are utilized to assess the execution and advance of business. KPI's oversee decentralized data proficiently and to address the visual and administrative holes existing in organizations. To make progress in the globalized world organizations ought to have the capacity to get and examine continuously the present circumstance of their business. Along this KPI's is a compelling and proactive method for guaranteeing a key way to deal with quality affirmation and administration advancement. Since there is a need to see well what is critical, different systems to evaluate the current situation with the business, and its key exercises, are related to the determination of indicators. These evaluations frequently prompt the distinguishing proof of potential upgrades, so execution pointers are routinely connected with 'execution change' activities.

4.1 Applied Statistical Techniques

Feature Selection is the way of expelling highlights from the information set that are immaterial regarding the task that will be performed. Feature Selection can be of great degree help in lessening the dimensionality of the information to be handled by the classifier, lessening execution time and enhancing prescient precision. Feature Selection was performed in two different ways.

At first, to think about how the factors were connected, two statistical techniques, to be specific Pearson's correlation coefficient and ANOVA-Test were applied. It examined the dependence between two different attributes. Table 1 demonstrates the valuations of both the examination. Furthermore to decide the level of dependence each attribute is examined across one single attribute of the high value of prediction. So, for this Table demonstrates the results.

4.1.1 Correlation

Karl Pearson's coefficient of relationship is the generally utilized strategy for estimating the level of connection between two factors. This gives a knowledge of the reliance of each of the traits on the 'result' quality. Correlation has a trend to be used when there is no illustrious response variable. It gauges the superiority of direct connection between at least two factors. The Pearson correlation coefficient measures the quality of the straight relationship between dual factors such as length and breadth, length and area, earnest money and registry amount etc.

Correlation is used to describe the linear relationship between two continuous variables. This relationship is measured on the basis of some well-defined value called range. This range is a dimensionless quantity and is independent of units of measurement that is P and Q. There are some set conditions of range and are explained as follows:

i. If the correlation is greater than zero then as the value of P increases the value of Q also increases. This defines that P and Q are positively correlated or r = 1 is "Perfect Positive Correlation".

ii. If the correlation is lesser than zero then as the worth of P increases the worth of Q decreases. This defines that P and Q are negatively correlated or r = -1 is "Perfect Negative Correlation".

iii. If the correlation is equal to 0 at that time, there is no direct association between P and Q. This defines that both the variables P and Q are uncorrelated that r = -1 is "Zero Correlation". Also, correlation is zero only when the co-variances between P and Q is 0.

In MATLAB, there is an inbuilt toolbox called Parallel Computing Toolbox. As soon as there is full benefit of this toolbox, manual calculations using the formula are not required using equation 4.1

$$r = \frac{\mathrm{n}\sum d1,d2-(\mathrm{d1})(\sum d2)}{\sqrt{[n\sum d1^2-(\sum d1\,)^2][n\sum d2^2-(\sum d2)^2]}} \tag{4.1}$$

In the command window of MATLAB, only a set of data for the variables d1 and d2 are provided and at last formula to find the correlation is written as in equation 4.2.

$$corr_{coeff} = corr2(d1, d2) \tag{4.2}$$

It automatically takes the inbuilt formula of MATLAB.

4.1.2 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a system for deciding if variety in the reaction variable emerges inside or among various populace gatherings. Insights and Machine Learning (ML) Toolbox of MATLAB give one-way, two-way, and N-way Analysis of Variance (ANOVA). ANOVA test is used to determine whether there are any statistical differences between the means and/or between the variances of two or more independent groups. Here, for house price index versus attributes ANNOVA value is calculated. If ANOVA value is greater than 0.05 then the difference is significant.

Here, E comes as E+ve or E-ve. E+ve value depicts too large value of ANOVAs P value and E-ve value depicts too small value of ANOVAs P value.

One-route Analysis Of Variance(ANOVA) will be performed on these variables, which will decide if there is any precise significant gap between the means of two or more unrelated (independent) attributes or not. ANOVA table is shown as under in table 5.6:

When ANOVA is evaluated

$Value > 0.05 \rightarrow then\ the\ difference\ is\ significant$

$Value < 0.05 \rightarrow then\ the\ difference\ is\ not\ significant$

Here,

$SS \rightarrow Sum\ of\ Squares$

$df \rightarrow Degree\ of\ Freedom$

$MS \rightarrow Mean\ Sum\ of\ squares$

$F \rightarrow ANOVA's\ F\ value$

$Prob > F \rightarrow ANOVA's\ P\ value$

4.2 Methodology of Descriptive Analysis

In this project domain, the focus is on analyzing and querying large dataset of real estate problems using cube technology and Microsoft SQL Server Business Intelligence Development Studio, Developer Edition tool, and later on, visualizing and creating dashboards using Power BI tool.

1. Extract, Transform and Load (ETL)
2. Create data sources in SQL Server Management Studio
   - i. Form Design ➔ STAR SCHEMA design
   - ii. Creating Dimensions and cubes
   - iii. Creating a dimensional data store
     - a. Creating dimensions
     - b. Creating attributes
   - iv. Mapping table with levels of dimensions
     - a. Creating cube
   - v. Cube Mapping with the relational data source
     - Import data to SQL Server Analysis Services
     - Load data in SSAS
     - Designing cube in BIDS 2012
     - Applying OLAP operations in the browser window of BIDS

4.3 Steps followed to apply relevant technology: Cube Technology

4.3.1 Extract, Transform, and Load (ETL)

Extract, Transform and Load (ETL) process is applied on heterogeneous datasets [59,60] to convert it into a homogenous form.

- **Extract:** Extract real estate dataset in the form of .csv files, .xls files and .xlsx files from [59] and [60].
- **Transform:** Transformations are applied for converting heterogeneous datasets to homogenous format files (.xlsx in this work) where transformations could be data type conversions, formatting, merging and splitting applied on columns, etc.
- **Load:** Loading the homogeneous format (.xlsx) files into SQL Server Analysis Services (SSAS) tool to map these datasets with multidimensional analysis model.

4.3.2 Creating data sources in SQL Server Management Studio

Creating data sources involves data entries to some particular columns selected for various tables. For creating a particular database related to real estate, the first step is to select various dimensions/tables to be created as shown in Table 4.1- 4.5

In this research work, the dimension and related levels nominated in the dataset include, (A) Customer Details, (B) Property Details, (C) Property Description, (D) Property Major Description and (E) Document File, as shown in tables. The tables are created in SQL Server Management Studio 2012 using create table query. The data entries to these tables are made using insert query. And lastly, the table is viewed using a select query.

Table 4.1: The basic illustration of dimensions and related levels of customer detail

DIMENSION	LEVELS
Customer Details	Customer Id
	Customer Name
	Customer Business
	Customer City
	Customer Number
	Customer Type

Table 4.2: The basic illustration of dimensions and related levels of property detail

DIMENSION	LEVELS
Property Details	Property Id
	Property Type
	Property Description
	House Number
	Property Address
	Property City
	Property Facing
	Property Neighbours
	Property Major Description
	Number of Owners
	Gender

Table 4.3: The basic illustration of dimensions and related levels of Property Description

DIMENSION	LEVELS
Property Description	Description Id
	Length
	Breadth
	The area in Square Foot
	Area in Marla
	Area in Guz
	The area in Square Inches
	Area in Sarsahi
	Area in Acre

Table 4.4: The basic illustration of dimensions and related levels of Property Major Details

DIMENSION	LEVELS
Property Major Details	Major Description ID
	Number of rooms
	Number of washrooms
	Number of kitchens
	Number of storerooms
	Dining room
	Drawing room
	Garage
	Garage capacity
	Garden
	Basement
	Basement Type
	Usage
	Electricity Supply
	Water Supply
	Number of Floors

Table 4.5: The basic illustration of dimensions and related levels of Document File

DIMENSION	LEVELS
Document File	File Id
	Site Map
	Old Registry
	Mutation
	Agreement
	Collector Rate per Marla
	Price per collector rate per Marla

	Price closed per Marla
	Price of property per closed value
	Earnest money
	Stamp duty on the agreement
	Registry amount
	Stamp duty on the registry
	Number of owners for registry
	Time is taken for registry
	Commission per sale
	Registration fee

4.3.3 Creating Cube

Cube creation, mapping and the validation of the ETL process are done in Business Intelligence Development Studio 2012. The main aim in using this software is the formation of multidimensional data mart which provisions the task of real estate data analysis is shown in Figure 4.3. So, the whole process is carried out in the following steps:

- **Creating Dimensions:** Dimensions selected for a particular business domain are unique and are level-based dimensions. The dimensions selected for our domain are (A) Customer Details, (B) Property Details, (C) Property Description, (D) Property Major Description and (E) Document File.
- **Creating Levels:** In this domain, Customer Profile dimension consists of nine levels named as profile id, name, business name, category, contact number, residential address, business address, city and pin code. Property dimension consists of twenty levels namely: property id, profile id, category, city, state, contact number, property description, property major description, facing, segment, area, units, length, breadth, number of floors, left neighbor, right neighbor, collector rate, property rate, number of owners. Document dimension consists of thirteen levels named as owner id, profile id, earnest money,

agreement, stamp duty on the agreement, mutation, old registry, site plan, bank loan, balance, registration time, stamp duty on registry and commission.

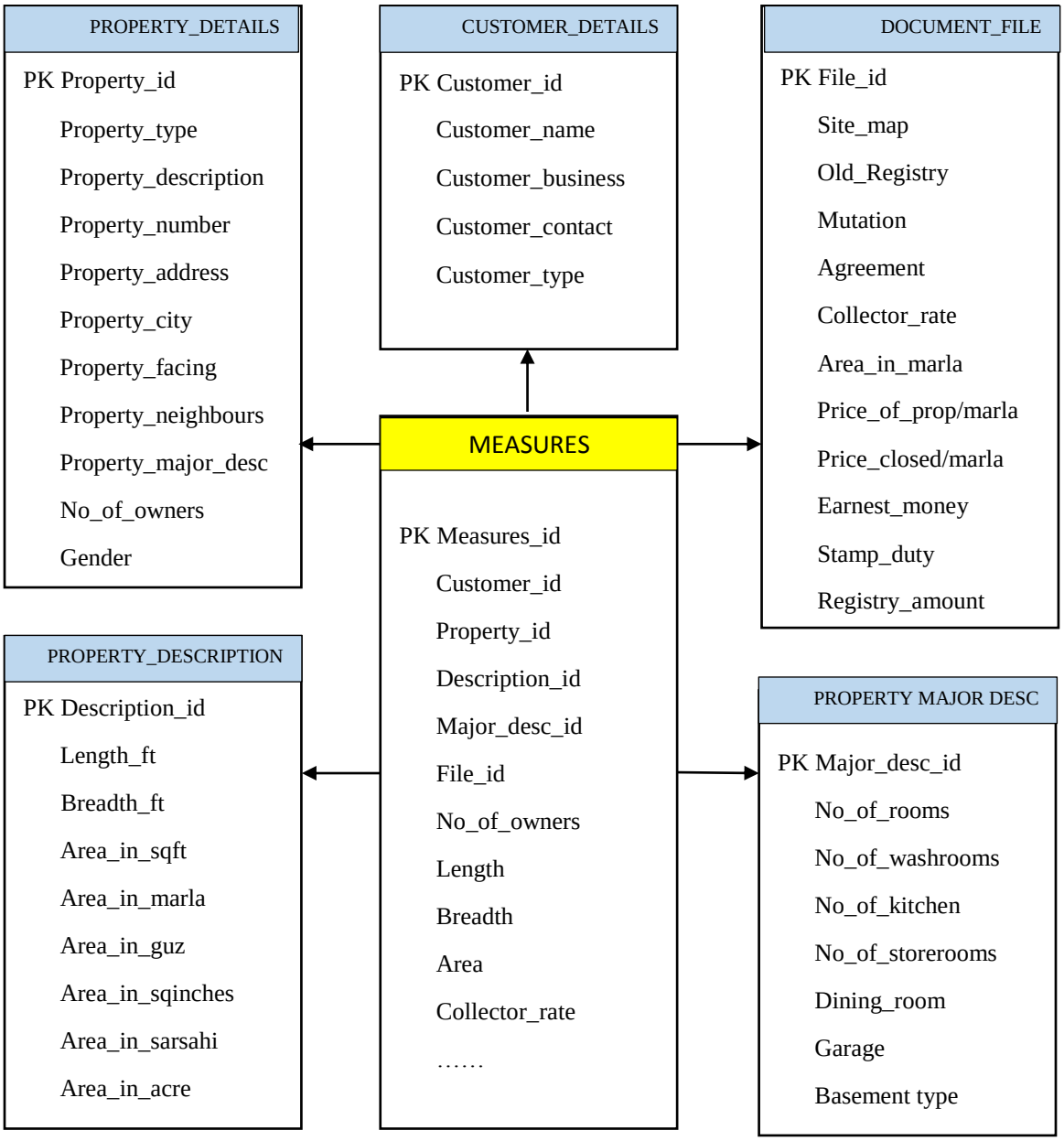
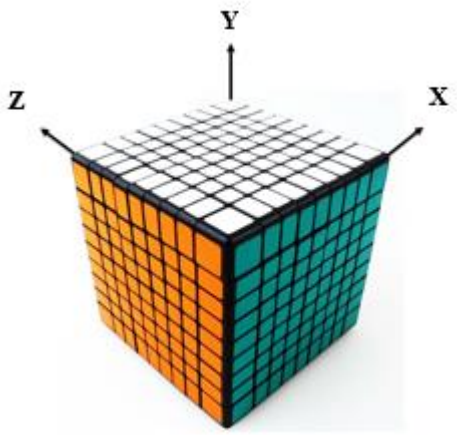


Figure 4.3: STAR Schema Design

- **Creating hierarchies:** Dimensions selected for this particular domain have one hierarchy and is a level based hierarchy. Likewise, the hierarchy for customer profile follows the sequence: profile id, name, business name, category, contact number, residential address, business address, city and pin code, the hierarchy for property dimension be property id, profile id, category, city, state, contact

number, property description, property major description, facing, segment, area, units, length, breadth, number of floors, left neighbour, right neighbour, collector rate, property rate, number of owners and the hierarchy for document dimension be owner id, profile id, earnest money, agreement, stamp duty on agreement, mutation, old registry, site plan, bank loan, balance, registration time, stamp duty on registry and commission.

- **Mappings:** In mapping part, each column of the relational database table is mapped to the member part of each level in the analysis model.
- **Loading data:** A data connection is built between SSMS and SSAS tools. The dataset is loaded in SSAS and further accessed in BIDS 2012 for the cube creation.
- **Cube Creation:** Cubes is created by firstly creating the measures and then mapping the measures to the relational database table to create a final design. In this research work, STAR schema design has been implemented for the selected dataset as shown in Figure 4.3.



X: CITY

- Jalandhar
- Ludhiana
- Patiala
- Amritsar

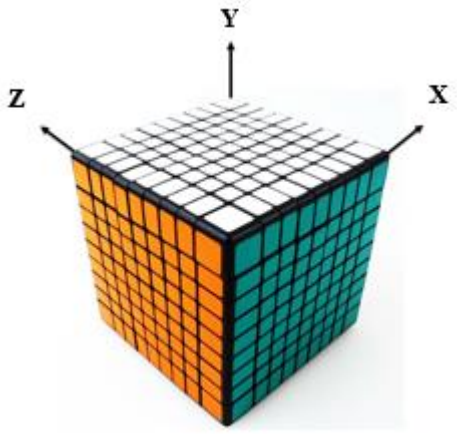
Y: GARAGE

- Lentered
- Non-Lentered
- None

Z: NEIGHBOURS

- Right-Left-back
- Right-Left
- Left-Back

- Right-Back



X: ADDRESS

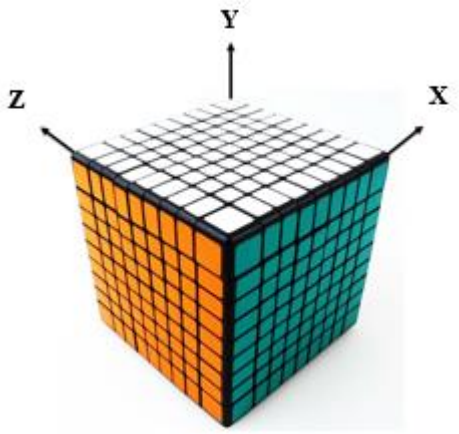
- Name of colony

Y: CUSTOMER TYPE

- Seller
- Buyer
- Rental

Z: GARDEN

- L
- NL
- N



X: OLD REGISTRY

- Yes
- No

Y: CUSTOMER TYPE

- Seller
- Buyer
- Rental

Z: PROPERTY TYPE

- Commercial
- Residential

Figure 4.4: Various Cube Structures

CHAPTER 5

IMPLEMENTING PREDICTIVE ANALYSIS

This section portrays the idea of J48 and Support Vector Machine (SVM) model in predictions. The connection between the existing and proposed methodology is built. Results are reviewed in form of confusion matrix to take a clear view about the property. Also, the hybridization of J48 and SVM model is done on the Secondary dataset.

5.1 J48 Decision Tree

In a decision tree, a data point is classified with subsequent paths from the root node through the tree, capturing the edges equivalent to the values of attributes; until the split is completed or no over fit occur. Decision trees utilize a pre-arranged dataset to order information utilizing existing patterns and examples. After the tree is delivered, the learning from the decision tree can be incorporated into various distinctive interruption location innovations including firewalls and IDS marks. It has the perspective to support an intrusion detection team with many disputes of protecting a network. Due to the expanding volume of data, decision trees have the potential to save time for security experts and helps in the analysis of malicious data. With this analysis, decision tree algorithms find out the abnormalities of the network and afford customized response to hold intrusion detection.

Steps followed are:

i. Dataset is loaded as Comma Separated Value (CSV) file. Also, this file must contain numeric values

ii. This file is read by function ‘csvread’ and stores its values in function a.

iii. Then this data is divided into input features and target class. Here division of data is done as (:, 1:end-1) and (:, end), where (:, end) means last or we say the last row.

iv. Then data is divided into training sets and testing sets. For this one function named Holdout is used. This function is standard function and can be taken from the toolbox of MATLAB. In this dataset is passed.

v. The dataset that is divided into training and testing dataset, it is stored in form of atr and ate. Where atr is data for training and eating is data for testing.

vi. Then data in atr is divided into two parts to separate test data and test class. Similarly, data in ate is divided into two parts to separate train data and train class. This training data and testing data is separated from target class. And target class is kept in test class and train class.

vii. Here the Naïve Bayes model is applied to take train data as input and train class is expected. Because training output is known.

viii. Now, the confusion matrix is generated. Confusion Matrix is a table that is often used to describe the performance of a classification model or classifier on a set of test data for which the true values are known

ix. Here J48 will predict upon test data. So we get the predicted results using J48

x. Now accuracy is evaluated to validate if the results of test class are equivalent to j48 predicted results or not. Because in test class actual values are obtained but in predicted those values are obtained that our model had predicted.

xi. If in both the cases are same, then accuracy increases otherwise decrease.

xii. Initially, TP = TN = FP = FN = 0

a. True Positive (TP): These are cases in which test class and J48 predicted the result, both prove that property is bought.

b. True Negative (TN): These are cases in which test class says the property is bought but J48 predicted result says that property is not bought.

c. False Positive (FP): These are cases in which test class says the property is not bought but J48 predicted result says that property is bought.

d. False Negative (FN): These are cases in which test class says the property is bought but J48 predicted result says that property is not bought.

xiii. If the predicted result is equal to J48 predicted results, then all four constraints that are TP, TN, FP, and FN are increased by 1. And lastly, Sensitivity (TPR), Specificity (TNR), Precision and F_Measure are evaluated. This is how the confusion matrix is obtained

5.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning model that is characterized as the limited dimensional vector spaces where each measurement portrays an element of a specific protest. Along these lines, SVM has been demonstrated as a successful technique in high-dimensional space issues. Because of its computational capability on enormous datasets SVM is ordinarily utilized in report arrangement, assessment investigation and expectation based errands.

Support Vector Machine (SVM) is noticed as the first choice for classification problems. Support Vector Machines (SVMs) are nothing but machines constructed for classifying positive and negative classes. These are mainly dependent on support vectors, which are decisive points for classifying data. The attractiveness of SVMs lies in its nice mathematical equations, pretty pictorial representations, excellent generalization abilities. They give optimal and global solutions, with low overfitting and overcomes the curse of dimensionality problem. SVMs are designed based on optimization methods and are the most significant tools for solving the problems of Machine Learning (ML) with finite training data. These are dependent on the Structural Risk Minimization (SRM) principle. They exploit large mathematical foundations to avoid overfitting and better empirical results. Mostly SVMs adjust Machine Learning (ML) problems to optimization problems; specifically, the convex optimization problems are used in the early era of SVMs in the 1990s. Without Statistical Learning Theory (SLT) the definition of SVMs is incomplete.

There are three aspects which make SVM more successful namely maximal margin hyperplane construction using support vectors, dual theory, and kernel trick. SVMs are straightaway used in various applications like handwriting recognition, intrusion detection, Speech recognition, Bioinformatics, information extraction, face detection and many more. And also they are treated as a remarkable technique because of its scope to handle data with larger dimensions. Unlike neural networks, it will not sustain the local minima problem, and consider very less modeling parameters, producing stable results. But, they have slower training times particularly with non-linear data and huge input data. SVMs are generally used as binary classifiers. Steps followed while implementing as shown in Figure 5.1 are:

i. Firstly data is loaded as a Comma Separated Value (CSV) file.

ii. This file is read by function 'csvread' and stores its values in function a.

iii. Then this data is divided into input features and target class. Here division of data is done as (:, 1:end-1) and (:, end), where (:, end) means last or we say the last row.

iv. Then data is divided into training sets and testing sets. For this one function named Holdout is used. This function is standard function and can be taken from the toolbox of MATLAB. In this dataset is passed.

v. The dataset that is divided into training and testing dataset, it is stored in form of atr and ate. Where atr is data for training and eating is data for testing.

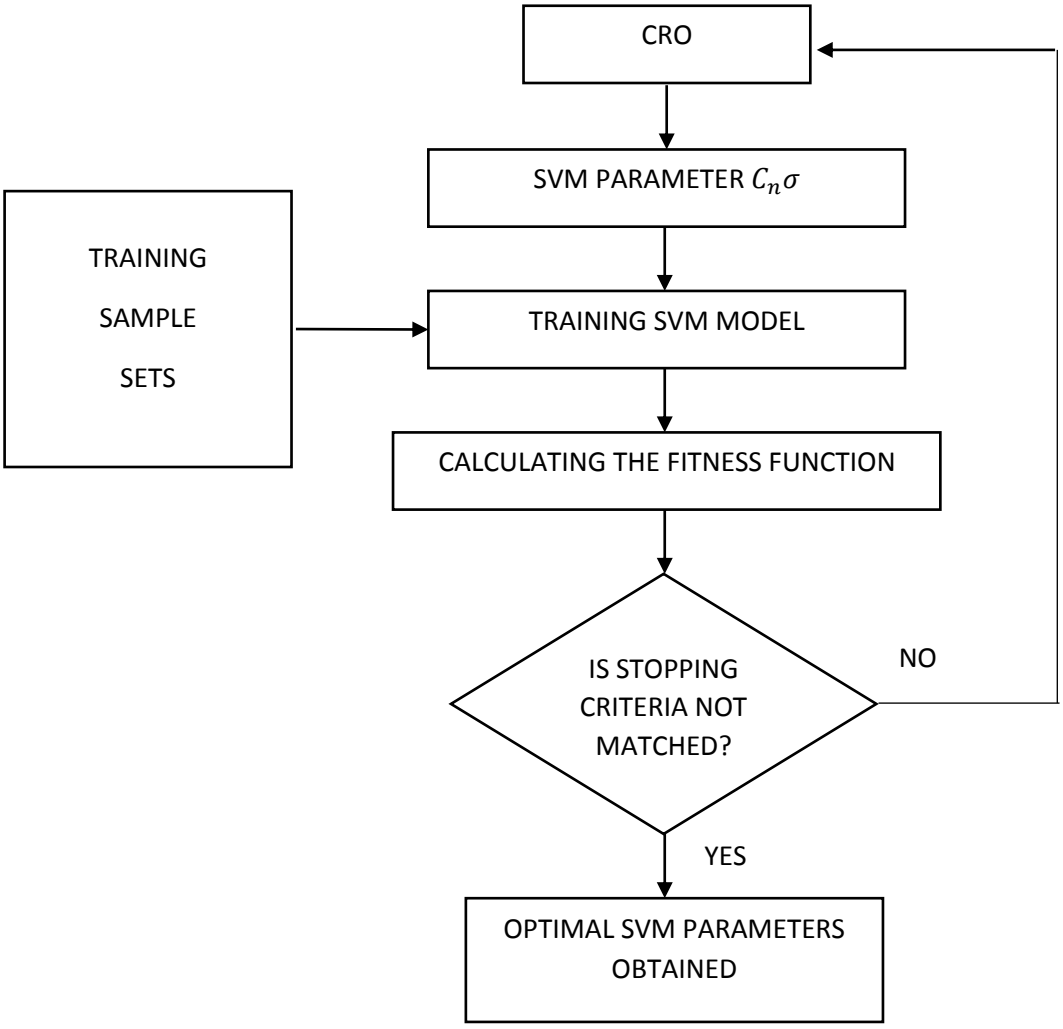


Figure 5.1: Flowchart of Support Vector Machine (SVM)

vi. Then data in atr is divided into two parts to separate test data and test class. Similarly, data in ate is divided into two parts to separate train data and train class. This training data and testing data is separated from target class. And target class is kept in test class and train class.

vii. Here classification is done using Support Vector Machine model.

viii. Training is done using function svm.train and prediction is done using model svm.predict.

ix. Again confusion matrix is created to validate the actual and predicted results.

Also, once the system is ready to respond to queries in fractions of seconds, it can also predict using one or the other model. Two models that are J48 & SVM are implemented. One dataset is used in both the models. A single variable is known as 'holdout' to fraction the data into training and testing data. In both training and testing data, we separate the training class and test class. While training our model, train data and train class both are passed but while testing our model, only test class is passed and predicted values are taken. Also from training and testing, we separate the input variable and test class. These predicted values are compared with the actual value of test data and we create confusion matrix and we get all five parameters that are accuracy, true positive rate, true negative rate, F-measure, and precision.

After completing the predictive analysis we conclude that there is one issue of "Parameter tuning" that is what type of input parameters should be there so that our technique is good. Firstly, what researchers do is, they consider any of the parameters. Later, if there is any little change in a fraction, accuracy is changed widely. These parameters were selected on the basis of hit and trial method. But, resulting formal parameter tuning is time-consuming and may not provide accurate results, whenever training and testing scaling is there.

CHAPTER 6

IMPLEMENTING PRESCRIPTIVE ANALYSIS

This section portrays the idea of NSGA-III model. This is a proposed technique over J48 and Support Vector Machine model to optimize the results and to acquire the best accurate results. The generally new field of prescriptive examination enables clients to endorse various distinctive conceivable activities to and control them towards an answer. More or less, these investigations are tied in with giving guidance. Prescriptive investigation endeavor to measure the impact of future choices keeping in mind the end goal to prompt on conceivable results previously the choices is really made. Taking care of business, prescriptive examination predicts what will happen, as well as why it will happen giving proposals with respect to moves that will make the favorable position of the expectations. These examinations go past distinct and prescient investigation by suggesting at least one conceivable approaches. Basically, they anticipate various prospects and enable organizations to evaluate various conceivable results in light of their activities. Prescriptive investigation utilizes a blend of strategies and devices, for example, business rules, calculations, Machine Learning (ML) and computational displaying techniques. These methods are connected against contribution from a wide range of informational indexes including chronicled and value-based information, ongoing information sustains, and huge information. The prescriptive investigation is generally intricate to manage, and most organizations are not yet utilizing them in their day by day course of business. At the point when actualized effectively, they can largely affect how organizations decide, and on the organization's main concern. Bigger organizations are effectively utilizing prescriptive investigation to enhance creation; booking and stock in the inventory network to ensure that are conveying the correct items at the ideal time and improving the client encounter.

6.1 Non-dominating Sorting Genetic Algorithm (NSGA-III)

NSGA-III is a well-known metaheuristic technique which can find the optimal path between a given set of nodes with sink as a destination. So, NSGA-III based tree beggar is proposed. Tree beggar here is the creation of multiple trees same as random forest.

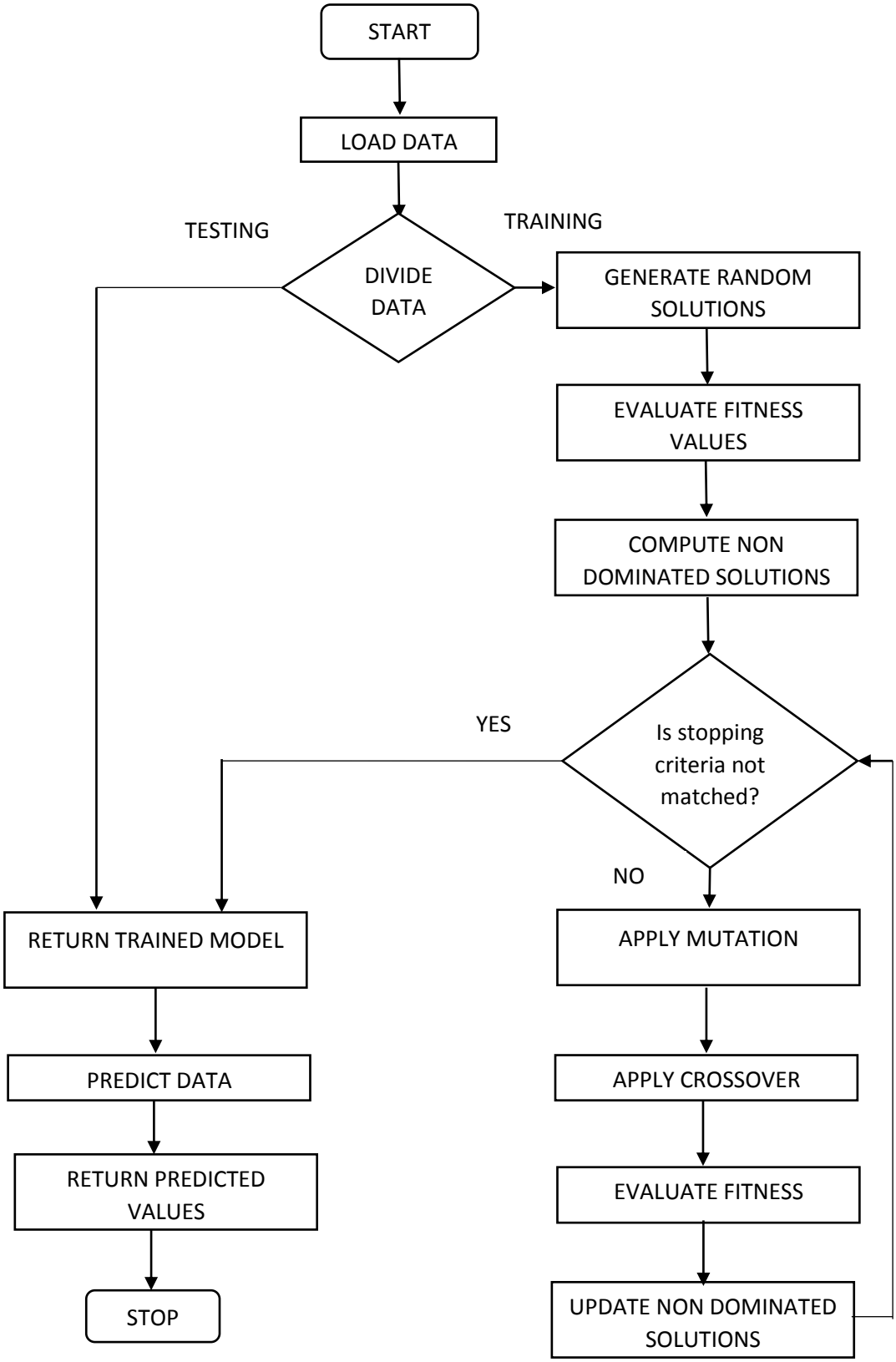


Figure 6.1: Flowchart of NSGA-III

Also training the model needs all the input parameters that are taken as input. But now tuning would not be done manually, rather is done using optimization techniques. Genetic Algorithm (GA) was found during research because it's easy to implement. But the Genetic Algorithm (GA) has issues:

i. Poor convergence speed
ii. May stuck in local optima
iii. May suffer from premature convergence

Poor convergence speed comes if crossovers and mutation are checked over and over again. Due to this, it gets stuck in one single solution which results in an issue that is it may be stuck in local optima. Suffering from premature convergence implies that if the time gets completed still we do not get the accurate results. Figure 6.1 shows Non-Dominating Sorting Genetic Algorithm (NSGA-III) and further steps explain the flowchart:

i. Firstly data is loaded as a Comma Separated Value (CSV) file.
ii. This file is read by function 'csvread' and stores its values in function a.
iii. Then this data is divided into input features and target class. Here division of data is done as (:, 1:end-1) and (:, end), where (:, end) means last or we say the last row.
iv. Then data is divided into training sets and testing sets. For this one function named Holdout is used. This function is standard function and can be taken from the toolbox of MATLAB. In this dataset is passed.
v. The dataset that is divided into training and testing dataset, it is stored in form of atr and ate. Where atr is data for training and eating is data for testing.
vi. Then data in atr is divided into two parts to separate test data and test class. Similarly, data in ate is divided into two parts to separate train data and train class. This training data and testing data is separated from target class. And target class is kept in test class and train class.
vii. Once divided into training and testing dataset, training dataset is allowed to go through step of conversions but testing dataset is directly returned to training model to predict data and return predicted values
viii. In training dataset random solutions are generated.

ix. Furthermore, fitness values are evaluated.

x. Multiple solutions are evaluated out of which non-dominated solutions are computed which means those solutions which do not clash with one another.

xi. If stopping criteria is matched, then model turns to testing part, else applies mutations and crossovers to update non-dominated solutions.

xii. If the mutated version is better than the parent solution it is selected and the competing, less solutions are allowed to die.

xiii. Once, updated non-dominated solutions returns to training model, then data is predicted and returns predicted values.

To get rid of these issues, researchers of GA found variants of GA that is NSGA-I, NSGA-II, and NSGA-III. Each variant had one or another feature but we carried NSGA-III because it overcomes all the issues of GA as shown in the figure. Also, it can optimize up to thirteen parameters at once likewise accuracy, true positive rate, true negative rate, F-measure, and precision are just five parameters. Although, optimizing thirteen parameters at once may reduce a little speed. NSGA-III creates random solutions and on the basis of random solution checks the tree beggar. Here predicted results are again checked. The best known accuracy that came so far is saved. Then again re-combinations are done and the child is created that is there accuracy is checked. If a child has better results than a parent, then the model upgrades the parent's accuracy. This how optimization is done.

CHAPTER 7

RESULT ANALYSIS

7.1 Descriptive Analysis

This section portrays the idea of Data Cubes and Data Warehousing in the setting of Online Analytical Processing (OLAP). The standards of data cubes are clarified and in addition how they are composed and what they are utilized for. The connection between data cubes, data distribution centers, and social databases is additionally inspected. Also, statistical techniques are applied for selecting features out of bulk.

7.1.1 Visual Analysis

The purpose of this section is to recognize and understand the visual outputs by observing the code implementation. Figure 7.1 shows the evaluated results for area and price of property per closed value and Figure 7.2 shows the ANOVA table.

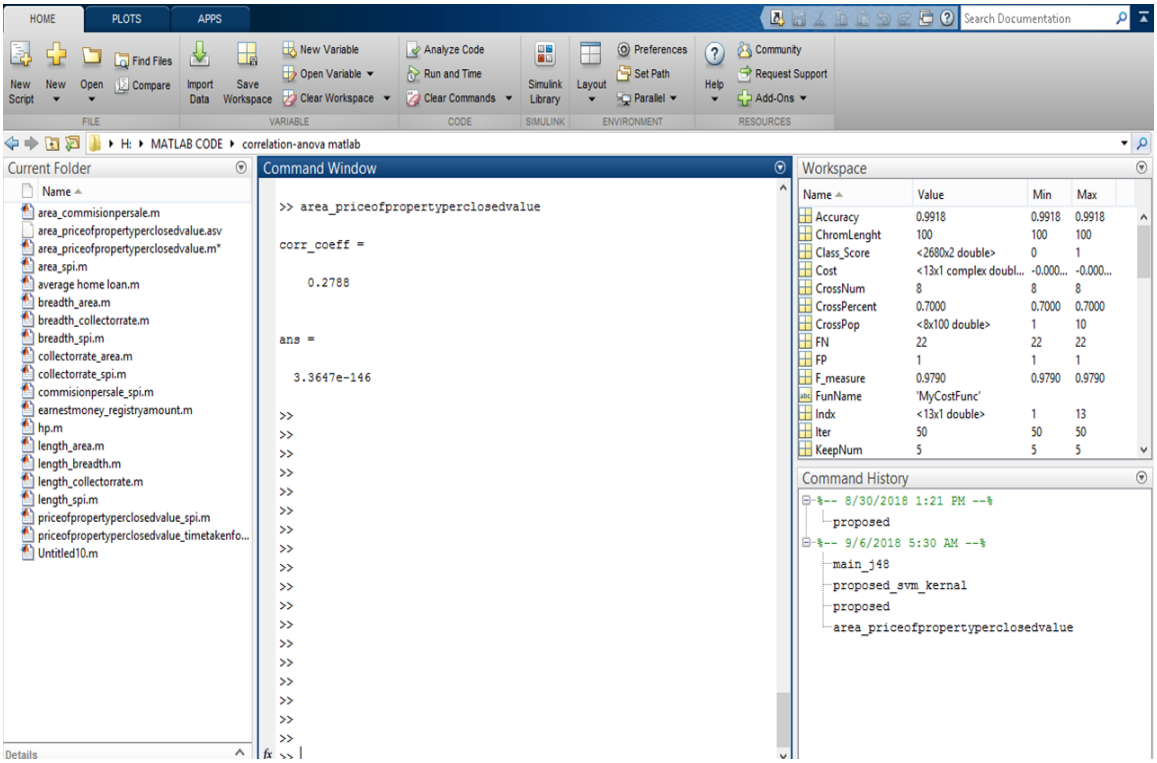


Figure 7.1: Correlation and ANOVA evaluated for Area and Price of property/closed value

ANOVA TABLE					
Source	SS	df	MS	F	Prob>F
Groups	28211.3	547	51.5745	14.5	3.36467e-146
Error	1607.9	452	3.5574		
Total	29819.2	999			

Figure 7.2: ANOVA evaluated for Area and Price of property per closed value

Figure 7.3 shows the evaluated cube results based on various levels such as customer_id, price of property per closed value, collector rate per marla, price closed per marla, area in square foot, length in square foot and breadth in square foot and Figure 7.4 shows the evaluated cube results based on various levels such as customer id, customer city, property facing, garage capacity, length in square foot, breadth in square foot and area in square foot.

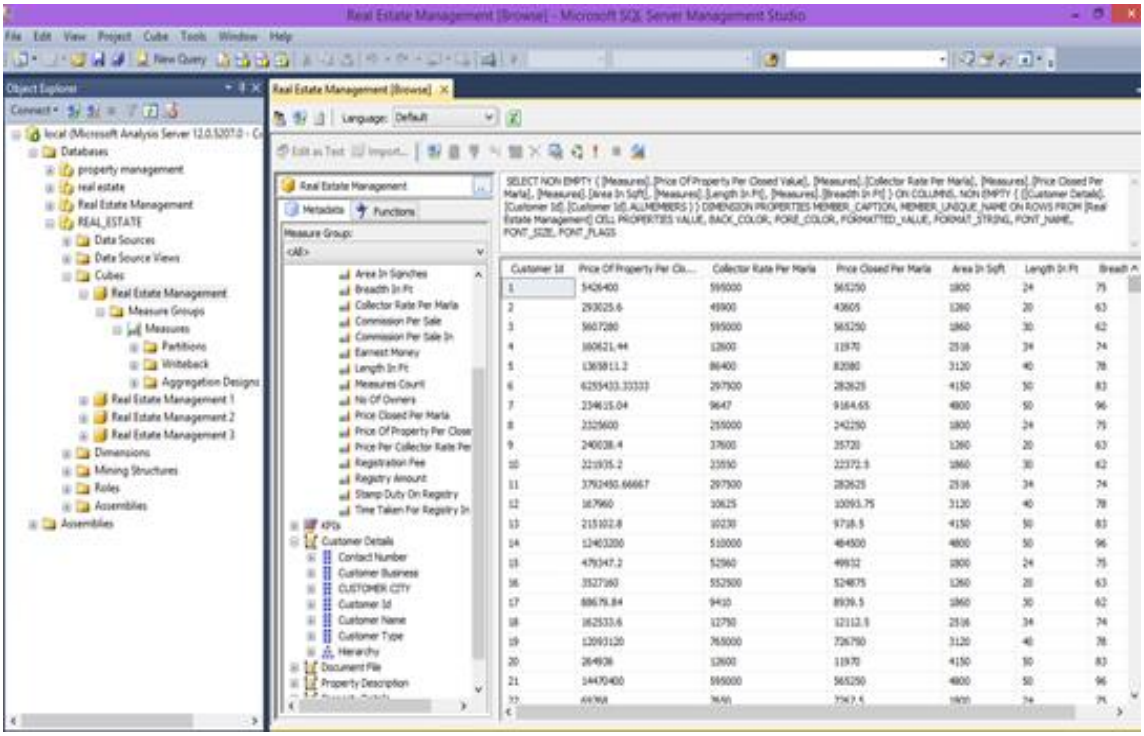


Figure 7.3: Cube Visualization in SQL Server Management Studio based on levels such as customer id, length, breadth, area, collector rate/marla and price of property/marla

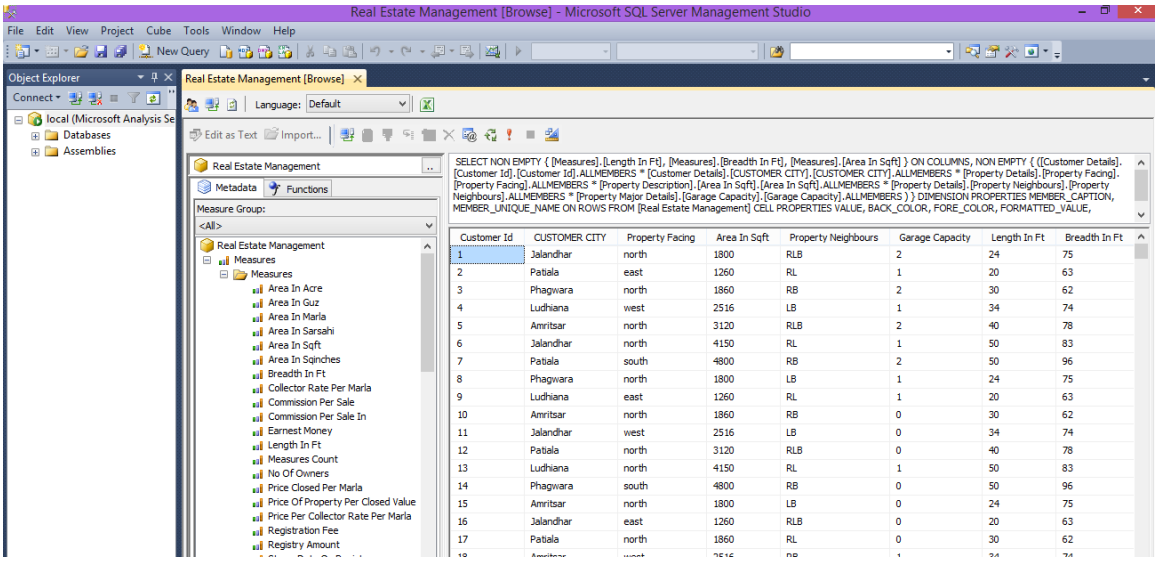


Figure 7.4: Cube Visualization in SQL Server Management Studio based on levels such as customer id, customer city, property facing, garage capacity, length, breadth and area

Visualizations (also called portrayals) demonstrate encounters that have been found in the data. A Power BI report may have a singular page with one visual or it might have pages stacked with visuals. In Power BI, visuals can adhere from reports to dashboards.

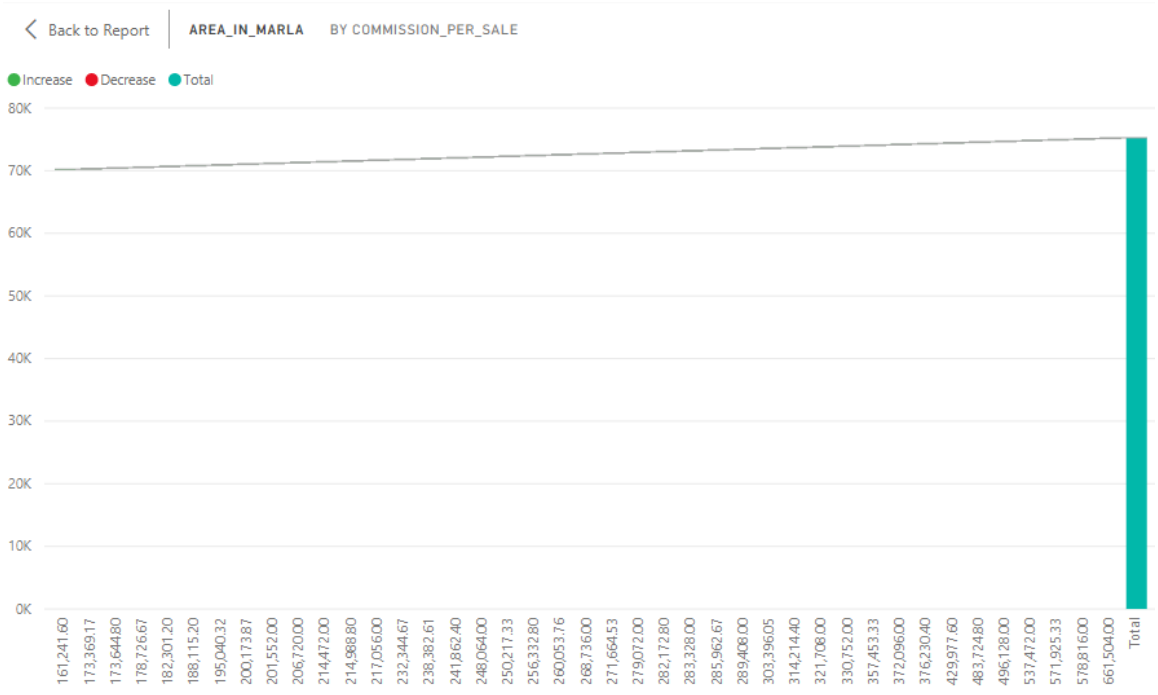


Figure 7.5: Visualization showing area in marla by commission per sale

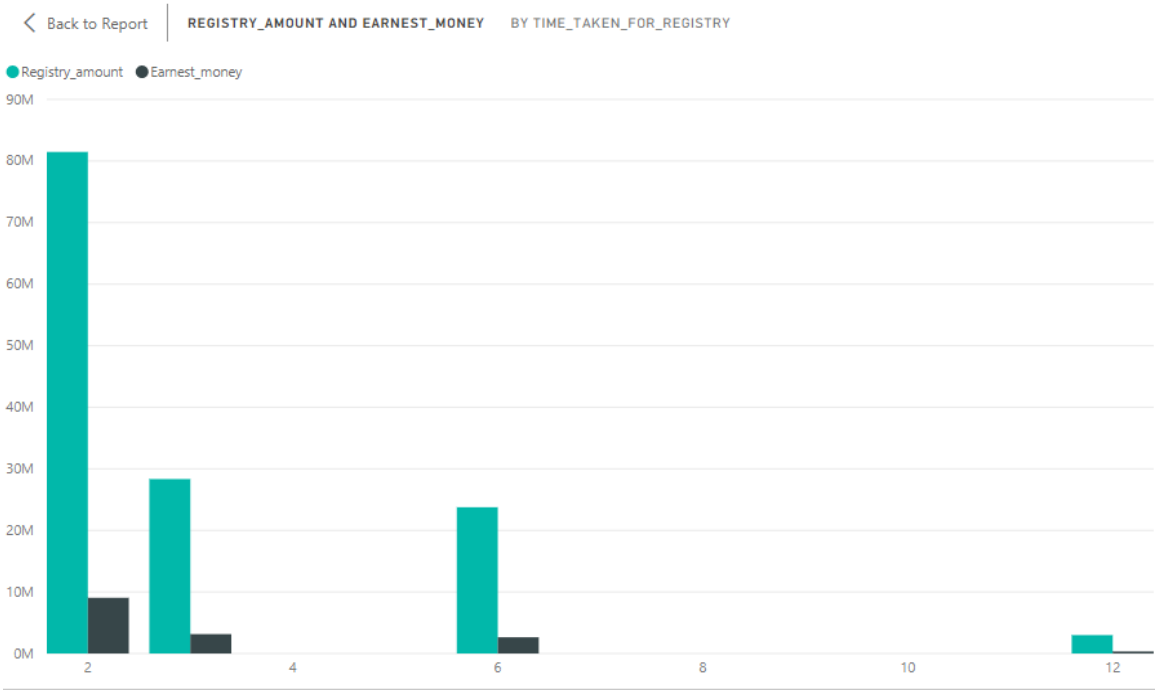


Figure 7.6: Visualization showing registry amount and earnest money by time is taken for registration

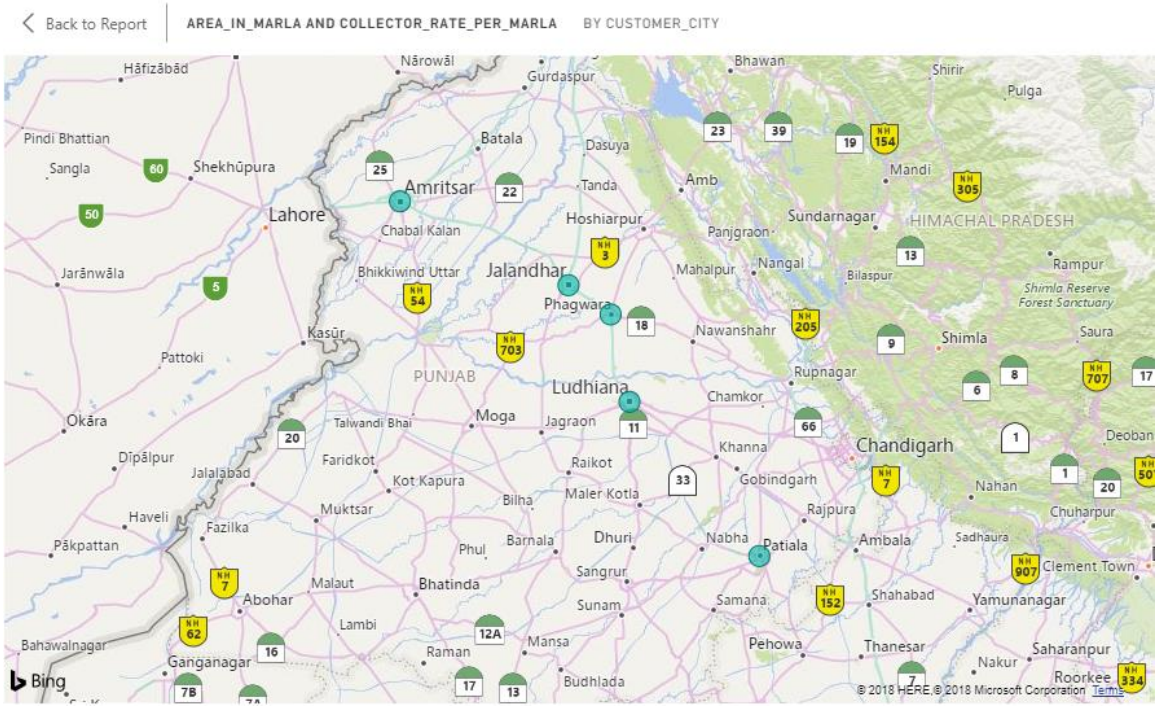


Figure 7.7: Visualization showing area in marla and collector rate per marla by customer city

It's imperative to make the refinement between report producers and report buyers. If an individual is building or changing the report, by the individual is a creator. In Power BI Desktop, this infers that opening the dataset in data view and make visuals in Report view. In Power, BI dataset can be viewed or reported in the report publication director in the editing view. There are an extensive variety of visual sorts open clearly from the Power BI visualizations sheet.

Various graphs are created on the basis of real estate data as shown in Figure 4.6, Figure 4.7 and Figure 4.8 for a graphical representation of data. These dashboards created helps to easily analyze the trend and saves our time to view the data manually.

7.1.2 Quantitative Analysis

Correlation and ANOVA values are evaluated for primary dataset and to relate with secondary dataset [60], mean of house prices is taken and same way mean of house price index is taken. Mean of house prices is divided with a mean of house price index to get resultant value. This resultant value of the secondary dataset is multiplied by house prices of the primary dataset to get the house price indexes of the primary dataset. Again the ANOVA test and Correlation is applied to the results to evaluate accurate results. Quantitative outcomes shown in Table 7.1 and 7.2

Table 7.1: Evaluating Correlation and ANOVA using Primary Dataset

S No	Attributes		Correlation Coefficient	ANNOVA's F Value	ANNOVA's P Value	Remarks
	d1	d2				
1	Length	Breadth	0.7859	1037.55	0	Average correlation Differ significantly
2	Breadth	Area	0.8908	-3.88411e+16	1	Good correlation Does not Differ significantly
3	Length	Area	0.9751	6.41519e+16	0	Good correlation Differ significantly
4	Collect or rate	Area	-0.0099	0.27	0.9833	Poor correlation Differ significantly

5	Earnest money	Registry amount	1.0000	-3.84241e+15	1	Good correlation Does not Differ significantly
6	Area	Price of property per closed value	0.2788	14.5	3.36467e-146	Poor correlation Differ significantly
7	Price of property per closed value	Time is taken for registry	0.0057	0.58	0.6277	Poor correlation Differ significantly
8	Area	Commission per sale	0.2917	7.06	1.1744e-81	Poor correlation Differ significantly
9	Length	Collector rate	-0.0036	0.84	0.9355	Poor correlation Differ significantly
10	Breadth	Collector rate	-0.0225	0.9	0.8302	Poor correlation Differ significantly

Table 7.2: Evaluating Correlation and ANOVA using Secondary Dataset

S No	**Attributes**		**Correlation Coefficient**	**ANNOVA's F Value**	**ANNOVA's P Value**	**Remarks**
	d1	**d2**				
1	Length	House price index	0.2793	14.42	6.11467e-146	Poor Correlation Differ Significantly
2	Breadth	House price index	0.2366	16.63	2.02844e-158	Poor Correlation Differ Significantly
3	Area	House price index	0.2788	14.51	1.71007e-146	Poor Correlation Differ Significantly
4	Collector Rate	House price index	0.8723	49.78	2.43562e-259	Good Correlation Differ Significantly
5	Commission per sale	House price index	0.9020	5.81	2.42385e-73	Good Correlation Differ Significantly

7.2 Predictive Analysis

This section covers the cross-authentication between existing and proposed techniques. Some familiar algorithms constraints have been chosen to show that the performance of the proposed algorithm is superior to the existing techniques.

For experimentation and execution, the proposed method is assessed utilizing MATLAB tool 2013 and measurements and Machine Learning (ML) tool stash. Here we will contrast the parameters of existing and proposed calculation i.e. converging of SVM and J48 Decision tree.

The unthinkable and graphical correlation has been done amongst existing and proposed a procedure based on parameters like TP-Rate, TN-Rate, Accuracy, F measure, and Precision. True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) are required to calculate all five parameters.

So, to calculate TP, TN, FP and FN, confusion matrix was created. Confusion matrix is a table that is often used for describing the performance of classification models on a set of test data for which the true values were known.

Table 7.3: Confusion Matrix

N=Testing Data	**Predicted (NO)**	**Predicted (YES)**	**Totals**
Actual – NO	TN	FP	TN+FP
Actual – YES	FN	TP	FN+TP
Totals	TN+FN	FP+TP	N

Here, are the parameters such as Accuracy, True Positive Rate, True Negative Rate, Precision and F_Measure which compare the performance of the existing and the proposed techniques.

These parameters allow us to evaluate the performance of the Machine Learning (ML) techniques quantitatively. Therefore, these parameters have been used to evaluate the effectiveness of the proposed technique over other techniques.

7.2.1 Visual Analysis

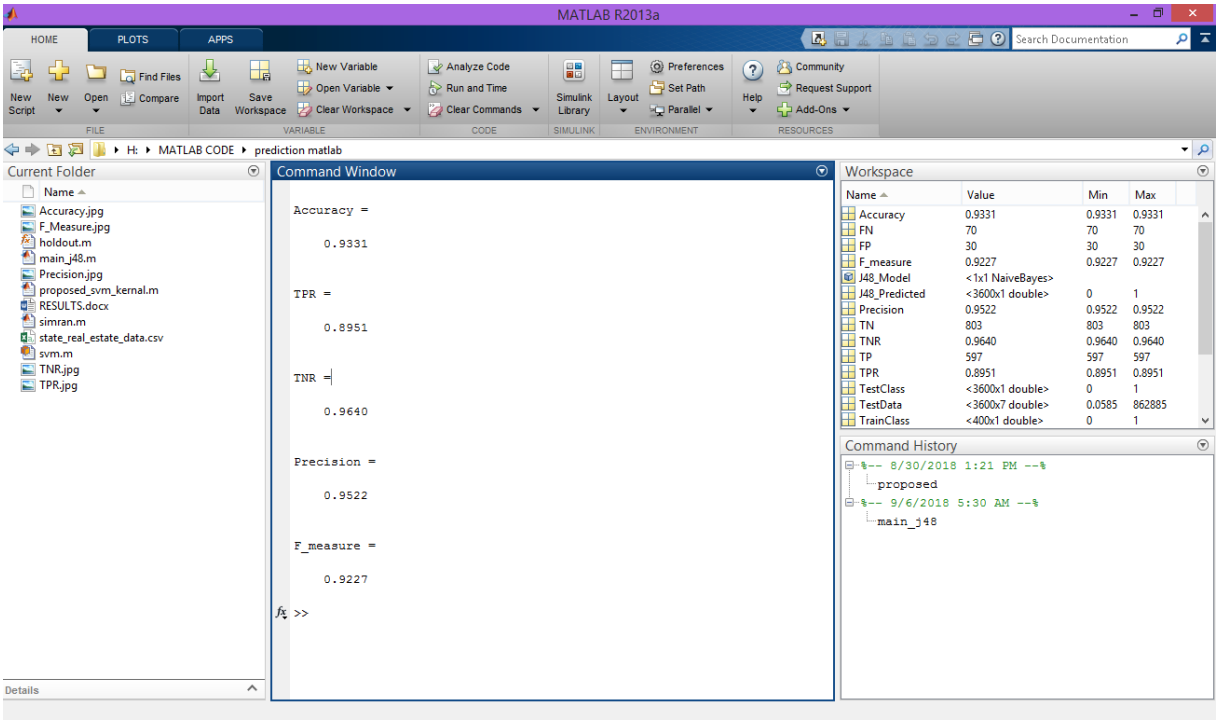


Figure 7.8: Evaluated J48 Model

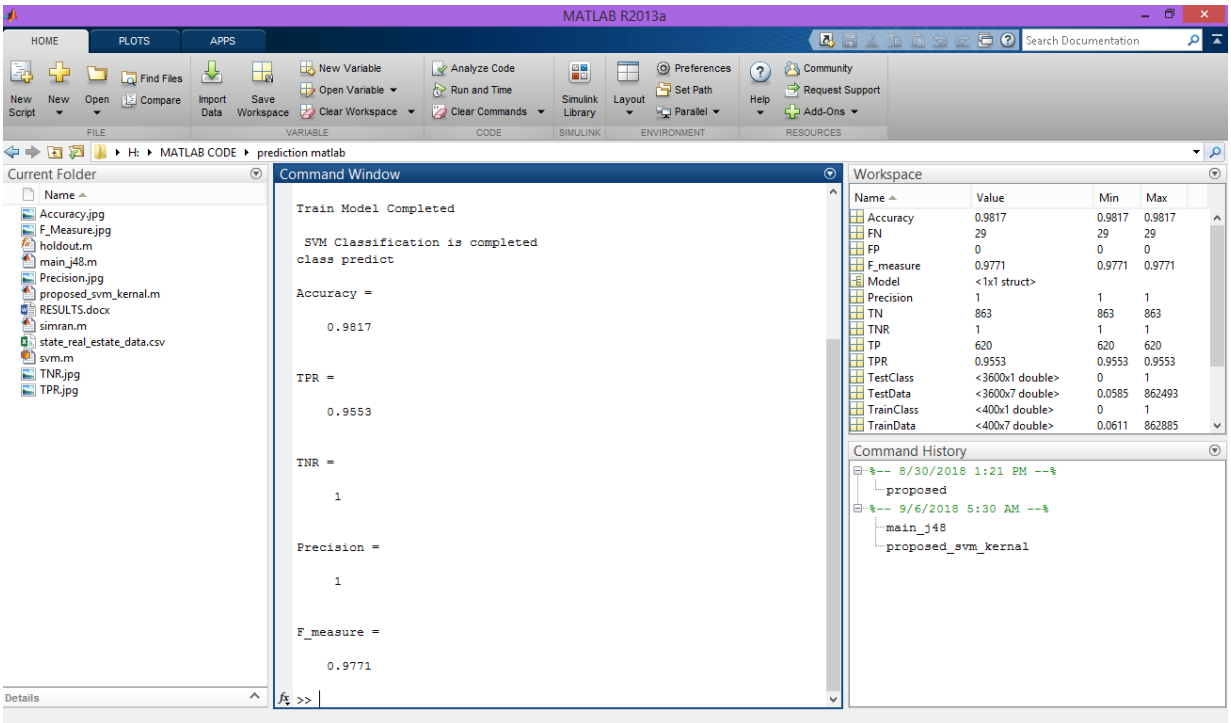


Figure 7.9: Evaluated Support Vector Machine (SVM) Model

7.2.2 Quantitative Analysis

This section covers the cross authentication between existing and proposed techniques. Some familiar algorithms parameters have been chosen to show that the performance of the proposed algorithm is superior to the existing techniques.

For experimentation and implementation, the proposed technique is evaluated using MATLAB tool u2013a and statistics & machine learning toolbox. Here we will compare the parameters of existing with proposed algorithm i.e. merging of SVM and J48 decision tree.

The tabular and graphical comparison has been done between existing and proposed methodology on the basis of parameters like TP-Rate, TN-Rate, Accuracy, F measure and Precision.

7.2.2.1 ACCURACY

Accuracy is the proportion of true results among number of cases. Accuracy is also called as rand accuracy or rand index. Accuracy is measured with respect to reality. Accuracy is calculated by following equation. In which true positive (TP), true negative (TN) and false positive (FP), false negative (FN) is considered for the calculation.

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN})} \times \mathbf{100}$$

Table 7.4 is indicated about quantized research into the Accuracy. As Accuracy ought to be higher which implies proposed algorithm is indicating the superior results when compared to access methods as the Accuracy is higher in each case.

Figure.7.10: indicates about comparison of Accuracy between existing and the proposed method wherever x-axis indicates size of training data as well as y- axis indicates Accuracy. Here, red line indicates the proposed technique and blue line indicate the previous one. In our case the proposed Accuracy are comparatively higher than existing one.

Table 7.4: Comparative analysis of Accuracy

PERCENTAGE OF TRAINING DATA	EXISTING TECHNIQUE (J48)	PROPOSED TECHNIQUE (SVM)
10	0.9500 $\pm$ 0.003	0.9825 $\pm$ 0.0435
20	0.9387 $\pm$0.005	0.9825 $\pm$ 0.0523
30	0.9550 $\pm$0.004	0.9917 $\pm$ 0.0256
40	0.9463 $\pm$ 0.002	0.9912 $\pm$ 0.0221
50	0.9560 $\pm$ 0.018	0.9830 $\pm$ 0.0126
60	0.9529 $\pm$ 0.051	0.9833 $\pm$ 0.0568
70	0.9468 $\pm$ 0.048	0.9850 $\pm$ 0.0156
80	0.9459 $\pm$ 0.045	0.9850 $\pm$ 0.0459
90	0.9414	0.9756

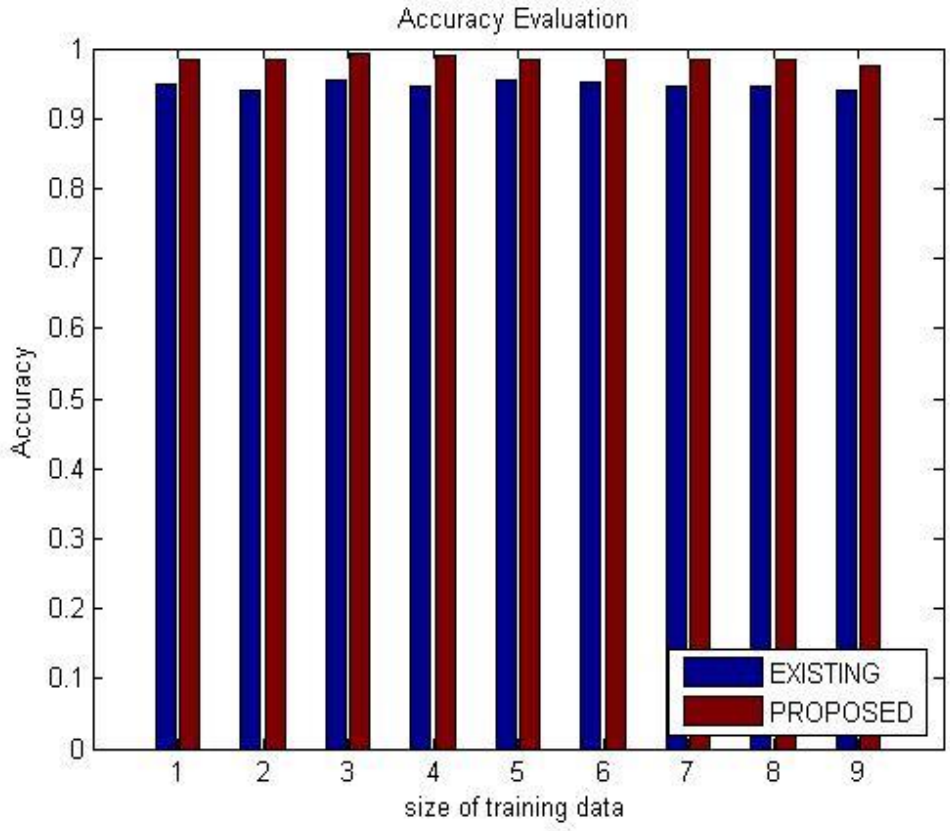


Figure 7.10: Accuracy Evaluation

7.2.2.2 TRUE POSITIVE RATE

True positive rate are needed functions of classifier. It defines as how many correct positive occur among all positive results. It is known as the hypothesis of the correct results that has been configured during the system working. True positive rate is calculated as:

$$\boldsymbol{True\ Positive\ Rate} = \frac{\boldsymbol{TP}}{\boldsymbol{TP}} + \boldsymbol{TN}$$

Table 7.5: Comparative analysis of True Positive Rate

PERCENTAGE OF TRAINING DATA	EXISTING TECHNIQUE (J48)	PROPOSED TECHNIQUE (SVM)
10	0.9143 ± 0.053	0.9692 ±
20	0.9518 ±0.023	0.9662 ±
30	0.9139 ±0.059	0.9909 ±
40	0.9000 ±0.056	0.9565 ±
50	0.9577 ±0.022	0.9456 ±
60	0.9404 ±0.056	0.9688 ±
70	0.9157 ±0.054	0.9713 ±
80	0.9119 ±0.045	0.9785 ±
90	0.9130 ±0.047	0.9739 ±

Table 7.5 is indicated about quantized research into the True positive rate. As True positive rate ought to be higher which implies proposed algorithm is indicating the superior results when compared to access methods as the True positive rate is higher in each case. Figure.7.11: indicates about comparison of True positive rate between existing and the proposed method wherever x-axis indicates size of training data as well as y- axis indicates True positive rate. Here, red line indicates the proposed technique and blue line indicate the previous one. In our case the proposed True positive rate are comparatively higher than existing one.

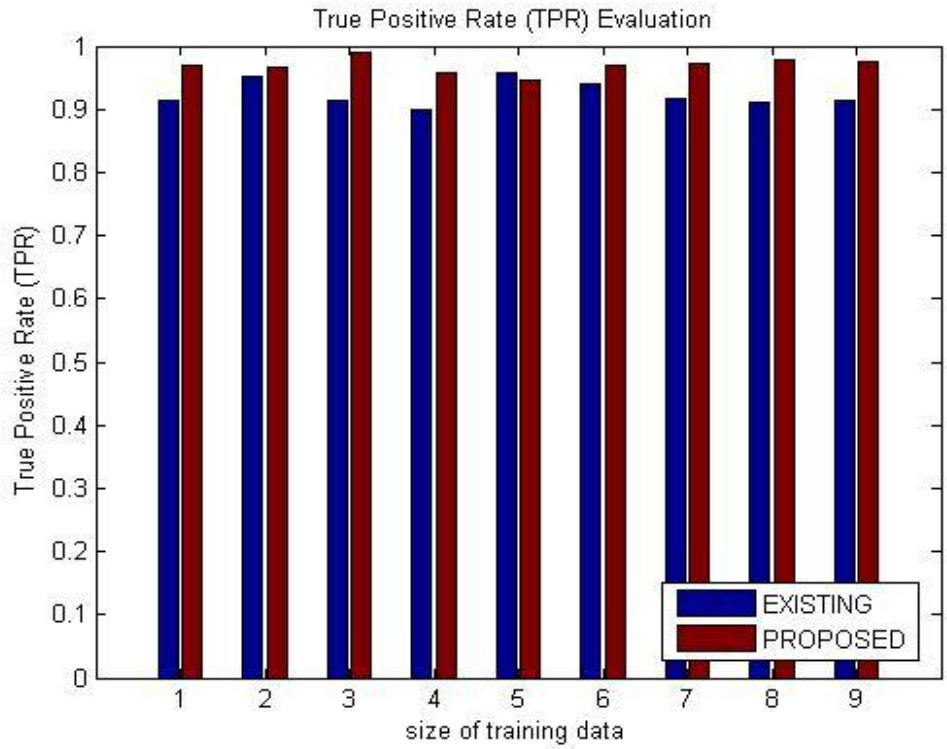


Figure 7.11: True positive rate Evaluation

7.2.2.3 TRUE NEGATIVE RATE

True negative rate is amount of negatives that are correctly identified as negatives. It defines as how many negative occurs among the all results. These are the amount of outcomes that are predicted negative but actually negative. The formula for calculating true negative rate is:

$$\boldsymbol{True\ Negative\ Rate} = \frac{\boldsymbol{TN}}{(\boldsymbol{FP} + \boldsymbol{TN})}$$

Table 7.6 is indicated about quantized research into the True negative rate. As True negative rate ought to be higher which implies proposed algorithm is indicating the superior results when compared to access methods as the True negative rate is higher in each case. Figure.7.12: indicates about comparison of True negative rate between existing and the proposed method wherever x-axis indicates size of training data as well as y- axis indicates True negative rate. Here, red line indicates the proposed technique and blue line indicate the previous one. In our case the proposed True negative rate are comparatively higher than existing one.

Table 7.6: Comparative analysis of True Negative Rate

PERCENTAGE OF TRAINING DATA	EXISTING TECHNIQUE (J48)	PROPOSED TECHNIQUE (SVM)
10	0.9667 ±0.076	0.9998 ±0.049
20	0.9458 ± 0.749	0.9997 ± 0.014
30	0.9858 ± 0.016	0.9899 ± 0.046
40	0.9728 ± 0.016	0.9979 ± 0.049
50	0.9710 ± 0.079	0.9349 ± 0.079
60	0.9673 ± 0.046	0.9930 ± 0.079
70	0.9703 ± 0.046	0.9985 ± 0.046
80	0.9826 ± 0.016	0.9961 ± 0.079
90	0.9599 ± 0.019	0.9846 ± 0.046

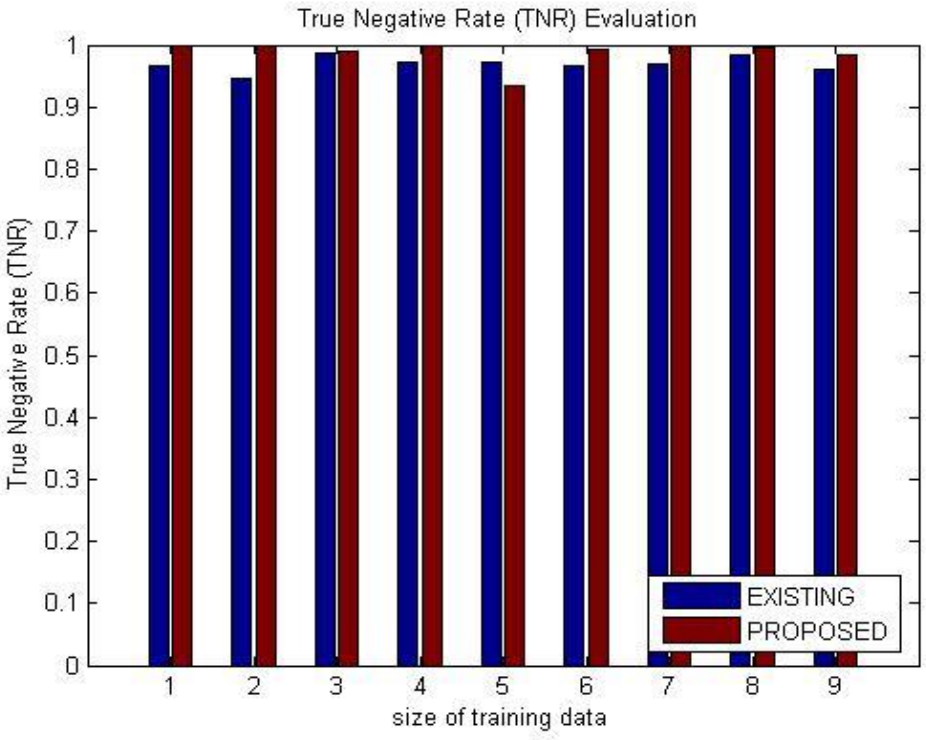


Figure 7.12: True negative rate Evaluation

7.2.2.4 F-MEASURE

F-Measure is also called F1 score. It contains both precision and recall. It is generally use to check the accuracy and reliability. It computes the mean of precision and recall. Basically, it uses as best and 0 as worst when both precision and recall are used. F-measure can be calculated with using the formula given as:

$$\mathbf{F-Measure} = 2 * \frac{\mathbf{P} * \mathbf{R}}{\mathbf{P} + \mathbf{R}}$$

Table 7.7: Comparative analysis of F_Measure

PERCENTAGE OF TRAINING DATA	EXISTING TECHNIQUE (J48)	PROPOSED TECHNIQUE (SVM)
10	0.9343 ± 0.049	0.9844 ± 0.016
20	0.9433 ± 0.049	0.9828 ± 0.079
30	0.9455 ± 0.045	0.9954 ± 0.046
40	0.9305 ± 0.016	0.9778 ± 0.046
50	0.9492 ± 0.019	0.9720 ± 0.003
60	0.9480 ± 0.016	0.9793 ± 0.016
70	0.9363 ± 0.016	0.9844 ± 0.046
80	0.9430 ± 0.045	0.9866 ± 0.016
90	0.9282 ± 0.016	0.9753 ± 0.016

Table 7.7 is indicated about quantized research into the F-Measure. As F-Measure ought to be higher which implies proposed algorithm is indicating the superior results when compared to access methods as the F-Measure is higher in each case. Figure.7.13: indicates about comparison of F-Measure between existing and the proposed method wherever x-axis indicates size of training data as well as y- axis indicates F-Measure. Here, red line indicates the proposed technique and blue line indicate the previous one. In our case the proposed F-Measure are comparatively higher than existing one.

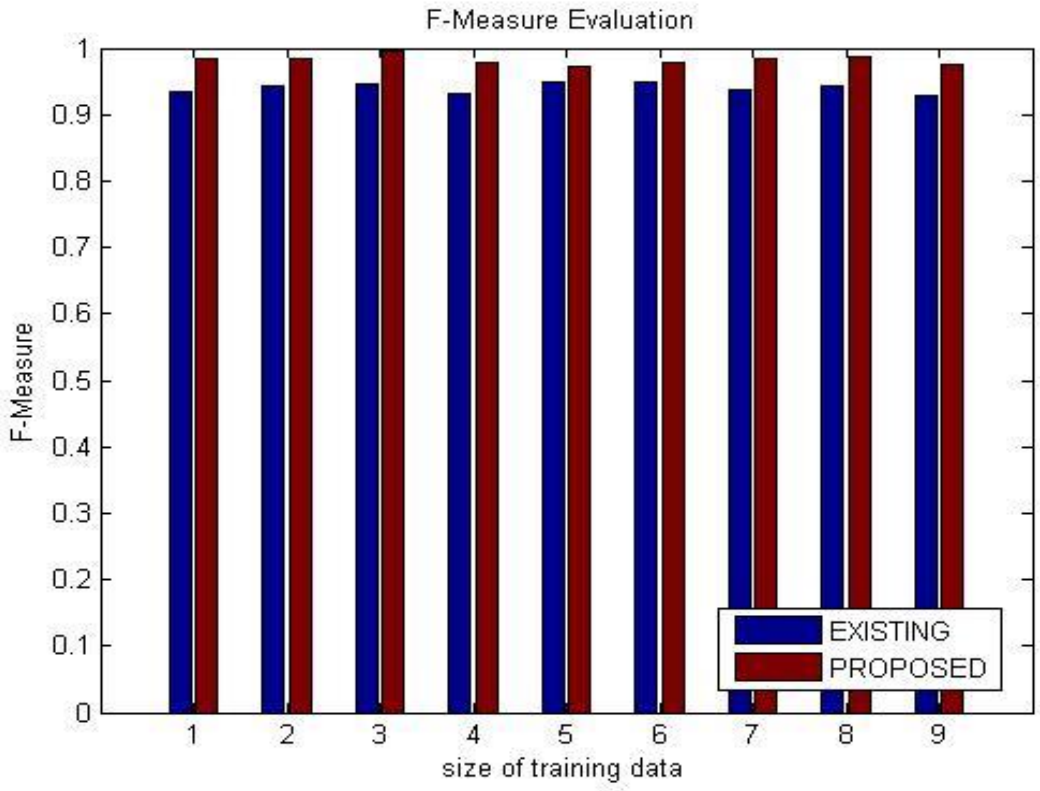


Figure 7.13: F-Measure Evaluation

7.2.2.5 PRECISION

Precision is defined as measurement of all positive cases that are identified when making calculations. Precision is also known as positive predictive value. Higher Precision signifies that an algorithm significantly returned more relevant results when compared to irrelevant. Precision can be calculated by using the formula:

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{FP} + \textbf{TP}}$$

Table 7.8 is indicated about quantized research into the Precision. As Precision ought to be higher which implies proposed algorithm is indicating the superior results when compared to access methods as the Precision is higher in each case. Figure.7.14: indicates about comparison of Precision between existing and the proposed method wherever x-axis indicates size of training data as well as y- axis indicates Precision. Here, red line indicates the proposed technique and blue line indicate the previous one. In our case the proposed Precision are comparatively higher than existing one.

Table 7.8: Comparative analysis of Precision

PERCENTAGE OF TRAINING DATA	J48	PROPOSED SVM
10	0.9552 ± 0.003	0.9982 ± 0.0013
20	0.9349 ± 0.004	0.9987 ± 0.003
30	0.9795 ± 0.009	0.9934 ± 0.013
40	0.9631 ± 0.013	0.9987 ± 0.045
50	0.9591 ± 0.012	0.9874 ± 0.049
60	0.9557 ± 0.045	0.9902 ± 0.003
70	0.9580 ± 0.003	0.9979 ± 0.079
80	0.9764 ± 0.097	0.9949 ± 0.046
90	0.9438 ± 0.006	0.9769 ± 0.078

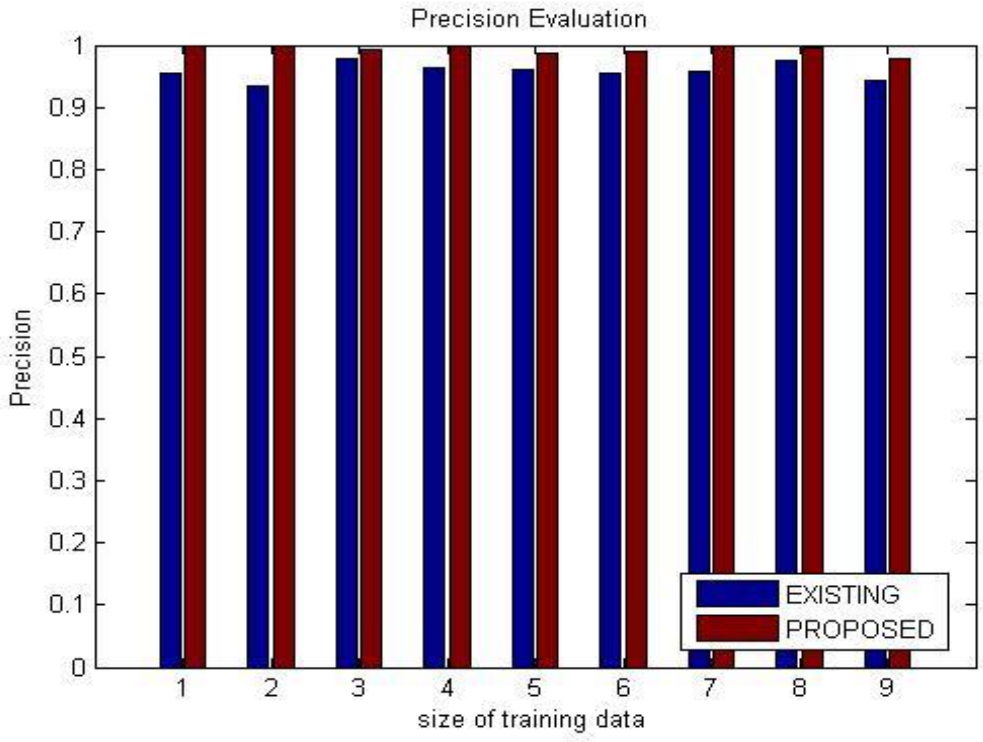


Figure 7.14: Precision Evaluation

7.3 Prescriptive Analysis

This section covers the cross authentication between existing and proposed techniques. For experimentation and implementation, the proposed technique is evaluated using MATLAB tool u2013a and statistics & machine learning toolbox. Here we will compare the parameters of existing with proposed algorithm i.e. merging of J48 Decision tree, SVM and NSGA-III.

7.3.1 Visual Analysis

The purpose of this section is to recognize and understand the visual outputs by observing the code implementation. Figure 7.15 shows the evaluated results for optimized accuracy and other parameters.

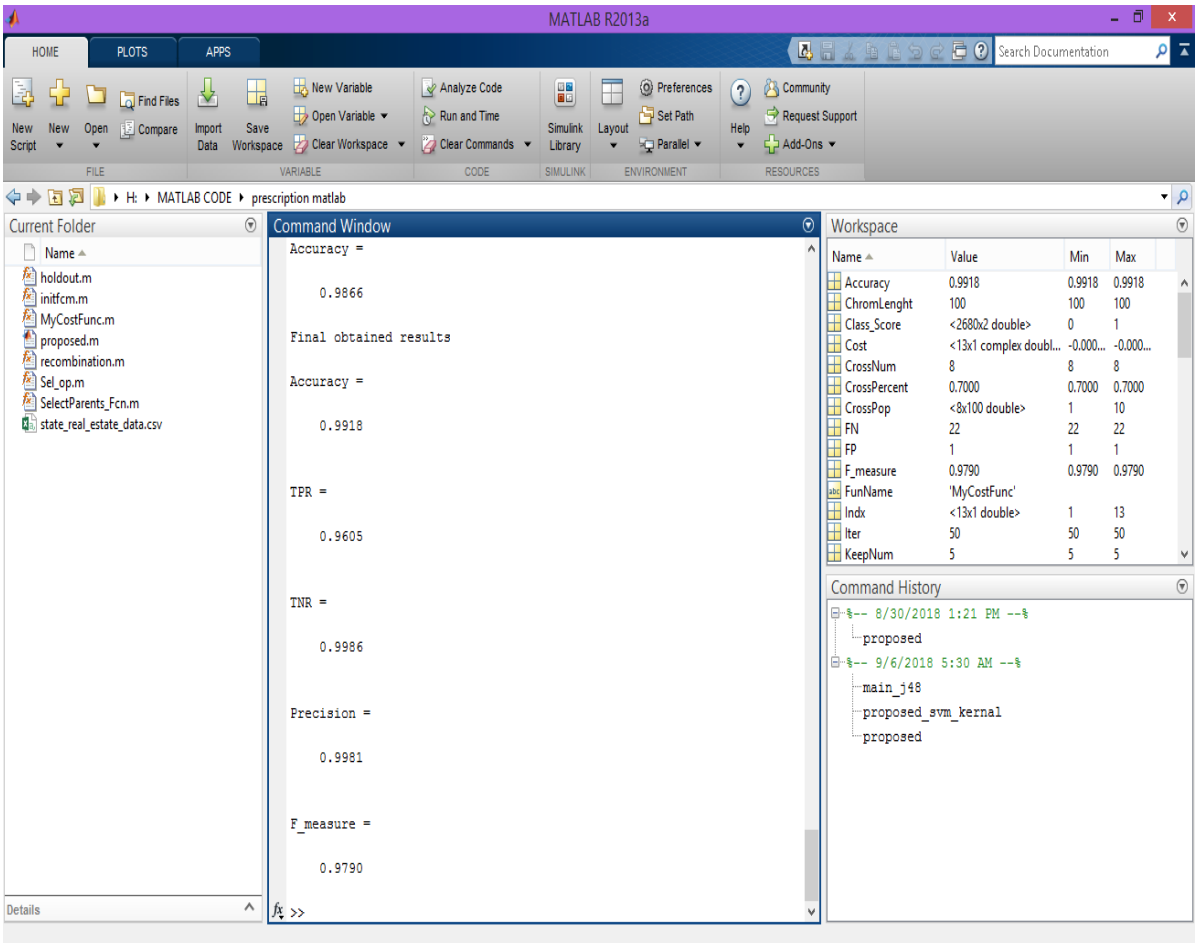


Figure 7.15: Evaluated NSGA-III Model

7.3.2 Quantitative Analysis

This section covers the cross authentication between existing and proposed techniques. Some familiar algorithms parameters have been chosen to show that the performance of the proposed algorithm is superior to the existing techniques.

For experimentation and implementation, the proposed technique is evaluated using MATLAB tool u2013a and statistics & machine learning toolbox. Here we will compare the parameters of existing with proposed algorithm i.e. merging of SVM and J48 decision tree.

The tabular and graphical comparison has been done between existing and proposed methodology on the basis of parameters like TP-Rate, TN-Rate, Accuracy, F measure and Precision.

7.3.2.1 Accuracy analysis

Accuracy is the proportion of true results among a number of cases. Accuracy is also called as rand accuracy or rand index. Accuracy is measured with respect to reality. Accuracy is calculated using equation. In which True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) is considered for the calculation.

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN})} \times \mathbf{100}$$

In Table 7.9, the first value that is 0.9525 indicates the mean value of the solution when testing is performed 30 times with J48 model. 0.003 indicates variance between first 30 outcomes that are performed using J48. 0.9850 indicates the mean value of the solution when testing is performed 30 times with SVM model.

0.007 indicates variance between 30 outcomes that are performed using SVM. 0.9918 indicates the mean value of the solution when testing is performed 30 times with NSGA-III model.

0.004 indicates variance between 30 outcomes that are performed using NSGA-III. All the three are performed on 10 percent training data and 90 percent testing data.

Table 7.9: Comparative Analysis of Accuracy

Percentage of Training Dataset	J48	SVM	NSGA-III
10	0.9525±0.003	0.9850±0.007	0.9918±0.004
20	0.9563±0.087	0.9850±0.005	0.9912±0.003
30	0.9500±0.004	0.9842±0.004	0.9925±0.008
40	0.9444±0.001	0.9850±0.002	0.9922±0.009
50	0.9505±0.005	0.9885±0.001	0.9917±0.009
60	0.9429±0.009	0.9871±0.003	0.9912±0.007
70	0.9489±0.007	0.9850±0.002	0.9924±0.004
80	0.9437±0.004	0.9841±0.003	0.9928±0.005
90	0.9383±0.003	0.9753±0.004	0.9920±0.008

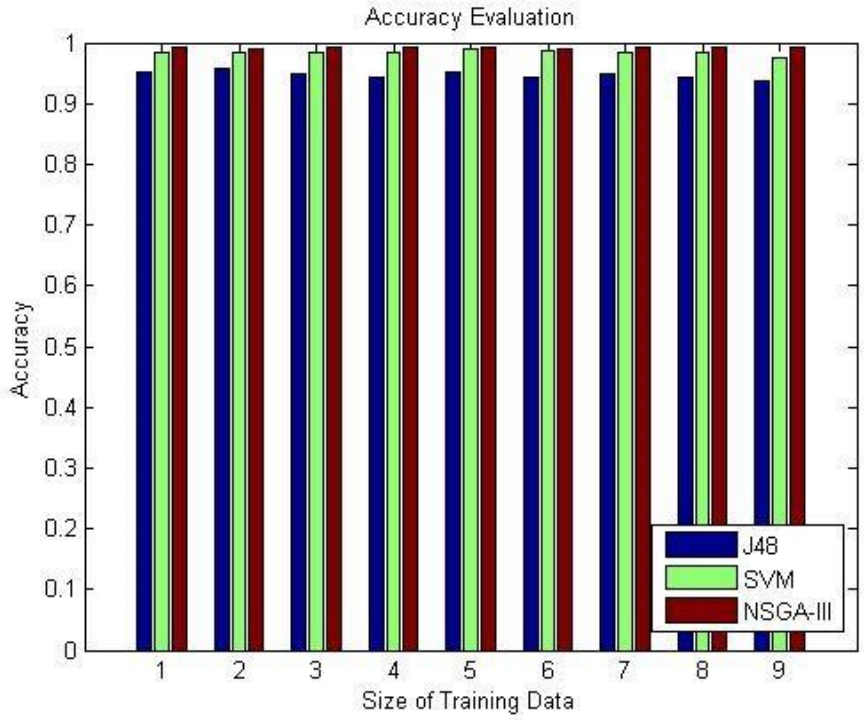


Figure 7.16: Accuracy Evaluation

7.2.2.2 True Positive Rate Analysis

The true positive rate is needed functions of the classifier. It defines as how many correct positive occur among all positive results. It is known as the hypothesis of the correct results that have been configured during the system working. The true positive rate is calculated using equation:

$$\boldsymbol{True\ Positive\ Rate} = \frac{\boldsymbol{TP}}{\boldsymbol{TP}} + \boldsymbol{TN}$$

In Table 7.10, the first value that is 0.9254 indicates the mean value of the solution when testing is performed 30 times with J48 model. 0.003 indicates variance between first 30 outcomes that are performed using J48. 0.9231 indicates the mean value of the solution when testing is performed 30 times with SVM model. 0.003 indicates variance between 30 outcomes that are performed using SVM. 0.9605 indicates the mean value of the solution when testing is performed 30 times with NSGA-III model. 0.009 indicates variance between 30 outcomes that are performed using NSGA-III. All the three are performed on 10 percent training data and 90 percent testing data.

Table 7.10: Comparative Analysis of True Positive Rate

Percentage of Training Dataset	**J48**	**SVM**	**NSGA-III**
10	0.9254±0.003	0.9231±0.003	0.9605±0.009
20	0.9296±0.089	0.9542±0.004	0.9579±0.009
30	0.9342±0.002	0.9611±0.006	0.9587±0.009
40	0.9272±0.002	0.9593±0.006	0.9605±0.002
50	0.9235±0.006	0.9724±0.007	0.9607±0.002
60	0.9033±0.002	0.9788±0.009	0.9579±0.008
70	0.9254±0.007	0.9832±0.006	0.9592±0.002
80	0.9202±0.009	0.9692±0.008	0.9592±0.001
90	0.9227±0.007	0.9754±0.001	0.9592±0.005

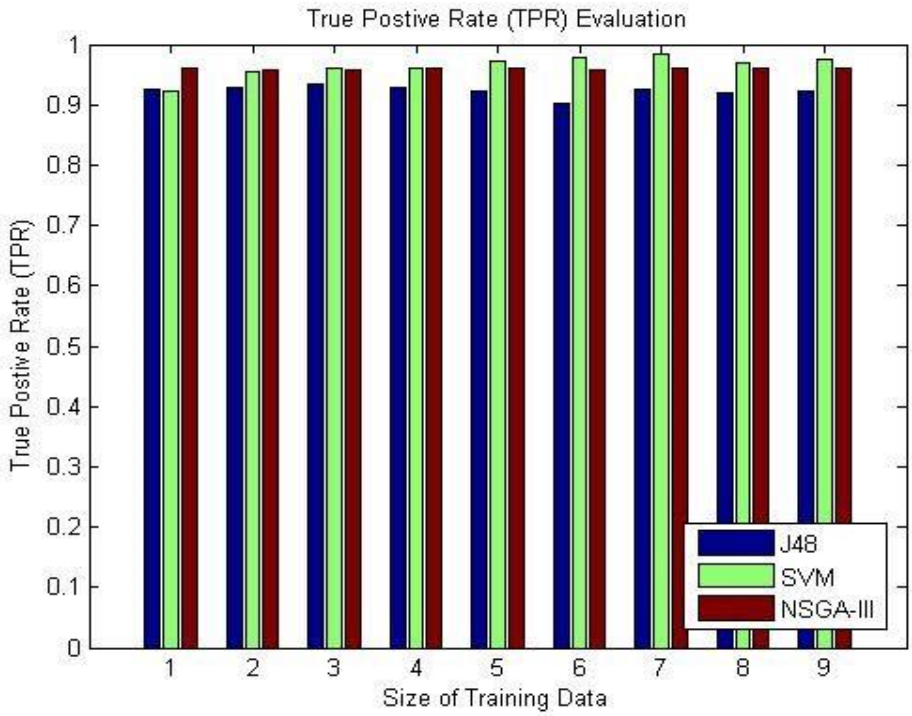


Figure 7.17: True Positive Rate Evaluation

7.2.2.3 True Negative Rate analysis

The True Negative Rate is analyzed for the proposed and existing competitive techniques. It shows that the proposed technique outperforms others because it has better sensitivity rate compared to other techniques. It is calculated using equation:

$$\boldsymbol{True\ Negative\ Rate} = \frac{\boldsymbol{TN}}{(\boldsymbol{FP} + \boldsymbol{TN})}$$

In Table 7.11, the first value that is 0.9688 indicates the mean value of the solution when testing is performed 30 times with J48 model. 0.002 indicates variance between first 30 outcomes that are performed using J48. 0.9899 indicates the mean value of the solution when testing is performed 30 times with SVM model.

0.001 indicates variance between 30 outcomes that are performed using SVM. 0.9986 indicates the mean value of the solution when testing is performed 30 times with NSGA-III model. 0.087 indicates variance between 30 outcomes that are performed using NSGA-III. All the three are performed on 10 percent training data and 90 percent testing data.

Table 7.11: Comparative Analysis of True Negative Rate

Percentage of Training Dataset	J48	SVM	NSGA-III
10	0.9688±0.002	0.9899±0.001	0.9986±0.087
20	0.9626±0.005	0.9999±0.006	0.9984±0.008
30	0.9820±0.009	0.9542±0.020	0.9986±0.009
40	0.9601±0.009	0.9974±0.006	0.9986±0.008
50	0.9660±0.006	0.9979±0.087	0.9985±0.005
60	0.9754±0.076	0.9965±0.003	0.9984±0.077
70	0.9583±0.033	0.9954±0.050	0.9986±0.021
80	0.9618±0.061	0.9946±0.060	0.9986±0.003
90	0.9532±0.003	0.9720±0.005	0.9986±0.006

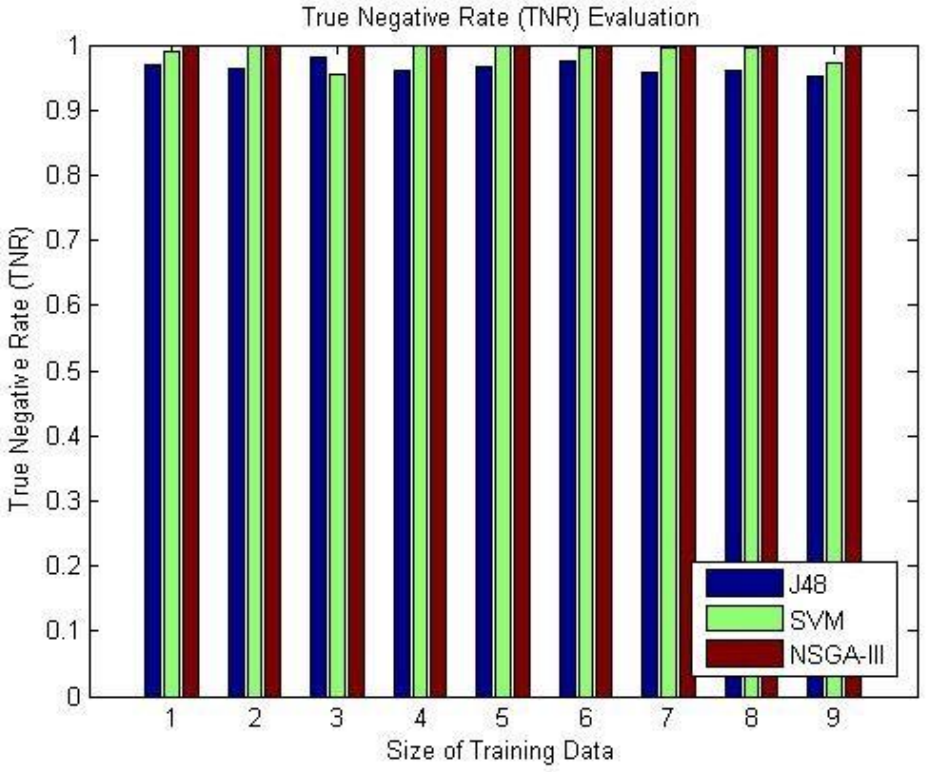


Figure 7.18: True Negative Rate Evaluation

7.2.2.4 Precision Evaluation

The Precision is analyzed for the proposed and existing techniques. It is found that the proposed Machine Learning (ML) technique outperforms others in terms of specificity because it has better specificity values compared to other techniques. It is calculated using equation:

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{FP} + \textbf{TP}}$$

Table 7.12: Comparative Analysis of Precision

Percentage of Training Dataset	**J48**	**SVM**	**NSGA-III**
10	0.9688±0.002	0.9899±0.001	0.9986±0.087
20	0.9626±0.005	0.9999±0.006	0.9984±0.008
30	0.9820±0.009	0.9542±0.020	0.9986±0.009
40	0.9601±0.009	0.9974±0.006	0.9986±0.008
50	0.9660±0.006	0.9979±0.087	0.9985±0.005
60	0.9754±0.076	0.9965±0.003	0.9984±0.077
70	0.9583±0.033	0.9954±0.050	0.9986±0.021
80	0.9618±0.061	0.9646±0.060	0.9986±0.003
90	0.9532±0.003	0.9720±0.005	0.9986±0.006

In Table 7.12 the first value that is 0.9688 indicates the mean value of the solution when testing is performed 30 times with J48 model. 0.002 indicates variance between first 30 outcomes that are performed using J48.

0.9899 indicates the mean value of the solution when testing is performed 30 times with SVM model. 0.001 indicates variance between 30 outcomes that are performed using SVM.

0.986 indicates the mean value of the solution when testing is performed 30 times with NSGA-III model. 0.087 indicates variance between 30 outcomes that are performed using NSGA-III. All the three are performed on 10 percent training data and 90 percent testing data.

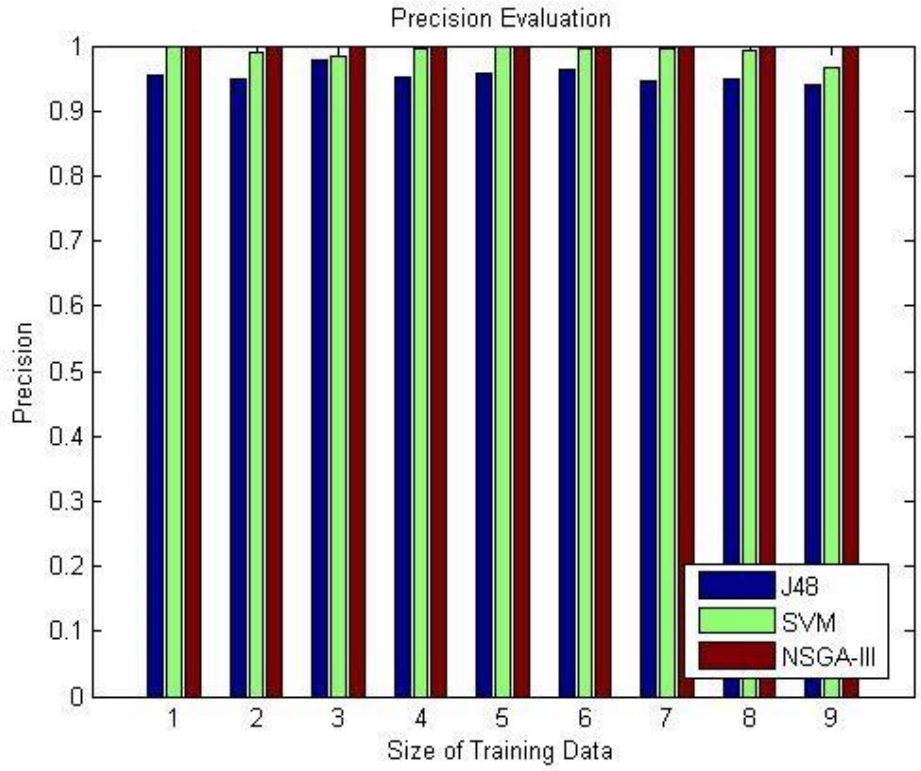


Figure 7.19: Precision Evaluation

7.2.2.5 F-Measure analysis

F-measure is also called the F1 score. It contains both precision and recall. It is generally used to check the accuracy and reliability. It computes the mean of precision and recall. Basically, it uses 1 as best and 0 as worst when both precision and recall are used. F-measure can be calculated by using the formula using equation:

$$\mathbf{F-Measure} = \mathbf{2} * \frac{\mathbf{P} * \mathbf{R}}{\mathbf{P} + \mathbf{R}}$$

Table 7.13: Comparative Analysis of F-Measure

Percentage of Training Dataset	J48	SVM	NSGA-III
10	0.9394±0.009	0.9600±0.007	0.9790±0.001
20	0.9395±0.006	0.9466±0.005	0.9775±0.007
30	0.9552±0.007	0.9802±0.001	0.9780±0.009
40	0.9391±0.008	0.9775±0.006	0.9790±0.008
50	0.9409±0.006	0.9846±0.006	0.9791±0.085
60	0.9330±0.007	0.9869±0.003	0.9775±0.082
70	0.9353±0.007	0.9888±0.003	0.9782±0.092
80	0.9345±0.007	0.9810±0.006	0.9782±0.062
90	0.9310±0.007	0.9694±0.003	0.9782±0.002

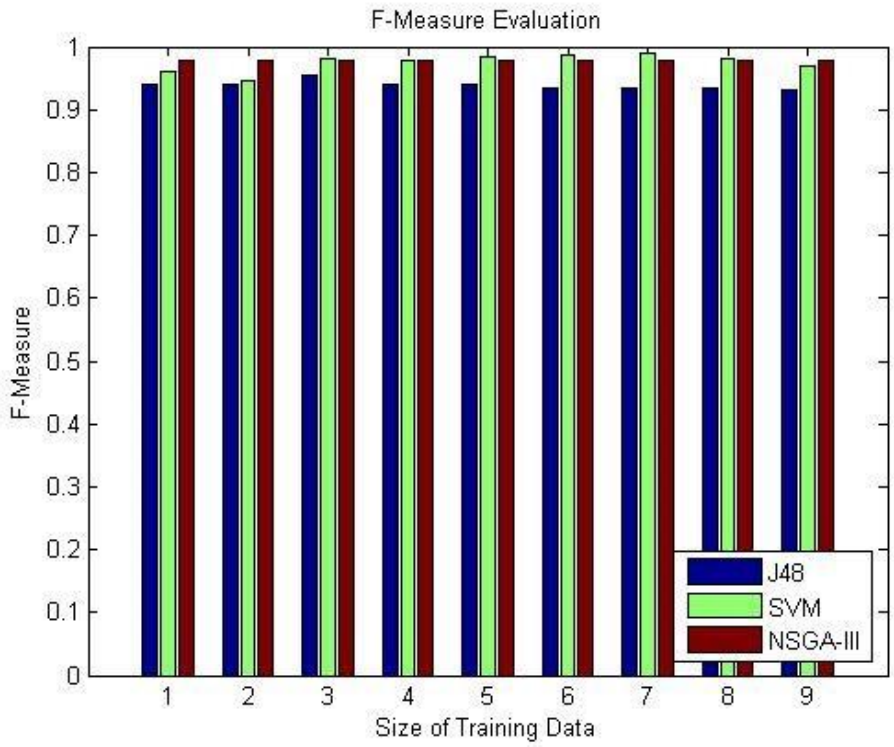


Figure 7.20: F-measure Evaluation

In Table 7.13, the first value that is 0.9394 indicates the mean value of the solution when testing is performed 30 times with J48 model. 0.009 indicates variance between first 30 outcomes that are performed using J48. 0.9600 indicates the mean value of the solution when testing is performed 30 times with SVM model. 0.007 indicates variance between 30 outcomes that are performed using SVM. 0.9790 indicates the mean value of the solution when testing is performed 30 times with NSGA-III model. 0.001 indicates variance between 30 outcomes that are performed using NSGA-III. All the three are performed on 10 percent training data and 90 percent testing data.

CHAPTER 8

CONCLUSION AND FUTURE WORK

In this work, to do cubing without OLAP is obviously to compose SQL queries that concentrate the outcome sets (i.e., wanted) and that contains similar information (i.e., would come about because of identical OLAP activities). There are anyway some significant disadvantages with this approach. As a matter of first importance, the execution would be inadmissible when the database is extensive with numerous relations required. Although, the tests were performed with real-time estimations and the review response time could be advanced with emerged. On the other hand, the questions were executed against the STAR schema formed data distribution center, which is loaded with excess information to limit the number of joins required.

To do similar inquiries against a standardized information source would break down execution. In any case, since the OLAP devices are particularly produced for these sort of questions, they are obviously improved for short question reaction times. A portion of these advancements exploit the read for the most part nature of OLAP models and can scarcely be found in a generally useful social database motor. Second, the announcing would be constrained. An awesome preferred standpoint of OLAP devices is that the client sees is multidimensional and the detailing is extremely adaptable.

OLAP is exceptionally adaptable with both section and line marks, and regardless of whether it isn't so normal, revealing in excess of two measurements is completely conceivable. Adding to this the bore down and move up tasks makes these sort of devices better than social databases with regards to breaking down information. Obviously, the requirement for multidimensional information examination for a little association with a constrained database may not require all the broad limit of OLAP devices, which regularly are costly despite the fact that there are open source options for BI arrangements also.

This work proposes a real-estate mining process that is performed with the aid of J48 and Support Vector Machine (SVM) classification technique. Here, input dataset is

high dimensional real-estate data which is a great barrier for classification. Therefore, initially feature dimension reduction using KPIs have been applied to reduce features space without losing the accuracy of classification. Here, unitary method has been used for selecting basic features from primary (self-created) dataset and secondary (taken from Kaggle website) datasets. Once the feature reduction is performed, the classification is applied based on J48 Decision tree and Support Vector Machine (SVM) classifier. This is how, the obtained data is transformed into a classification problem that states whether the property has been purchased or not. To train the classification data, J48 and SVM have been implemented. Although, these models perform significantly to classify real estate purchasing, but, suffer from the parameter tuning issue.

This issue has been solved by considering the well-known technique i.e., Non-dominated Sorted Genetic Algorithm – III (NSGA-III) based Meta-Decision tree. It iteratively optimizes the meta-J48 model to improve the classification rate by considering mutation and crossover operator. The obtained solutions that are non-dominated in nature, therefore, the proposed model can provide better accuracy as well as other parameters such as true positive rate, true negative rate, precision and f_measure concurrently. Extensive experiments have been performed. It has been found that the proposed technique for optimization outperforms in terms of Accuracy, TP-Rate, TN-Rate, Precision, and F_Measure. Therefore, the proposed technique is applicable for real-time real-estate users.

There are various directions in which this proposed work can be carried forward in future. Firstly, one can extend this work for forecasting house prices. Secondly, parameter tuning can also be done for other machine learning techniques such as SVM, J48, and Neural network.

REFERENCES

[1] Larose, & Daniel T., "Data mining methods & models", *John Wiley & Sons*, 2006.

[2] Larose, Daniel T., & Chantal D. Larose, "Discovering knowledge in data: an introduction to data mining", *John Wiley & Sons*, 2014.

[3] Wu, Xindong, "Data mining with big data", *IEEE transactions on knowledge and data engineering*, vol. 26, no.1, pp. 97-107, 2014.

[4] Kimball, "R. Data Warehouse (DW) Toolkit", 3rd Edition, 1996.

[5] Morgan Kaufinann, "Data Mining: Concepts and Techniques", 2006.

[6] Gnanapriya, S., "Data mining concepts and techniques." *Data Mining and Knowledge Engineering*, vol. 2, no. 9, pp. 256-263, 2010.

[7] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. "Data Mining: Practical Machine Learning (ML) tools and techniques" *Morgan Kaufmann*, 2016.

[8] Lei-da Chen, Toru Sakaguchi, and Mark N. Frolick. "Data mining methods, applications, and tools." *Information systems management*, vol. 17, no. 1, pp. 67-68, 2011.

[9] Koh, Hian Chye & Gerald Tan, "Data mining applications in healthcare", *Journal of healthcare information management",* vol. 19, no.2, pp. 65, 2011.

[10] Videla-Cavieres, Ivan F., & Sebastian A. Rios, "Extending market basket analysis with graph mining techniques", *Expert Systems with Applications*, vol. 41, no. 4 pp. 1928-1936, 2014.

[11] Berland, Matthew, Ryan S. Baker, & Paulo Blikstein, "Educational data mining and learning analytics: Applications to constructionist research", *Technology, Knowledge and Learning*, vol. 19, no.1-2, pp. 205-220, 2014.

[12] Khodakarami, Farnoosh & Yolande E. Chan, "Exploring the role of customer relationship management (CRM) systems in customer knowledge creation", *Information &* Management, vol. 51 no. 1, pp. 27-42, 2014

[13] Feng, Wenying "Mining network data for intrusion detection through combining SVMs with ant colony networks", *Future Generation Computer Systems*, vol. 37 pp. 127-140, 2014.

[14] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", *Information Sciences*, vol. 275, pp. 314-347, 2014.

[15] Alpaydin & Ethem, "Introduction to Machine Learning (ML)" *MIT press*, 2014.

[16] Dougherty, James, Ron Kohavi, and Mehran Sahami. "Supervised and unsupervised discretization of continuous features", *Machine Learning (ML) Proceedings*, pp. 194-202, 2008.

[17] Galatzer-Levy & Isaac R, "Quantitative forecasting of PTSD from early trauma responses: A Machine Learning (ML) application", *Journal of psychiatric research* 59 pp. 68-76, 2014.

[18] Deng, Li, and Dong Yu, "Deep learning: methods and applications." *Foundations and Trends® in Signal* Processing, vol. 7, no. 3–4 pp. 197-387, 2014.

[19] Schmidtler, Mauritius AR, Roland Borrey, & Anthony Sarah. "Data classification using Machine Learning (ML) techniques." No. 8, pp. 719,197, 2014.

[20] Kornelson, K.P., Vajjiravel, M., Prasad, R., Clark, P.D. & Najm, T., "Method and system for developing Extract Transform Load (ETL) systems for Data Warehouse (DW)s." Microsoft Corp, 2006.

[21] Leonardi, L., Orlando, S., Raffaetà, A., Roncato, A., Silvestri, C., Andrienko, G., & Andrienko, N., "A general framework for trajectory data warehousing and visual OLAP" *GeoInformatica*, vol. 18, no. 2, pp. 273-312, 2014.

[22] Mavroforakis, Michael E., and Sergios Theodoridis. "A geometric approach to support vector machine (SVM) classification." *IEEE transactions on neural networks*, vol. 17, no. 3 pp. 671-682, 2006.

[23] Keogh & Eamonn. "Naive Bayes classifier", 2017

[24] Bhargava &Neeraj, "Decision tree analysis on j48 algorithm for data mining", *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3 no .6, 2013

[25] Mkaouer, Wiem, et al. "Many-objective software remodularization using NSGA-III." *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 24 no. 3, 2017.

[26] Mühlenbein, Heinz, M. Schomisch & Joachim Born, "The parallel genetic algorithm as function optimizer", *Parallel Computing*, vol. 17 no. 6-7, pp. 619-632, 2012.

[27] Crossover Christensen, Pernille H., Spenser Robinson, and Robert A. Simons. "The application of mixed methods: using a crossover analysis strategy for product development in real estate", *Journal of Real Estate Literature*, vol. 24, no. 2, pp. 429-451, 2016.

[28] Ying-Ping, A., & J. Hu., "OLAP application in enterprise marketing management system", *3rd International Conference on Systems and Informatics, ICSAI*, pp. 588-592, 2017.

[29] Wang, P., H. Zareipour, & W. Rosehart., "Descriptive models for reserve and regulation prices in competitive electricity markets", *IEEE Transactions on Smart Grid*, pp. 471-479, 2014.

[30] He, Z., & B. LIN., "Application of Data Warehouse (DW) and OLAP in the merchandising system", *Proceedings-2011 International Conference on Computational and Information Sciences, ICCIS*, pp. 449-451, 2011.

[31] W.Zhenyuan, & H. Haiyan., "OLAP Technology and its Business Application", *Second WRI Global Congress on Intelligent Systems,* pp. 92-95, 2010.

[32] Li, C., Y. Wang, & X. Guo., "The application research of OLAP in college decision support system", *International Conference on Multimedia and Information Technology*, MMIT 2010, pp. 74-77, 2010.

[33] Shimura, M., F. Monma, & S. Mitsuyoshi., "Descriptive analysis of emotion and feeling in voice", *6th International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2010*, pp. 4-7, 2010.

[34] Zhao, H., "Application of OLAP to the analysis of the curriculum chosen by students", *2nd International Conference on Anti-counterfeiting, Security, and Identification, ASID 2008*, pp. 97-100, 2008.

[35] Quafafou, M., S. Naouali, & G. Nachouki., "Knowledge Data Warehouse (DW)s: Web usage OLAP application", *IEEE/WIC/ACM International Conference on Web Intelligence, WI 2005*, pp. 334-337, 2005.

[36] He Y, Wang C, Jiang C., "Discovering canonical correlations between topical and topological information in document networks", *IEEE Transactions on Knowledge and Data Engineering*, vol. 1, no. 30(3), pp. 460-73, 2018.

[37] Kadar JA, Napitupulu D, Jati RK, "Analysis of factors influencing the quality of intranet website based on WebQual approach case study in agency X", *3rd International Conference on Science in Information Technology (ICSITech)*, pp. 526-532, IEEE, 2017.

[38] Qiao F, Chen K., "Correlation and Visualization Analysis of Large-Scale Dataset GDELT", *International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI)*, IEEE, pp. 68-72, 2016.

[39] Tan Z, Qin Z, "Empirical study of the correlation between real estate prices and bank loans based on VAR model", *Computer Science and Service System (CSSS), 2011 International Conference*, IEEE, pp. 2212-2217, 2011.

[40] Fu Q, Jia T, Cao L, Li S. Gray, "Correlation Model-Based Evaluation of Real Estate Projects", *International Conference on Management and Service Science (MASS)*, IEEE, pp. 1-4, 2010.

[41] Singh A, Sathyaraj R., "A Comparison Between Classification Algorithms on Different Datasets Methodologies using Rapidminer", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5 no. 5, 2016

[42] Kaur G, Chhabra A, "Improved J48 classification algorithm for the prediction of diabetes", *International Journal of Computer* Applications, pp. 98, 2014.

[43] Rahman RM, Afroz F., "Comparison of various classification techniques using different data mining tools for diabetes diagnosis." *Journal of Software Engineering and Applications*, vol. 6, no. 3, pp. 85, 2013.

[44] Vaithiyanathan V, Rajeswari K, Tajane K, Pitale R., "Comparison of different classification techniques using different datasets" *International Journal of Advances in Engineering & Technology*, vol. 1, no. 6(2), pp. 764, 2013.

[45] Weis M, Rumpf T, Gerhards R, Plümer L., "Comparison of different classification algorithms for weed detection from images based on shape parameters", *Bornimer Agrartechn. Ber*, vol. 69, pp. 53-64, 2009.

[46] Köknar-Tezel S, Latecki LJ, "Improving SVM classification on imbalanced data sets in distance spaces", *Ninth IEEE International Conference on Data Mining ICDM'09*, IEEE, pp. 259-267, 2009.

[47] Tang Y, Zhang YQ, Chawla NV, Krasser S., "SVMs modeling for highly imbalanced classification", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39 no. 1, pp. 281-288, 2009

[48] Davut Hanbay, "An expert system based on least square support vector machines for the diagnosis of heart disease", *Expert Systems with Applications: An International Journal*, vol. 36, no. 3, pp. 4232-4238, 2009

[49] Zhao CY, Zhang HX, Zhang XY, Liu MC, Hu ZD, Fan BT, "Application of support vector machine (SVM) for prediction toxic activity of different data sets", *Toxicology*, vol. 217, no. 2-3, pp. 105-19, 2006.

[50] Rokach L, Maimon O., "Decision trees", *Conference on Data mining and knowledge discovery,* pp. 165-192, 2005.

[51] Li K, Wang L, Wu J, Zhang Q, Liao G, Su L., "Using GA-SVM for defect inspection of flip chips based on vibration signals", *Microelectronics Reliability*, vol. 81, pp. 159-66, 2018.

[52] Yi JH, Deb S, Dong J, Alavi AH, Wang GG, "An improved NSGA-III algorithm with an adaptive mutation operator for Big Data optimization problems", *Future Generation Computer Systems*, 2018.

[53] Hu C, Dai L, Yan X, Gong W, Liu X, Wang L., "Modified NSGA-III for sensor placement in water distribution system", *Information Sciences*, 2018.

[54] Elarbi M, Bechikh S, Said LB., "On the Importance of Isolated Infeasible Solutions in the Many-objective Constrained NSGA-III", *Knowledge-Based Systems*, 2018.

[55] Chahardoli S, Hadian H, Vahedi R., "Optimization of hole height and wall thickness in perforated capped-end conical absorbers under axial quasi-static loading (using NSGA-III and MOEA/D algorithms)", *Thin-Walled Structures*, vol. 127, pp. 540-55, 2018.

[56] Tavana M, Li Z, Mobin M, Komaki M, Teymourian E., "Multi-objective control chart design optimization using NSGA-III and MOPSO enhanced with DEA and TOPSIS Expert Systems with Applications", vol. 50, pp. 17-39, 2016.

[57] Liu, Huan, and Hiroshi Motoda, "Feature selection for knowledge discovery and data mining", *Springer Science & Business Media*, vol. 454, 2012.

[58] Retrieved from Collector Rate in Punjab: http://punjabrevenue.nic.in/collectorrateframe.htm

[59] Retrieved from: https://www.kaggle.com/javico101/datamiami

[60] Retrieved from Justdial - Local Search, Order Food, Travel, Movies, Online Shopping: https://www.justdial.com/

LIST OF PUBLICATIONS

[1] Gursimran Kaur, Harkiran Kaur, (2017). Comprehensive Survey of OLAP Models. **ICHSA 2018** *(SCOPUS Indexed-SPRINGER)* [Published].

[2] Gursimran Kaur, Harkiran Kaur, (2018). Descriptive Data Analysis of Real Estate Using Cube Technology. **ERCICA 2018** *(SCOPUS Indexed-AISC Series-SPRINGER)* [Accepted].

[3] Gursimran Kaur, Harkiran Kaur, (2018). Factual Dimension Identification and Usage for Real Estate Framework. **ICMETE 2018** *(SCOPUS Indexed-IEEE Proceedings-IEEE)* [Accepted].

[4] Gursimran Kaur, Harkiran Kaur, (2018). Hybridizing J48 and SXM for Predicting Sale and Purchase of Property. **ICMETE 2018** *(SCOPUS Indexed-IEEE Proceedings-IEEE)* [Accepted].

[5] Gursimran Kaur, Harkiran Kaur, (2018). Non-Dominated Sorting Genetic Algorithm-III Based Meta Decision Tree. **ICMETE 2018** *(SCOPUS Indexed-IEEE Proceedings-IEEE)* [Accepted].

ORIGINALITY REPORT

8%	5%	7%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	blog.appliedinformaticsinc.com Internet Source	<1%
2	Mohammed Aashkaar, Purushottam Sharma, Naveen Garg. "Performance analysis using J48 decision tree for Indian corporate world", 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS), 2016 Publication	<1%
3	Maha Elarbi, Slim Bechikh, Lamjed Ben Said. "On the Importance of Isolated Infeasible Solutions in the Many-objective Constrained NSGA-III", Knowledge-Based Systems, 2018 Publication	<1%
4	Jiao-Hong Yi, Suash Deb, Junyu Dong, Amir H. Alavi, Gai-Ge Wang. "An improved NSGA-III algorithm with adaptive mutation operator for Big Data optimization problems", Future Generation Computer Systems, 2018 Publication	<1%