Sales Prediction Using Machine Learning Algorithms: The Case of Transsion (Tecno, Itel) Mobile Phone Manufacturing PLC.

A Thesis presented

By

Hailemicael Tenkir

The faculty of Informatics

Of

St. Mary's University

In partial fulfillment of the requirements
For the Degree of Master of Science

in

Computer science

July, 2024

Addis Ababa, Ethiopia

ACCEPTANCE

Sales Prediction Using Machine Learning Algorithms: The Case of Transsion (Tecno, Itel) Mobile Phone Manufacturing PLC.

By

Hailemicael Tenkir

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

Name	Signature	Date
Alemebante Mulu, PhD	_____	11/07/2024
Internal Examiner		
Minale Ashagrie, PhD	_____	11/07/2024
External Examiner		

_____

Dean, Faculty of Informatics

Name	Signature	Date
Alemebante Mulu (PhD)	_____	11/07/2024

July, 2024

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

_____

Hailemicael Tenkir

_____

Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Million Meshesha (PhD)

Million

Signature

Addis Ababa

Ethiopia

July, 2024

Contents

Acknowledgment

First of all, I would like to thank the almighty God and his Mother for giving me the strength, peace of my mind, and good health. Second, I would also like to express my deepest gratitude to my advisor Dr. Million Meshesha for his unreserved follow-up, invaluable comments, and constructive guidance throughout conducting this study.

Finally, I would also like to express the deepest gratitude to my family and friends who have been providing their advice and encouragement always including those hard times.

List of Abbreviations

AI	Artificial Intelligent
ERR	Error Rate
FNR	False Negative Rate
FPR	False Positive Rate
KNN	K-Nearest Neighbor
MAE	Mean Absolute Error
MES	Manufacturing Execution System
ML	Machine Learning
RMSE	Root Mean Square Error
SAP	System Analysis Program
SVM	Support Vector Machine

List of Figures

List of Tables

Abstract

The traditional approach of sales and marketing goals no longer help the companies to manage up with the pace of the competitive market, as they are carried out with no insights to customers' purchasing patterns. Major transformations can be seen in the domain of sales and marketing as a result of Machine Learning advancements. Due to such advancements, various critical aspects such as consumers' purchase patterns, target audience, and predicting sales for the recent years to come can be easily determined, thus helping the sales team in formulating plans for a boost in their business. The aim of this study is to utilize machine learning algorithms to develop a sales prediction model for Transsion Manufacturing PLC. In this study an attempt is made to apply machine learning algorithms for mobile phone sales prediction. After performing business and data understanding the data preparation task is done to clean and make the data ready for experimentation. For the experiment and construct predictive model, machine learning algorithms such as Random Forest, KNN, Naïve Bayes and SVM are selected based on their advantages and past performance seen in different literatures, it has been reported that they were widely used classifier algorithms for prediction and classification. The Jupyter Notebook with python programming is employed to simulate all the experiments. Confusion matrix is used to calculate the accuracy, precision and evaluate the performance of the models.

The results of the experiment show high accuracy, so that the models can be used to predict mobile phone sales either ITEL or TECNO Brand and either FEATURE phone or SMART phone accurately. Experimental results show that the Random Forest classifier outperforms other algorithms with an accuracy of 99.6%, 96.8% in experiment one and two respectively. Therefore, the Random Forest classifier is proposed for constructing mobile phone sales prediction models for Transsion Manufacturing. Based on the proposed optimal models in this study, we recommend future research to integrate mobile phone sales predictive models with mobile phone production systems.

Keywords: *Mobile Phone*, *Machine Learning*, *Transsion Manufacturing*, *Prediction Model*

CHAPTER ONE

INTRODUCTION

1.1 Background

Mobile phones, or cell phones, have grown more sophisticated and larger over time due to their structure as a cellular network. US researchers at Bell Laboratories in the 1970s started working on the cellular concept of a phone network. There are currently millions of phones in operation worldwide [1].

Transsion Manufacturing was established in 2006 as a global enterprise that focuses on providing the best mainstream mobile communication products and mobile internet services to local consumers in developing markets. The company's flagship mobile brands include Tecno, Itel and Infinix each of which has a significant presence in emerging economies. Headquartered in China, Transsion's market share in Sub-Saharan African countries exceeds 40% and has developed as a powerful competitor in the mobile industry after only ten years of growth. Overall over 246 million Dual-SIM devices have been sold. For developing markets that promote improved products, Transsion is invested in cutting-edge technology and provides locally customized merchandise by the wise expression "Think Globally, Act Locally" to have the strongest impact. Phantom, Camon, spark, povanewo, tablets, and accessories are the most significant [2].

Machine learning is one technology that emerged from AI. Machine learning is a technology that relies on the integration of computer science, statistics, and optimization to solve various problems such as regression, clustering, and classification. In its simplest form, machine learning is the ability of systems to learn from data while managing hidden and incomplete information and making decisions about new data guided by similar past computations. Machine learning, on the other hand, is learning on the part of a computer to learn and comprehend some unique parameters, on the other hand, data mining is the learning of rules from vast databases. In other words, machine learning employs trained systems to perform difficult operations and advance their skills utilizing accumulated data and experience, whereas data mining involves conducting research and examining the findings to forecast the outcomes based on the obtained data [3].

A challenge that confronted companies that invested heavily in collecting consumer information, other than big data technology, was that they had no idea how to leverage it. They were unable to leverage their vast customer and product feature databases to supplement a competitive edge [4]. Sales forecasting is an important part of corporate efficiency. The forecast can assist in anticipating market trends and where resources such as labor and money should be allocated. It is widely accepted that forecasting is a core component of marketing companies; hence, vendors must forecast authority. Through the use of this strategy, companies would be able to forecast future sales patterns and rely on data to drive strategy to enhance their sales tactics. It is both time-consuming and error-prone to manually conduct sales forecasting, causing future harm to the company's integrity, particularly in the constantly changing environment that we live in. Companies, ranging from industries are all critical to the worldwide economy because they produce products that sum up to satisfy requests on a broad-based level [5].

In business, sales forecasting is one of the vital business aspects since vendors have the responsibility of predicting the market while marketing to ensure the fleet of the market. Manual forecasting poses many hazards to business management today. The business sectors are crucial to the global economy, with varying units producing items and services as to the need in the market. In firm targeting one of the goals of individual and legal entities is to attract more market and forecasting tools come in handy. With more data sources, including consumer forms and industry marketing notes, firms have added more knowledge to financial units and forecast accurately. Utilizing forecasting methods simplifies the task of estimating product demand and sales figures within time frames. In this area machine learning shows potential as computers excel at following instructions and delivering outcomes contributing to the progress of modern society. Machine learning is rooted in principles allowing for the creation of models that closely align, with desired results. It has proven to be quite beneficial in sales forecasting enhancing the accuracy of sales projections. The goal is to anticipate product quantities and sales trends by identifying characteristics within the data. In order to grasp the information accurately and guide businesses in making choices during key moments, in their marketing strategies it is essential to conduct a comprehensive examination and review of the data. By leveraging data analysis and machine learning techniques sales forecasting, through machine learning predicts the sales outcomes of a product or service by analyzing data trends. The method involves the construction of predictive

models that yield reliable sales approximations with regard to factors such as brand product characteristics and customer type. For instance, machine learning algorithms such as regression analysis, decision trees, random forests, KNN, SVM, Naïve Bayes, and neural networks can help identify phenomena in a set of data and the relationship therein, providing an accurate speculation about future sales. The system, which has been trained with historical sales data, can adequately predict future sales trends. In sum, utilizing machine learning to predict sales helps businesses enhance their sales strategies as it informs teams about their clients' patterns and tendency, enabling them to make informed decisions regarding what to sell [6] [7].

Transsion Holdings Co., Ltd. aims to become the most popular mobile services and smart device supplier for consumers living in emerging areas around the globe. The company's premium multi-brand smart devices are its most well-known B 2 C goods. While mobile phones are the company's principal business, it also offers mobile Internet services built on an operating system that it created [2].

To enhance production and sales operations, currently uses SAP (System Applications and Products in Data Processing) and MES (Manufacturing Exclusion System). These systems produce essential statistics that Transsion Manufacturing PLC uses as its fundamental supply of facts. The employer wants to use this information to categorize, forecast, and examine approximately the behavior of its products. It is feasible to predict which cellular telephones are possible to sell properly by the use of predictive machine learning approaches, which include assessing cellphone brand, type, model, color, market type, and sales quantity. The aim of this research is to apply various machine learning techniques for sales forecasting in Transsion Manufacturing, thereby converting many sales datasets into beneficial fashions.

1.2 Statement of the Problem

In the past, companies would manufacture goods without considering sales numbers and market demand. To decide whether to scale up or scale down production, manufacturers now rely on data about the demand for their products in the market. Ignoring these crucial factors can lead to financial losses for companies as they strive to compete in the market. Each company adopts distinct criteria to assess their demand and sales figures. In today's highly competitive environment and ever-changing consumer landscape, accurate and timely forecasting of future revenue, also

known as revenue forecasting, or sales forecasting, can offer valuable insight to companies engaged in the manufacture, distribution or retail of goods. Short-term forecasts primarily help with production planning and stock management, while long-term forecasts can deal with business growth and decision-making. Sales forecasting is particularly important in the industries because of the limited shelf-life of many of the goods, which leads to a loss of income in both shortage and surplus situations. Too many orders lead to a shortage of products and still too few orders lead to a lack of opportunity. Therefore, competition in the market is continuously fluctuating due to factors such as pricing, advertisement, increasing demand from the customers [8].

Managers usually make sales predictions randomly. Professional managers, however, become hard to find and not always available (e.g., they can get sick or leave). Sales predictions can be assisted by computer systems that can play the qualified managers' role when they are not available or allow them to make the right decision by providing potential sales predictions. One way of implementing such a method is to try and model the professional managers' skills inside a computer program. Alternatively, the abundance of sales data and related information can be used through Machine Learning techniques to automatically develop more accurate sales predictive models. This approach is much simpler. It is not prejudiced by a single sales manager's particularities and is flexible, which means it can adapt to data changes. It has, however, the potential to overestimate the accuracy of the prediction of a human expert, which is normally incomplete. For example, once companies used to produce the products without taking into consideration the number of sales and demand as they faced several problems. Since they don't know how much to sell, for any manufacturer to decide whether to increase or decrease the number of units, data regarding the consumer demand for products is essential. If companies do not consider these principles when competing in the market, they will face losses. Different companies choose different parameters to determine their market and sales create a reliable prediction model that can project future sales for a certain time period utilizing past sales data and other pertinent information, while accounting for numerous factors that can Seasonality, marketing initiatives, the state of the economy, and rival actions all have an impact on sales. The model should offer useful insights and be assessed using the right measures [9].

Due to the incompatibility of the market demand with the sort of mobile phone production, Transsion Manufacturing encountered the previously mentioned problem when implementing

their sales strategy. For instance, the company manufactures mobile phones under the ITEL brand, but the market requires TECNO. In the same way, the company produces SMART phones, but FEATURE phones are what the market wants. This issue costs Transsion Manufacturing a lot of things, that is why this research was done to address this problem.

In an increasingly competitive market landscape, businesses are faced with the challenge of accurately forecasting sales to optimize resource allocation, inventory management, and marketing strategies. Traditional forecasting methods may fall short in capturing the complex relationships between various factors influencing sales. Therefore, there is a need to develop a machine learning algorithm-based sales prediction model that can leverage historical sales data, customer behavior patterns, market trends, and other relevant features to provide accurate and timely sales forecasts. The objective is to empower businesses with actionable insights to make informed decisions, drive revenue growth, and gain a competitive edge in the market [10].

The purpose of this research is therefore to create a sales prediction model for Transsion Manufacturing PLC by using machine learning strategies. The study intends to investigate and answer the following research questions.

1. Which attributes are more critical for Predicting Transsion Manufacturing mobile phone sales?
2. Which classification algorithm is suitable for constructing a model that predicts mobile phone sales?
3. What is the performance of the model in predicting mobile phone sales?

1.3 Objectives of the Research

1.3.1 General Objective

The fundamental objective of this research is creating a sales prediction model by using machine learning techniques.

1.3.2 Specific Objectives

Establishing the subsequent specific objectives is essential in order to achieve the overall objective of the study.

- To review related literature so as to identify suitable methods and algorithms for the study.
- To collect and prepare data sets for experimentation via establishing a robust pipeline for preparing records to manage outliers, missing values, and function engineering to enhance model training.
- To assess and identify suitable machine learning algorithms for sales forecasting.
- To create machine learning models that predict mobile BRAND.
- To create machine learning models that predict mobile PHONE TYPE.
- To optimize the hyper parameters of the model to beautify its expected accuracy and typical overall performance.
- To evaluate the performance of the model and decide its prediction accuracy by utilizing applicable measures.

1.4 Scope and Limitation of the Research

1.4.1 Research Scope

This research applies machine learning algorithms for sales prediction which can be particularly custom made to Transsion Manufacturing PLC's operations. In order to create a specific sales prediction model, we collect and analyze past sales data, customer conduct styles, market traits, and different relevant variables. The scope of the study involves assessing various machine learning methods, together with classification and class, to check the great approach for sales forecasting. In order to assist actual-time forecasting and decision-making, the research will include the sales prediction models into Transsion Manufacturing PLC's modern structures.

1.4.2 Research Limitations

The reliability and accuracy of the sales prediction models can be impacted by way of the availability and quality of historic sales information. Transsion Manufacturing PLC's information may be essential to the system studying algorithms' overall performance of the study depending on the information provided by Transsion Manufacturing. Uncertainties inside the sales prediction method may be introduced by outside variables such as economic status of the clients, the government economic system, competition in the market, and unexpected occasions. The research may be limited by the computational resources and expertise required to implement and optimize

complex machine learning algorithms for sales prediction. The scope of the study may not cover all possible variables and factors influencing sales, leading to potential limitations in the predictive capabilities of the models.

1.5 Significance of the Research

As an academic exercise, this study has a great contribution towards getting an understanding of the design and implementation of sales prediction by using ML technique. In addition, this study contributes its part for improving sales production, forecasting accuracy, and sales strategy optimization as listed below:

- **Better Forecasting:** Machine learning models can produce extra accurate and regular sales predictions, enabling companies to better expect call for and allocate sources as it should be.
- **Optimized Marketing Strategies:** Businesses can greater efficiently target the right customers with the proper items on the proper time via forecasting sales patterns.
- **Inventory control:** Businesses may also optimize their stock levels, limit surplus stock or out-of-stock and decorate typical supply chain performance via the usage of correct sales estimates.
- **Revenue Growth:** Businesses can pinpoint growth possibilities, high-quality-song pricing policies, and eventually enhance revenue producing by using sales prediction models.
- **Cost Savings:** By slicing waste, decreasing the value of inventory protecting, and enhancing aid allocation, correct sales forecasts can store fees.
- **Businesses possessing:** the capability to precisely forecast sales patterns experience a competitive gain inside the marketplace, which enables them to preserve a bonus over competitors and directly modify to evolving instances.
- **Consumer Insights:** By delivering useful insights into customer behavior and possibilities, sales prediction models assist companies enhance client happiness and customize their offerings.
- **Risk Management:** Accurate sales forecasting enables companies to plan investments, control financial dangers, and make assured strategic decisions.

- **Operational Efficiency:** Businesses may optimize resource allocation, streamline operations, and boom average enterprise overall performance with correct sales forecasts.
- **Data-Driven Decision Making:** Machine learning based sales forecast encourages an information-pushed approach to decision making, permitting businesses to base selections on quantitative analysis in preference to intestine feeling or speculation.

1.6 Research Methodology

Research methodology is the overall principle that leads the research. In order to conduct a good research, a precise approach and principle has to be followed.

1.6.1 Research Design

The studies layout of this research is experimental, with a focus on operation and managed testing to analyze causal linkages and identify institutions among special variables. Relevant literature sources, which includes books, journals, magazines, conference papers, manuals, and online sources which include Transsion manufacturing manuals are examined to meet the studies objectives and make sure an intensive expertise of the research issue and a successful investigation's end.

1.6.2 Data Preparation

The corporation of the records for machine learning to know programs is the main consciousness of this level that is every so often called data filtering. Machine learning of data filtering processes consist of Business Understanding (product definition, classification, and attribute choice), Data Understanding (records representation), Data Collection (information nice guarantee), and Data Preprocessing (data cleansing and dataset splitting).

1.6.3 Implementation tools

In order to generate dataset outputs, the usage of Python programming in Jupyter Notebooks, this look at required employing software development environments and doing experiments on a proposed approach. Python is the most extensively used language in facts science due to its considerable series of software program gear for computation, statistics manipulation, and visualization all of that are vital for studying experimental datasets. Python is desired because of

a variety of factors: The interface of Jupyter Notebook is simple to use, whilst Interface-Studio facilitates the display of datasets the usage of a number of figures and hyperlinks datasets in not unusual formats inclusive of, Technical computing advantages substantially from CSV's simplicity, ease of processing and storing big amounts of information, range of operators for array calculations, and rich device set for intermediate statistics evaluation. In this look at, MS Excel was applied for dataset coaching and Python turned into selected for guidelines mining.

1.6.4 Evaluation

Using a confusion matrix, the recommended studies was assessed by contrasting its findings with phenomena that have been manually located. Moreover, comparisons were made with different algorithms that have been typically used in in advance studies. The confusion matrix become used to evaluate the performance of the classification model on a selected dataset. It is a 2x2 table with four possible outcomes: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

1.7 Organization of this Thesis

There will be five chapters in the research thesis. A brief overview of the research will be given in Chapter One, which will include the Problem Statement, General and Specific Objectives, Research methods, Research Scope, Limitations and Significance. A review of related literature, an overview of Transsion Manufacturing, sales forecasting, and a brief overview of machine learning algorithms and techniques, model evaluation, reviews of related works, and gap analysis will all be covered in Chapter Two. The third chapter, which covers methodology, will address business understanding, data understanding, and data preprocessing. The overview of the experiment and discussion of the results, the proposed architecture, the experiment dataset, the modeling using KNN, Random Forest, Naïve Bayes, and SVM, the comparison of machine learning models, and a detailed discussion of the results are all covered in Chapter Four. The research's conclusion, recommendations, and suggestions for the future will all be included in the last chapter.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

Transsion Holdings Co., Ltd. Is dedicated to being a famous dealer of mobile offerings and clever gadgets to clients in emerging regions throughout the globe. The enterprise, which is famous for its wonderful selection of smart devices below multiple manufacturers, focuses mostly on mobile phones and offers mobile Internet services which might be based on a proprietary working system. Well-known mobile cell phone manufacturers in growing nations along with Tecno and Itel are included in Transsion's emblem portfolio, similarly to Syinix for domestic appliances, Carlcare for publish-sale services, and Oraimo for clever add-ons [2].

Transsion has turned out to be a main player in the cell sectors of global rising countries after years of boom. The commercial enterprise offered 156 million mobile phones globally in 2022, continuing to grow substantially in new markets and retaining a huge marketplace proportion in Africa. Notably, Transsion ranked 6^{th} in the international cell phone market with a 7.6% marketplace proportion and 3^{rd} within the international cell phone market with a 13.9% market percentage within the first half of 2023, according to IDC's Worldwide Quarterly Mobile Phone Tracker. Transsion positioned 6^{th} in India and primary in Africa, Pakistan, and Bangladesh in phrases of phone shipments [11].

Transsion's cellular smartphone brands Tecno, Itel and smart accent employer Oraimo are most prominent manufacturers in Africa in 2023, in line with a ranking by using African Business. The employer prioritizes technological innovation, and it has R and D facilities in Chongqing, Shanghai, and Shenzhen, China. Transsion constantly boosts research and improvement (R and D) spending to make its products more competitive and to widen its gain in localized generation innovation in developing international locations. Customers gain from this superior consumer revel in and value [2].

Transsion has advanced significantly in a number of technical fields in recent years, including visual perception, AI voice recognition, intelligent charging, photography, and more. The business has won multiple accolades in recognition of its innovations in AI speech technology and

photography. Embodying the brand essence of "Stop at Nothing" to enable people to pursue their best selves and futures, Tecno and Itel is an innovative technology brand that operates globally and offers a diverse range of products and services, including smartphones, smart wearable, laptops, tablets, the HiOS operating system, and smart home products [11].

2.2 Background of sales prediction

In order to help decision making enhance making plans, production, delivery, and advertising strategies, sales forecasting entails projecting future sales. Businesses use a whole lot of techniques to maintain income ranges over the route of many economic quarters. For instance, they will run income campaigns in which more than a few merchandises are furnished to traders at discounted expenses, increasing the amount of merchandise sold at some point of a given term. Precisely calculating a product's sales extent inside a given timeframe, region, and rate range is essential. In order to accurately forecast future sales, we can adopt a predictive look at Transsion production sales and the usage of massive datasets in this research [12].

In the past, forecasting sales has been seen as a time series hassle that requires statistical modeling and evaluation to make destiny predictions. Alternatively, it is able to be considered as a regression problem, the usage of device machine learning techniques to find hidden styles and traits in beyond sales information and project destiny sales, whether or not they be quick-term or long-term. The accuracy and dependability of the predictive model may be progressed by using variables consisting of competitors, tendencies, demographics, and marketing campaigns [13].

For sales forecasting, several machine learning techniques can be applied. Supervised machine learning techniques inclusive of Random Forest, Support Vector Machines (SVM), and neural networks can be used when a good enough amount of statistics is to be had. On the other hand, supervised machine learning techniques of strategies like K-nearest neighbor can be useful for forecasting the sales of latest products. Achieving high prediction accuracy in actual-global settings calls for improving fashions the usage of methods like hyper parameter tuning; but, it's also important to ensure model generalization for actual-world records through the usage of strategies like model stacking and ensembles of fashions [14].

Sales forecasting is extra than simply making predictions; it is also about preparing for unknowns. Sales estimates can grow to be unsure due to numerous elements together with weather fluctuations, rival hobby, and promotions. Therefore, calculating ability uncertainties is a crucial part of the forecasting process. Furthermore, depending only on one model technology may result in deceptive conclusions down the street because predictor distributions and patterns in time series facts are dynamic in nature. As a result, it's essential to constantly develop new models using contemporary information, and automatic model choice may be pretty useful in keeping relevance and accuracy over the years [9].

In précis, using machine learning algorithms to historical time series information may help corporations complete an essential mission: sales forecasting. This task emphasizes model optimization, generalization, and accounting for uncertainties and precise residences of time series records, with a focus on the usage of a number of machine learning strategies to gain top-quality prediction accuracy. Business analytics is now an essential issue of all enterprise aid structures thanks to trends in data engineering (DE) and analytics. Accurate sales projections are necessary for firms to correctly manage their operations and sales strategies, and demand and sales forecasting are key components of enterprise analytics systems [15].

2.3 Introducing Machine Learning

Today's companies and enterprises are able to file their daily sports considering laptop technology is extra available and inexpensive. As a result of the development of present-day generation and the substantial use of databases, corporations and agencies gather a massive amount of facts in numerous codecs. Finding beneficial facts or insights in this fact to assist with decision-making is tough [16].Even though records are growing at a quick pace, a massive quantity of statistics is stored in the garage without any in addition use.

When well analyzed, data gathered from many sources might reveal formerly undiscovered statistics that spur development and help with decision making that advances the company. However, as Zen Tut [17] mentioned, an enterprise's potential to amplify may be hampered through the collection of needless records.

These factors make machine learning crucial for enhancing decision making in corporate settings. Machine learning is the technique of extracting understanding from big datasets, which can be any kind of records (internet, multimedia, textual content, etc.) [17]. It includes a number of methods for the use of the facts kept in statistics warehouses to discover new or surprising patterns in facts.

One crucial use of machine learning strategies which could help production companies is the insights that sales managers can use to hold tremendous products, entice new customers, and boom profits from current ones.

The multidisciplinary region of machine learning integrates quantitative statistics and artificial intelligence to derive models from datasets that are more complex than people who can be processed by using the SQL (Structured Query Language) language [18]. Furthermore, device learning requires the following steps: Preprocessing -> Model -> Validation.

Since actual datasets are frequently noisy, partial, unclean, and available in a variety of codecs, preprocessing is a vital step within the system getting to know the procedure. The strategies and algorithms that Machine Learning uses to extract insights from the records are called the Model. It gives a thorough rundown of all the different machine learning strategies, together with the A-priori algorithm and K-means clustering [19].

Verifying the output model produced by machine learning algorithms is validation, the closing section of the machine getting to know the system. Not every model generated with the aid of gadget mastering algorithms is considered valid. The Machine Learning algorithms have to be evaluated on a check set of records in order to overcome this obstacle. If the output model of the machine learning approach yields the expected results, it can be used to extract understanding from a larger dataset. Nevertheless, the pre-processing and Machine Learning set of rules steps must be reevaluated if the output model does no longer yield the expected outcomes [18].

2.3.1 Types of Machine Learning

In the language of the subject, machine learning is the ability of machines to study on their own without the want for express programming. It includes a pc program that learns from revel in E with respect to a certain set of responsibilities T and performance measure P. As a result of enjoy

E, the pc application turns into more talented at tasks interior T as decided with the aid of P. In popular, device mastering refers to an application which can examine and discover information to address plenty of responsibilities. Tasks like e-mail junk mail identity, sales forecasting, credit score card fraud detection, inventory marketplace predictions, digital assistants, personalized product hints, self-reliant cars, sentiment analysis, and more are examples of not unusual uses for machine learning [20].

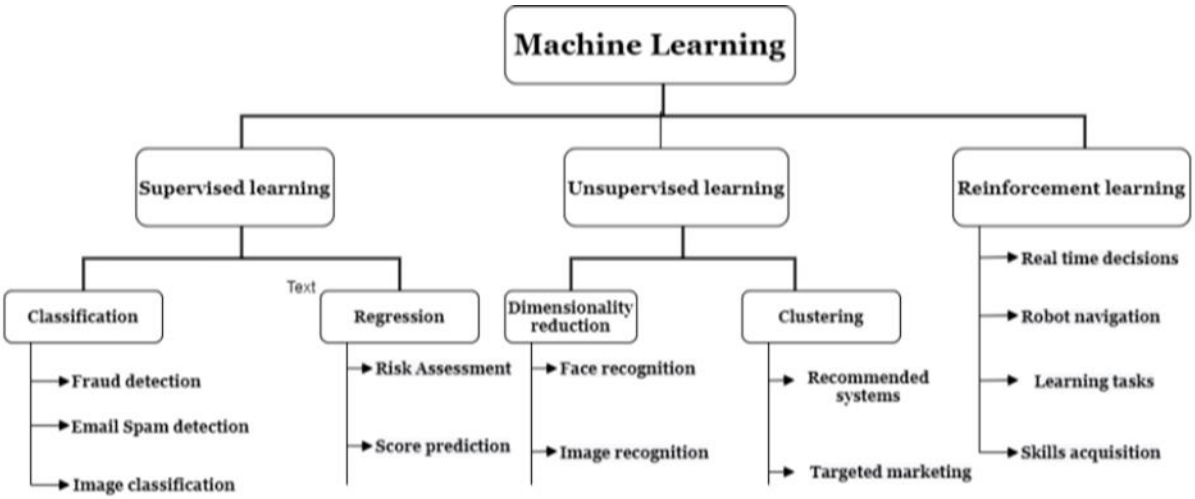


Figure 2.1: Types of Machine Learning [20]

As shown in figure 2.1 above there are three types of machine learning such as supervised learning, unsupervised learning and reinforcement Learning.

Supervised learning is the primary approach utilized in device learning tactics. It works best with datasets which have labeled dataset among the input and output information. This branch of device getting to know is worried with the use of classified test statistics to analyze a type or classification model [20].

In the case of unsupervised studying, this is, when there's virtually unlabeled facts, the records aren’t always explicitly labeled into separate training. From the facts, the Model can examine by locating implicit styles. Based on records, unsupervised mastering classifies densities, structures, related segments, and different similar attributes [20].

Reinforcement Learning is a subfield of machine learning that focuses on selecting behaviors so that it will maximize rewards in a certain scenario. To determine the first-rate route of movement or choice to make in a given state of affairs, several computational techniques and algorithms are implemented. Reinforcement mastering entails an agent that decides how to carry out a task without specific responses, in evaluation to supervised mastering, in which the education information provides the proper answers for the model to learn from. Instead of depending simply on training information, this method requires machine learning to enjoy [20].

2.3.2 Machine Learning Algorithm

For both supervised and unsupervised gaining knowledge of, a huge range of machine learning algorithmic techniques can be used to categorize a trouble in step with a collection of attributes. In order to discover supervised algorithms that paint properly for sales forecasting and market analysis, this takes a look at investigating them. This observation is seen numerous machines learning algorithms, including Random Forest, Naïve Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor (K-NN).

2.3.2.1 Support Vector Machine (SVM)

Supervised studying fashions known as support vector machines are used for issues associated with class, prediction, and clustering. An SVM builds a model that may classify clean observations into distinct businesses through reading a collection of entered observations and matching binary outputs. Support vector machines are desired by means of many due to the fact they will produce outcomes with amazing precision and little computing overhead [21].

A setting apart from the hyperplane defines a Support Vector Machine (SVM), a discriminative classifier. Essentially, the technique finds the high-quality hyperplane to categorize new times whilst given labeled training records (supervised machine learning). This hyperplane functions as a line that splits the aircraft into two portions in dimensional space, with each magnificence positioned on every side [22].

Finding a hyperplane that correctly divides the information factors into N-dimensional area in which N is the variety of functions is the main intention of the help vector device set of rules.

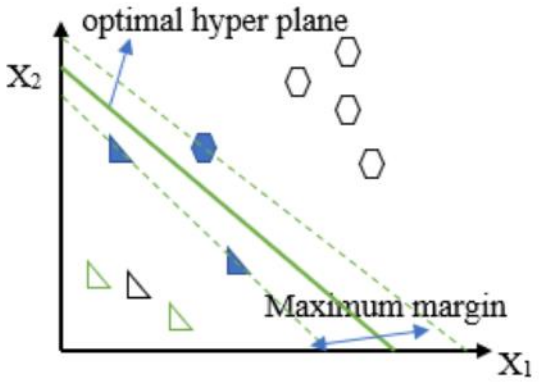


Figure 1.2: A linear line separating the data types [21]

When it comes to deciding on a hyperplane to divide classes of data points apart, there are loads of alternatives available. Finding a plane with the maximum margin, a plane that represents the maximum distance among information factors from every class is the valuable concept of support vector machines (SVMs). The SVM becomes stronger in its capability to optimistically categorize upcoming data points by optimizing this margin distance.

In actual-global applications, achieving a perfect category often comes with a fee, mainly when noise or different variables prevent the two lessons from being linearly separable. In those conditions, the need for the best hyperplane can be loosened by means of adding a massive training dataset as a new term.

It is often impossible to achieve a properly divided line in the facts area in real-global programs. As an end result, a curved selection boundary may be required. Although it's far more viable to create a hyperplane that divides the records, this will no longer be the first-rate option if noise is a gift in the records. In these varieties of conditions, the soft margin approach is used. With this device, it is viable for points to slip out of doors of the margin, with corresponding consequences increasing as the factors deviate from the margin. The hyperplane separation is used to maximize the space among the margin and the ultimate examples whilst minimizing the penalty of misclassified points.

Another technique, referred to as the SVM kernel, is utilized to segregate statistics that are not linearly separable by means of mapping the data into a better-dimensional area. Through this mapping, including x = (x, x2), the records are converted right into a two-dimensional space, in

which a discernible linearly separable line emerges while graphed. The preference of mapping to elevate the trouble's dimensionality is contingent upon the particular facts space underneath examination. The computations worried in figuring out the maximum-margin separator can be expressed in terms of scalar products among pairs of statistics factors inside the excessive-dimensional feature area, with these scalar products serving as the sole aspect of the computation motivated by way of the dimensionality of the excessive dimensional area.

2.3.2.2 Naive Bayes

One popular categorization technique based on Bayes principle is Naive Bayes. A check statistics point's posterior class chance may be ascertained with the aid of using elegance conditional density estimation and class previous possibility, which allows it to be assigned to the class with the highest posterior magnificence probability [23]. The Naive Bayes algorithm that is seen as one of the maximum efficient techniques is a simple yet incredibly powerful device for predictive modeling [24].

Because the technique is simple to enforce in code and permits for the quick advent of predictive fashions, it has been widely adopted. As a result, real-time model forecasts are beneficial. The Bayes theorem, that is frequently used to compute posterior possibilities primarily based on observations, in Naive Bayes opportunity idea, connects the conditional and marginal probabilities of random events [23].

Consider x = (x1, x2..., xd) as a d-dimensional example without a class label. The intention is to build a classifier of the usage of the Bayes theorem to be expecting its unknown elegance label. The collection of class labels is denoted by way of C = C1, C2..., CK, and the previous possibility of Ck (okay = 1, 2..., K) is represented by P(Ck), that's determined before additional evidence is acquired) represents the conditional possibility of witnessing the evidence x within the event that the hypothesis Ck is accurate. The following method illustrates how the Bayes theorem is implemented within the procedure of making such classifiers: -

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{k'} P(x|C_{k'})P(C_{k'})} \quad \text{……..2.1}$$

The value of a particular feature inside a class is thought to be independent of the value of every other feature in a naive Bayes classifier [23].

$$P(x|C_k) = \prod_{j=1}^{d} P(x^j|C_k) \qquad \text{…….2.2}$$

2.3.2.3 K-Nearest Neighbor (KKN)

Nearest neighbor (KNN) classifiers are extensively used for the reason that they're recognized to be easy but powerful type strategies. The k-NN technique mainly is one such instance of this. The KNN algorithm's primary goal is to be expecting a new pattern point's categorization by use of a database containing facts points labeled into exceptional groups. An object's type is decided by means of looking at times which can be closest to it. The wide variety of friends taken into consideration at some stage in the voting manner is indicated with the support of the parameter K in KNN algorithms. Choosing a suitable fee for K is essential as it impacts the category accuracy. An item is given the equal magnificence as its closest neighbor while K=1. The similarity of all K instances furnished is used to classify facts as K rises [25].

Similarity may be measured by the use of numerous distance metrics or similarity functions, together with Euclidean, Cosine, and Jaccard. The holdout method is used to evaluate the accuracy and performance of the classifier as soon as it has been constructed using the training data documents. The authentic training dataset is split into halves the usage of the holdout method: the majority is used to generate classifiers, while the minority is utilized to evaluate classifier precision. Because of their relatedness, it is anticipated that papers within the same elegance could have nearby friends, signifying a smaller distance between them [25].

Two of the principle problems in the usage of the k-NN selection rule are the computational complexity that results from many distance computations and the critical undertaking of choosing a suitable value for K. A low K fee can also motive motivate over-fitting because it amplifies the effect of noise on the outcomes. On the other hand, an excessive K number increases computing requirements and could probably compromise the core concept of KNN, that's that neighboring points most probably belong to the identical magnificence. Setting K equal to the rectangular root of the entire number of occurrences k = n^ (1/2) is a trustworthy choice guiding principle [26].

Preparing the database so that observations can be compared is the first level within the trendy workflow of KNN algorithms. The observations are then considered as factors in area, and the diploma of similarity between them is measured with the aid of calculating the space among them using a suitable metric (the Euclidean distance is a frequently used measure). The following method is used to determine the Euclidean distance between instances (X1, X2, X3..., Xn) and (U1, U2, U3..., Un): -

$$\sqrt{(X_1 - U_1) + (X_2 - U_2) + \cdots + (X_n - U_n)} \quad \ldots\ldots\ldots\ldots \boldsymbol{2.3}$$

2.3.2.4 Random Forest

An improved model of the Bootstrap Aggregated (Bagged) Trees approach, which comes from the traditional Classification and Regression Tree (CART) algorithm first provided via Breiman et al. In 1984 the Random Forest algorithm. A dataset is regularly divided into maximize the homogeneity or "purity" of each resulting "leaf" Portions in Regression Tree if you want to maximize the homogeneity or maximize the homogeneity or "purity" of each resulting "leaf" of each subsequent set. Using a greedy method, this set of rules chooses the function at every node that describes the variation in the education set. Based on the fee or common of the reaction variable within the matching leaf that corresponds with the characteristics of the brand-new times, predictions are made for brand spanking new records instances. Because of their feature-structured splits, tree predictors are exquisite at shooting interplay results between capabilities, but they'll have problem capturing relationships that might be higher represented by way of non-stop features like logarithmic, linear, exponential, or quadratic. By collecting noise within the training records, unrestricted tree development would possibly cause over fitting, necessitating the use of stopping criteria or manual pruning to improve robustness [27].

Breiman (1996) advised an ensemble strategy to overcome those drawbacks, where several trees are grown on numerous subsamples, sampled separately, and all bushes in the "forest" have the equal distribution. Next, the usage of a way known as Bootstrap Aggregating, also referred to as Bagging, predictions crafted from fresh data are averaged over all trees. According to Breiman (2001), the power and correlation of person trees decide the generalization mistakes of a wooded area of tree regressors. Random Forests were created because of enhancing the bagged tree algorithm with the aid of including randomization to the tree-construction system as a way to

maximize the Bias/Variance stability. The robustness and predictive effectiveness of the technique are multiplied for the reason that every node in the manner can most effectively break up across a random subset of attributes, generating person trees which can be weaker however much less correlated [26].

Random Forest is a strong machine learning framework for predictive analytics that makes use of a collection of simple fashions, all of which can be decision trees. This ensemble technique, also called model meeting, combines multiple fashions to improve prediction performance. In a random wooded area, every base version is built personally with a wonderful set of functions [28].

In the initial setup, the model quantity is represented via the configuration. In this situation, every base classifier is a truthful selection tree. The procedure of combining many fashions to boom forecast accuracy is referred to as model assembling. In a random forest, every base model is constructed independently using a distinct subset of statistics [28].

2.4 Model Evaluation

A famous technique for assessing overall performance in the domain names of facts, system studying, facts mining, and synthetic intelligence is the confusion matrix. By giving a clear split of actual versus expected classifications, it is regularly used to evaluate the performance of binary class problems. The confusion matrix became selected as the assessment technique on this examination as it affords insightful data approximately how well the recommended model plays in classification and prediction responsibilities [29].

2.4.1 Confusion Matrix

The confusion matrix is used to evaluate how properly a binary class problem performs across the dataset. While FP (False Positive) and FN (False Negative) pick out instances wrongly, the diagonal factors, together with TP (True Positive) and TN (True Negative), accurately classify instances. The wide variety of instances which are incorrectly classified as terrible is known as the TN. FP indicates the range of cases that have been incorrectly categorized as high quality. FN stands for the number of instances that have been accurately categorized as bad. TP stands for the full wide variety of instances that are accurately categorized as high-quality [29].

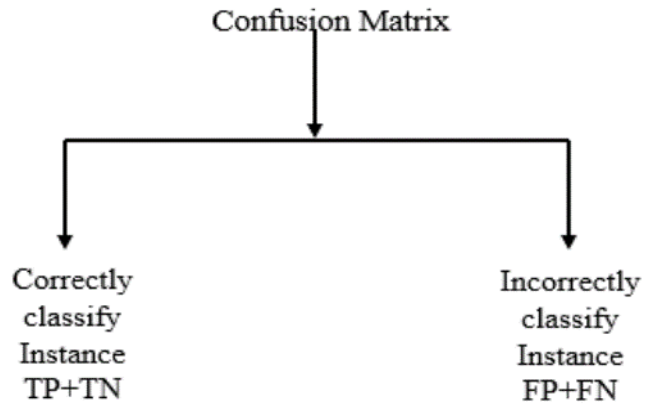


Figure 2.3: Example of Confusion Matrix [29]

The general range of occurrences is made from the wide variety of effectively and wrongly classified occurrences. Instances which are effectively categorized are computed as TP TN. FP FN determines instances which have been incorrectly categorized. The techniques indexed below can be used to decide the values of the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) [29].

Table 2.1: Confusion Matrix Example [29]

	Predicted	
Actual	Yes	No
Yes	TP	FN
No	FP	TN

2.4.1.1 Accuracy

The overall range of correct predictions (TP and TN) divided via the entire range of times in the dataset yields the accuracy. The exceptional viable accuracy rating is 1.0, which denotes perfect predictions, and the lowest price is zero. Which denotes the worst viable overall performance [30].

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{……..2.4}$$

2.4.1.2 Precision

The range of correct positive predictions divided by using the overall variety of nice predictions (TP and FP) yields the high-quality precision 1.0 is the best precision score, and zero is the bottom [30].

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{……2.5}$$

2.4.1.3 False Positive Rate (FPR)

This fee is decided by dividing the full range of negatives (TN and FP) by the range of wrong bad predictions [29].

$$FPR = \frac{FP}{TN+FP} \quad \text{……2.6}$$

2.4.1.4 False Negative Rate (FNR)

By dividing the total range of positives (FN and TP) by means of the number of incorrect tremendous predictions, you could get the fake-terrible charge. 1.0 is the precise fake-terrible fee, and 0.0 is the worst rate [29].

$$FNR = \frac{FN}{FN+TP} \quad \text{……..2.7}$$

2.4.1.5 Error Rate (ERR)

The overall quantity of faulty predictions (FP and FN) divided via the overall quantity of instances inside the dataset yields the error price 0.1 represents the premiere error charge, whereas 1.0 is the worst price [29].

$$ERR = \frac{FP+FN}{TN+FP+FN+TP} \quad \text{……2.8}$$

2.4.1.6 Sensitivity

The quantity of True positives (TP) divided by the whole quantity of positives within the dataset (TP FN) yields sensitivity, also referred to as True Positive Rate (TPR) [30].

$$sensitivity = \frac{TP}{TP + FN} \quad \text{.....2.9}$$

2.4.1.6 Specificity

By dividing the whole quantity of negatives TN and FP covered by the wide variety of proper negative predictions, specificity additionally called True Negative Rate, or TNR is calculated [30].

$$specificity = \frac{TN}{TN + FP} \quad \text{.....2.10}$$

2.4.1.7 Mean Absolute Error (MAE)

The suggest absolute errors, or the absolute amount of the prediction error of all test sales prediction errors being the distinction between the actual and anticipated values defines the average magnitude of errors in a hard and fast of predictions without taking instructions under consideration. When comparing fashions, this method is usually carried out to regression fashions. The accuracy of the model improves with a lower MAE cost [31].

2.4.1.8 Root Mean Square Error (RMSE)

One common place approach for calculating the inaccuracy of prediction facts in a model is to apply the root mean square. It is frequently referred to as the separation among the observed and anticipated values' vectors. This makes it possible to calculate the error popular deviation for an unmarried statement rather than the complete set of observations. A model's accuracy can be improved by means of adjusting its capabilities or adjusting its hyperparameters if the model predicts a better RMSE fee than 180, that's the best number for RMSE [31].

2.4.1.9 R-Squared (R^2)

R-Squared (R^2) a statistical metric referred to as R-squared is used to decide the number of statistical factors that are closer to the suit regression line. For multiple regression, it is also known as the coefficient of willpower or the coefficient of a couple of dedication. In essence, it expresses the percentage of response variable variants that a linear version explains. R-squared is made from the explained and general variant. R-squared lies among 0% and a 100% of variability, in which 0% suggests that response variable variability around imply is absent and a 100% shows that response variable variability round mean within the version is comprehensive. Therefore, a version that suits better has a better R^2 price [31].

2.4.1.10 Adjusted R-squared(R^2)

This is a changed shape of R^2 (R-squared), which estimates goal variable variance with the support of omitting all besides the maximum vast factors that are a useful resource in data prediction. It is hired due to the fact the R-Squared technique has limitations, inclusive of the truth that once extra variables are introduced to the model without first expertise how the version behaves, the price of r-square will increase and might have either an amazing or negative impact. Therefore, together with more traits within the model to predict the target variable might be penalized via adjusted R^2. Adjusted R-squared would upward push within the event that R^2 considerably expanded, and would fall inside the event that R^2 substantially reduced [31].

2.5 Related Works

Companies and businesses are constantly searching for better machine learning models and techniques that are vital for retaining their competitiveness and generating extra profits [32]. An incredible deal of examination has been executed within the region of product sales forecasting using information mining and machine learning techniques. This segment gives a concise overview of diverse associated research which has explored sales forecasting and related predictive challenges. Numerous statistical fashions and techniques, which includes regression, Auto-Regressive Integrated Moving Average (ARIMA), and Auto-Regressive Moving Average (ARMA), were utilized to expand various sales forecasting frameworks. However, forecasting

poses a complicated hassle stimulated by way of each internal and outside element, with limitations inherent within the statistical method.

Arif. et al. [33] offer a completely unique strategy: the use of machine learning to enhance prediction accuracy in an assessment evaluation. To ascertain the maximum green approach for sales forecasting, their technique entails collecting and evaluating ancient information from stores using diverse algorithms, which includes K-Nearest Neighbor (KNN), Gaussian Naïve Bayes, and Decision Tree Classifier. Based on their ancient information, they evaluated diverse algorithms and found that Gaussian Naive Bayes had first-class accuracy. Several machine learning algorithms and information mining strategies had been examined to determine the first-class technique for accurate sales forecasting. They emphasized how threatening it is for traditional forecasting algorithms to manipulate big datasets and guarantee forecast accuracy. They did, however, suggest that a variety of records mining techniques may be applied to overcome those limitations. Gather information from a superstore as well as a variety of 10 distinct products, then transform the raw information into processed information. 80 percent of the data is used to train the machine, with the remaining 20 percent being used for testing. The accuracy of the KNN, Gaussian Naïve Bayes, and Decision Tree classifiers are 35.71%, 58.92%, and 28.57%, in that order. Among other techniques, the Gaussian Naïve Bayes classifier has the highest accuracy at 58.92%. The model generates a preliminary demand for a specific good. This study did not consider consumer behavior, seasonal weather, time, occasion, month, or product category.

In work done by S. Cheriyan et al. [32], they implemented various machine learning algorithms and data mining techniques with the perception of choosing the best approach to forecast sales with a high degree of precision. However, they pointed out that traditional forecast systems are complicated when dealing with big data and the accuracy of sales forecasting. However, they clarified that these problems might be resolved by applying different data mining strategies. The results indicate that the Gradient Boost Algorithm performs at 98% overall accuracy, Decision Tree Algorithms comes in second with about 71% overall accuracy, and the Generalized Linear Model comes in third with 64% accuracy. Lastly, a comparison of the three selected algorithms' empirical evaluations reveals that Gradient Boosted Tree is the method that fits the model the best. Fields and qualities utilized in this research, however, were insufficient for additional investigation.

Comparably, Chand N et al. [34] give a unique setup methodology that emphasizes an algorithm that ensures steady precision while reducing variance from actual sales through the usage of an averaging approach. According to their studies, combining regression and time-series fashions reduces the risk of over or beneath-forecasting, intently aligns forecast values with real facts, and favors fashions which have done properly within the beyond. creating the ratios 70:30. 30 percent of the dataset is used for testing, while the remaining 70 percent is used for training. The accuracy of the time-series model is 63%, and the regression-based model is 60%. The sales-in data for a specific market was employed in this study's construction of the model, and it was then tested using various datasets.

Furthermore, to improve demand forecasting a crucial factor of supply chain management Kilimci Z H et al. [35] cautioned combining nine time series techniques, the Support Vector Regression (SVR) set of rules, and Deep Learning forecasting fashions. They highlighted a boosting strategy for more suitable demand forecasting model performance and used a singular integration strategy associated with boosting ensemble strategies to mix algorithms for finest decision making. With this approach, extra accurate projections that considered modifications in trend and seasonality had been assured. They used random forest, clustering algorithms, and neural networks in conjunction with a sentiment analysis approach to estimate product demand for a sure term for you to boom the profitability of businesses done extra accurate product call for forecasting. The raw data was gathered from Turkey's SOK market, a rapidly expanding business with 6700 outlets, 1500 goods, and 23 distribution centers. Compare several methods for the demand forecasting process, such as the statistical model, the winter model, and the radius basis function neural network (RBFNN) with SVM. They end the analysis by noting that, at the average MAPE results level, the SVM method outperforms the others with an augmentation of about 7.7%. The study did not consider sources such as economic studies, consumer trends, social media, social gatherings, and store demographic data based on geography.

Researchers from Dokuz Eylul University assessed the region-space fashions' and the ANFIS model's predicting accuracy in an extraordinary examination carried out in 2017 [36], adding significant new records to the body of current expertise. Five distinctive methodologies had been applied in their research to forecast the sales extent of a retail furnishings store: nation-space fashions, ARIMA fashions, ARFIMA fashions, ANN models, and ANFIS models. They sought to

grow the effectiveness of the supply chain gadget via helping departmental personnel in determining which forecasting method was great for diverse product sales through contrasting the performance of every forecasting technique. The observer's end, which highlights the fee of mixed strategies over standalone models, is that combining more than one forecasting methodologies can result in considerable gains in forecast accuracy. This study's application of sales forecasting is based on weekly sales data from a multinational furniture manufacturer that has a long history of operation in Turkey's retail market. This study examines sales information for ten products chosen from various product categories. In this study, the forecast accuracy of the ten-time series is assessed using a range of various forecast evaluation statistics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (sMAPE). The models with the lowest error value on the validation set were chosen as the final models for the ANN, ANFIS, ARIMA, and ARFIMA models as well as the ETS, ARIMA, and ARFIMA models. As a result, it makes sense that the combined approaches' overall results would be better than the solo models. The study omits any macroeconomic variables that provide information on the purchasing power and economic conditions of the nation in which it is conducted, as well as any indicators that provide a summary of the circumstances in the relevant industry.

The usefulness of Bayesian classifiers in sales forecasting was investigated in Gallagher et al.2015 take a look at [37].Specifically, the observer looked at how nicely they diagnosed agreements that were misclassified through the Qualitative Sales Predictor (QSP) and might probably win or lose. They used a greater sophisticated set of rules than QSP, combining qualitative and quantitative sales variables to enhance overall performance. According to their findings, throughout the checking out and validation tiers, Bayesian classifiers in particular, the TAN classifier displayed the highest prediction accuracy. The 90.6% accuracy rate for the TAN approach on the cease of the examiner tested a remarkable improvement above the 75.6% accuracy acquired via QSP on the same dataset.

In their work, Khan M.A et al. [38] built a demand forecasting model structured on business intelligence and machine learning. Thus, by analyzing past and current market data, businesses and organizations can effectively predict the future demand of goods and manufacture those goods that appear to have more demand in the future. The sales and inventory figures are derived on

actual data from prior years. For the forecasts, data from 2014 onwards is considered. All types of data are thrown into the shared database. Before data is processed and taken into consideration for the following step, purification is done. cleaning and preparing data for entry into machine learning (ML). When their model was used in stores, it produced an accuracy of 92.38%. This study is unable to consider the stock/product optimization stage.

2.6 Gaps Analysis

There are already a huge variety of producing facilities running everywhere in the global, every with their very own wonderful distribution techniques and product designs. Manufacturing company has specific obstacles and functions in its products while assessing which brands are most applicable for precise client groups. Company needs to adapt its sales techniques to other nations due to the fact extraordinary markets have distinct desires in terms of budget, training, and services.

While research on sales prediction in Ethiopia remains scarce, existing studies predominantly focus on foreign manufacturing companies and various industries. Directly applying the findings from these studies to local context is not feasible.

The purpose of the observing is to study an extensive model of literature on sales prediction from exceptional agencies. Various techniques had been used in previous research, along with distinct techniques, distinct datasets and preprocessing techniques, and unique residences linked to product and market classes. This variant emerges from the distinct sales problems that production organizations stumble upon globally, that are impacted through their distinct product offerings, sales techniques, and organizational setups.

The majority of previous studies concentrated on sales modelling and employed training data directly, without considering the relationship between the training and testing data. This results in numerous errors, which lower accuracy. In order to reduce computational time and obtain effective evaluation performance, recent studies have proposed clustering strategies to partition the full forecasting data into many clusters of predictable data prior to creating predictable models.

Previously a lot of sales and demand forecasting work was performed using Machine Learning. Most of the work will concentrate on the sales of food items. However, the frequency of food sales and mobile phone sales are totally different. The frequency of food sales very high because of daily consumption. This study encompasses various attributes such as phone brand, model, color, market type, phone type, and phone quantity. Furthermore, it utilizes distinct datasets that have not been previously employed in related studies, featuring balanced and sizable samples that adequately represent all time frames. Overall, this research incorporates larger sample size datasets and diverse attributes compared to the researchers mentioned in this study. It introduces a novel dataset from Transsion Manufacturing that has not been utilized in previous research. The resulting model will be directly applicable and beneficial for Transsion Manufacturing.

CHAPTER THREE

Methodology

3.1 Overview

A major factor in determining the outcomes of machine learning is data preparation. How exactly and extensively the required data is gathered, examined, and preprocessed will largely determine the model that is constructed. The original data description is followed by the following sections, which describe the business understanding's selection of the relevant attribute, the data understanding's construction of the task for the relevant attribute used in sales prediction, and the preprocessing tasks carried out to clean the dataset's attributes.

3.2 Research design

Experimental research is the method used in this study. Since the main method for examining causal (cause/effect) linkages and examining the relationship between one variable and another is experimentation, experimental research designs are chosen. Researchers compare two or more groups on one or more measures using experimental research [39].Establishing a connection between two variables, the dependent and independent variables is the aim of experimental research. After an experimental research study is finished, a correlation between a certain property of an entity and the variable under investigation is either confirmed or refuted. Statically a machine learning model is used in the study to do a large-scale experiment. This particular process model was chosen because it offers a more comprehensive and research-focused explanation of the processes, substituting the modeling step with a data processing step and adding multiple new explicit feedback mechanisms. The model consists of six steps: comprehending the business domain, comprehending the data, preparing the data, processing the data, assessing the result that was found, and using the result that was found. Thus, the ultimate goal of the research was to use the Transsion manufacturing dataset to create a model that could be used to forecast mobile phone sales. The dynamic and iterative features of this approach are among its most significant features. Because decisions and modifications made in one stage might affect subsequent steps, feedback loops are essential. Below, the researcher attempts to go into detail about the tasks completed and the methods utilized at each step.

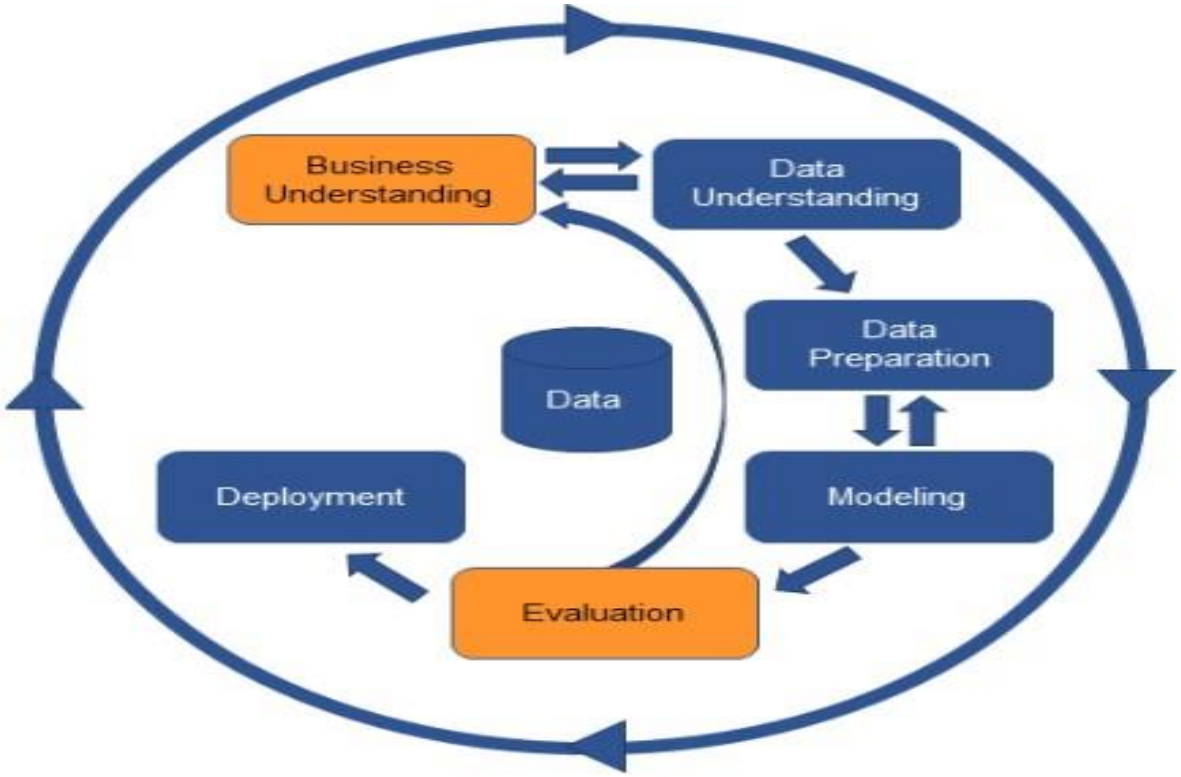


Figure 3.1: Stage of Machine Learning [32]

3.2.1 Business Understanding

Understanding the issue at hand is the first step in any practical investigation related to the business. A hawk's eye view of the possible factors that can be predicted is provided by business understanding. The manufacturing center business that was selected focuses on workers, customers, products, and inventory. With the customer at the center of the business, this study focuses on the daily sales of mobile phones. Following multiple phone conversations and meetings with the business's stakeholders, a general overview of the company was presented. Having great personnel and inventory management is essential to running a business smoothly. Overstaffing or poorly managed inventory raises operating costs, which eventually have an impact. Compared to analogous study subjects among other industries that are essential, there has been a lack of information on sales forecasting for small to big scale manufacturing centers.

Select six attributes (Brand, Model, Color, Phone Type, Market Type, Qty) from the list of nine attributes (Brand, Model, Color, Phone Type, Market Type, Qty, Materialcode, IMEI Number,

Carton Number) for this research that are most pertinent for mobile phone sales. The brand of a mobile phone is particularly critical for mobile phone sales. When a customer considers purchasing a mobile phone, their first thought is often about which brand is best. Similarly, customers are influenced by the model of the mobile phone. Mobile phone models frequently change, and so does customer demand. Another influential factor is color; the color of a mobile phone can affect mobile phone sales to some extent. The type of mobile phone also impacts sales, depending on customer preferences; some customers prefer smartphones, while others select for feature phones. Market type is also a significant determinant of mobile sales; certain mobile phones are in higher demand in local markets, while others are sought after in export markets. Finally, sales quantity is a crucial attribute for mobile phone sales analysis. Attributes such as Materialcode, IMEI number, and Carton number are not relevant to mobile phone sales.

3.2.2 Data collection

An essential step in any research project is data collection. Data determines the effectiveness and achievement of the research goals. Instead of being a single task, gathering data is a methodical, iterative process. This phase depends on the research's defined purpose and business understanding. The degree of research accomplishment is directly implied by planned decisions on data collecting. The expectations for the data requirements were established throughout the business understanding phase, and as a result, many CSV files containing the sales data were received and the information obtained was taken from the cloud storage system that the company utilized to run its SAP operations. The necessary elements for completing the assignment were addressed and provided in accordance with the shareholders. Sales information from November 2017 through January 2024 is gathered for this study from the SAP system of Transsion Manufacturing.

3.2.3 Data Processing

The raw data needed to be preprocessed because it could not be used directly for the various models. All of the previously listed raw data had to be transformed into multiple usable datasets, which required exploratory data analysis, aggregating daily sales, handling potential missing values, transforming features to become more useful, choosing features that actually contributed

to the quantity sold explanation, converting features into One Hot Encoding so that the features could be used in the models, and finally dividing the dataset into a train and a test dataset.

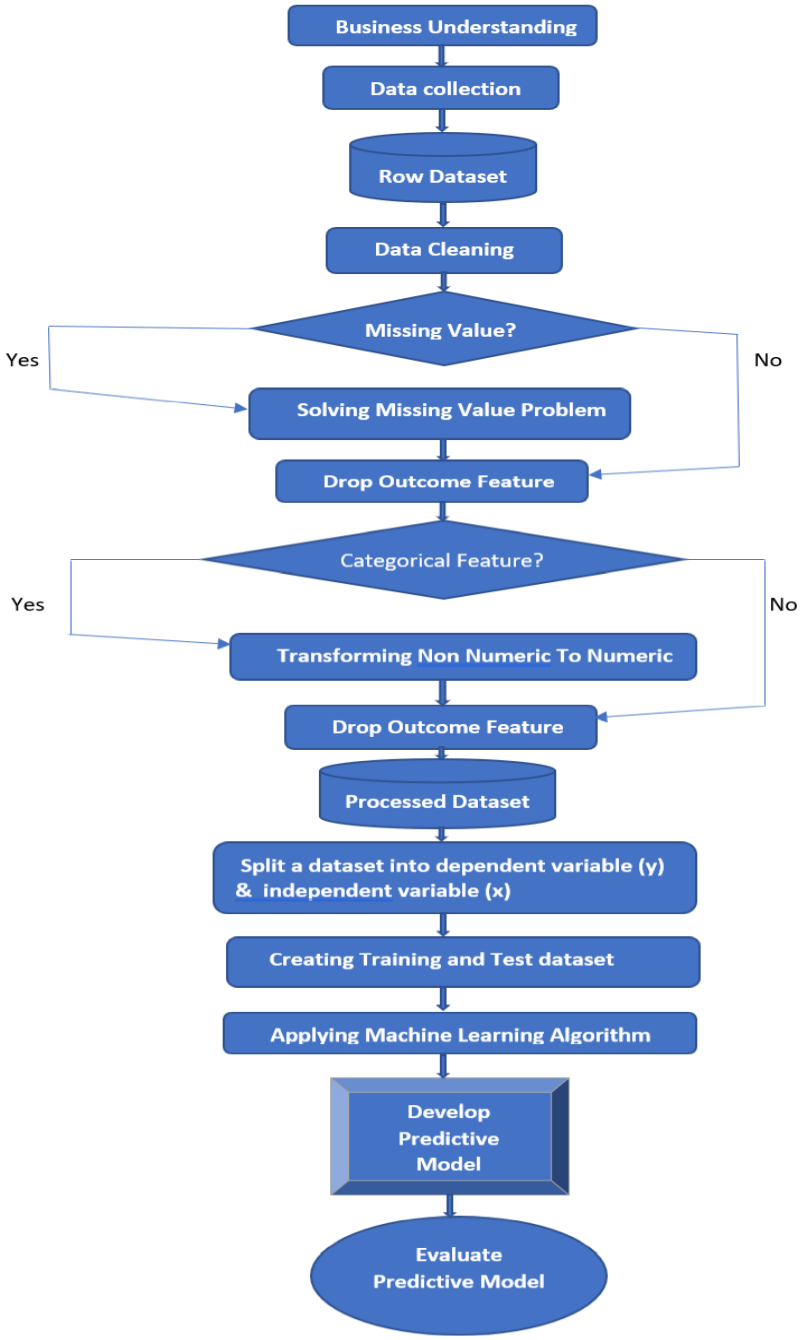


Figure 3.2: Proposed Framework

3.2.4 Exploratory Data Analysis

A first data analysis was necessary in order to comprehend the supplied dataset completely. The purpose of this data analysis was to gain a better understanding of the behavior of the various variables over the course of the seven-year period, as well as how the sales of various products compared to one another and any existing trends or patterns, such as weekly and monthly trends. Lastly, the analysis aimed to identify any potential outliers or missing values. It could be possible to identify underlying patterns in the data that are not immediately evident by conducting an exploratory data analysis. Therefore, it is imperative to ascertain whether the data requires manipulation prior to application and whether more information could be extracted for potential use.

3.2.5 Aggregation

Combining the daily sales was the second step. The data needs to be aggregated to show the overall sold quantity each day for each product and store combination because there could be hundreds of individual purchases of a certain product at a given store every day. There were more rows for each date during the course of the seven-year period since there were more stores, each with a variety of merchandise. Presumably, certain products might have been marketed as individual objects as well as in weight units. There were rows in the aggregated data where the Type of Discount information may have more than one value. This happened if the product was sold that day both with and without a promotion. Multiple sales could be documented if the specials were exclusive to a particular group or if the customer's purchase amount determined the special offers. But in this study, a product's promotion within a certain retailer was treated as a binary variable. Then, based on dates, the meteorological data and extra data were combined, as well as the SAP system data.

3.2.6 Missing Values

The next step was to deal with missing values in each associated feature when the aggregate was finished. A few values were missing from a few variables. One of two things happened: either no amounts sold for a certain product were recorded, or some store station was not operating, therefore some sales data was missing. The mean values of that feature were used to impute the missing values in the sales data. There are two possible explanations for missing sold quantity values: either the product was not sold at that store on that specific date, or the information was not there.

Moreover, it might also be the result of the supplier not making enough goods, which would have prevented any product from being sold even though the demand was probably comparable to previous days. All periods that had one or two days in a row with missing values were handled with imputed zero values. A rolling mean function that determined the mean value of that specific product at that specific store for that day window, beginning from one week prior to the missing day in question, was used to impute periods with three or more missing days in a row [40].

3.2.7 Feature Engineering

One aspect that appeared to hold potential in the scholarly literature was the matching weekday for the sales. Since this feature wasn't expressly included in the raw data, it had to be created using the dates and then added to the dataset. Previous study suggests that this feature may have a high explanatory power since monthly sales tend to follow monthly trends. As a result, sales in January of a given year may be connected with sales in January's past and present. Additionally, sales tended to be irregularly dispersed throughout the entire month, which made it a possible predictor for improving the models' performance an aggregate over the previous months was done to ensure that these facts applied to this particular set of data, and the results demonstrated unequivocally that there was not a uniform distribution of sales over the course of a month. Since time series analysis is the foundation of sales analysis, earlier values of the quantity itself, represented by lags, provide another feature that may be advantageous for some machine learning models [41]. These lags, which were obtained by copying the quantity and moving it ahead by the necessary number of days, represent prior values.

3.2.8 Feature Selection

Depending on the model and available processing capacity, adding a lot of features to machine learning models may or may not be advantageous. Since the model contains more information, adding more features could enhance its performance. But if the features that are added don't add enough predictive power, this might lead to overfitting and an unneeded boost in processing power. A big variance would be the result of over fitting if there were a large number of features relative to the number of data points. There is a possibility that the higher errors stem from the models' sensitivity to the training data. The goal of this thesis was to develop forecasting algorithms that would be useful for the sales of mobile phones alone, excluding other Transsion company products like televisions and other household appliances.

3.2.9 One Hot Encoding

Because they are unable to read categorical data as such, certain machine learning algorithms are not appropriate for use with it. Rather, it would be perceived as numerical data by the models, leading to inaccurate data interpretation and inferior forecasts. As an illustration, consider the mobile brands Tecno and Itel to which Tecno and Itel correlate. The model would interpret Tecno as greater than Itel if it were to employ a single variable with values ranging from Tecno to Itel. In the same vein, Itel would rank higher than Tecno, and so on. This is useless since categorical variables have to be used to model the data because the values could not be compared directly. One-hot encoding (OHE) is a technique for transforming a categorical variable with several distinct categories. Each categorical feature with r categories was converted to r new features by this algorithm. Following that, any observation belonging to category j would receive a 1 in the feature column j that corresponds to it and a 0 in all other columns. OHE therefore produced nine additional features for two category features that had three and six categories, respectively. After that, the original variable which was not one-hot encoded would be eliminated from the dataset [42].

3.2.10 Train-Test Data split

The prepared data had to be divided into two distinct datasets in order to be used in the machine learning models and to assess the models' performance. A set for testing and training. There was 80% training data and 20% test data. The time series characteristics should be considered because sales forecasting is by its very nature a time series problem. It involves segmenting the data so as to maintain the timestamps on the recorded data. To avoid utilizing data from the future to anticipate events from the past, it is crucial to keep the data organized. We just need to set up a way to allow the rest of the pipeline to call it as a service whenever needed. There isn't a different strategy for each layer of our pipeline in this section. The Sklearn library package is utilized in this project to create the train-test split function, which is then incorporated into our prototype.

3.3 Model Training and Evaluation

Following multiple iterations of data processing and preparation, we have at last obtained the desired dataset upon which to build a predictive model. The model training pipeline component should be designed to be able to switch between different machine learning algorithms and then optimize those algorithms in an effective way to determine which model performs the best. In this

research, we have developed a number of machine learning algorithms, each of which has gone through the process of hyperparameter optimization to determine which model performs the best. The process of hyperparameter optimization, which requires model training for a variety of combinations of hyper parameters, takes a long time and should be optimized itself due to the massive amount of data. Some of the most well-known machine learning techniques that employ various methods for model training are the machine learning algorithms that were employed in this application. These techniques have been carefully chosen so that the effectiveness of various strategies for applying sales forecasting may be compared. There have been four machine learning algorithms used: Random Forest, SVM, KNN, and Naïve bias. These algorithms are based on a number of methods, including statistical techniques, ensemble trees, bias theory, and gradient boosting trees. All algorithms were subjected to a hyper parameter optimization procedure, and the model with the highest performance among all approaches was identified. Subsequently, the models undergo comparative analysis to determine which one is optimal for future forecasting. The test data is retrieved from the pipeline's data split service as part of the model evaluation process, and the model's performance is assessed using the most dependable classification analysis metric the Adjusted F1-Score which can be used to gauge the classification model's correctness. Since this statistic is approaching the value of 1, it is performing better. The percentage of the target variable's variation that can be predicted from the independent variable is known as the F1-Score, or coefficient of determination [43]. The memory consumption of the processes in Spark requires careful management of a few configuration variables. The way Spark uses memory resources for data processing differs from typical Python library packages, which gives it an edge. While Spark loads the necessary portion of the data into memory only when it needs it for calculations, Python loads all of the data into memory at the start of the operation. With the aid of Spark memory management configuration settings, we must oversee the processes' memory management. In actuality, this is one of the difficulties in managing Big Data for this research. Displays the last set of choices that were employed. These settings were made by keeping an eye on how much memory the Windows operating system's applications were using, as well as how the Spark user interface's applications behaved.

3.3.1 Implementation Technique and Tools

Numerous development technologies are employed in this study to create and execute the suggested approach. Here is a description of the implementation tools together with their rationale.

Tools for Software a review of the software products that are now available on the market with their libraries is done in order to determine which software tool is best for predicting sales of Transsion manufacturing using various algorithms. During the investigation, we learned that although some tools, like SVM, are specific and can only be used with one type of machine learning algorithm, others, like Python, are general and can be used with both. Before selecting the tools, we considered the following factors since they help us locate the appropriate software tools and related libraries. The computer language that is utilized to execute the algorithm is the main contributing component. The second is to select technologies that have sufficient learning resources, such prior experience and free video courses. Utilizing the tools on computers with limited resources (such as CPU only) is the third prerequisite. Python has been utilized as the programming language in the Anaconda environment to build various algorithms, together with the Jupyter notebook and Anaconda. These tools operate in the widely used Python programming language and satisfy all requirements.

3.3.2 Jupyter Notebook

Users can create and share documents with live code, equations, graphics, and narrative text using the open-source web application Jupyter Notebook. It was created initially for Python and currently works with more than 40 programming languages, making it a flexible tool for a variety of tasks like machine learning, scientific computing, and data analysis. Because of its interactive features, which let users write and run code in short bursts, programming and analysis become more exploratory and iterative. Researchers, educators, and data scientists increasingly choose Jupyter Notebook for collaborative work and reproducible research because of its user-friendly design and support for rich media integration.

3.3.3 Hardware Tools

The hard drive is used to store the datasets, the GPU is used to increase processing power for effective training, the CPU is used to test the models, and RAM is used to train the model efficiently using the CPU and GPU working together. The specifications of the laptop utilized for the experiment are displayed in Figure 3.3.

- Computer Type: - Lenovo core i7 9th Gen
- Operating System: - Windows 10 Pro
- Install RAM: - 8.00 GB
- Storage Disks: - 1TB
- Processor: -Intel(R) Core (TM) i7-4510U CPU @ 3.30GHz 2.60 GHz

Figure 3.3: Laptop Specification

3.3.4 Evaluation

In order to assess a model's utility and performance, it is necessary to analyze it. In computational tasks such as classification and detection, evaluation metrics such as accuracy, precision, recall, and F1-score are used to identify which instance belongs to which class. Those measures were computed using the categorization metrics. The classification metrics give details about a model's performance for every class. The confusion matrix value (False Negative) served as the basis for the computation of all the metrics mentioned above, including TP (True Positive), TN (True Negative), FP (False Positive), and FN. The confusion matrix is shown in Table 3.1.

Table 3.1: Confusion matrix

		Actual Value	
		Negative	Positive
Predicted Value	Negative	TN	FN
	Positive	FP	TP

Where: TP: The instance's actual class is positive, which matches the prediction. TN: Predicted to be negative, but actual class of the instances is negative. FP: Predicted as positive, but actual class of the cases is negative. FN: Predicted as negative, but actual class of the instances is positive the following evaluation metrics are calculated using the confusion matrix's summarized data. The

answer to the issue of how frequently the model correctly predicts the classes, that is, which phone brand and type are more popular is found in accuracy, which is defined as the overall correct classification of instances into their belongingness. This formula is used to calculate it.
$\boldsymbol{Accuracy}$= TP+TN/TP+TN+FP+FN

The degree of precision provides information about the accuracy of positive value predictions. Example: Predicting the brand and type of phone and the frequency at which the prediction is accurate. This formula is used to calculate it.
$\boldsymbol{Precision}$= TP/TP+FP

Recall: Also referred to as sensitivity, recall expresses the classifier's sensitivity in identifying positive cases. This is how it is computed.
$\boldsymbol{Recall}$= TP/TP+FN

F1-account is the harmonic mean of precision and recall. The F1-score has a lowest value of 0, indicating that one of the metrics is zero. It denotes flawless recall or precision. This is how it is computed.

$\boldsymbol{F1 - Score}$ = 2*TP/2*TP+FP+FN

CHAPTER FOUR

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Overview

This chapter describes the process used to conduct the experiment and analyzes the results using the stages outlined in the suggested framework design. The proposed framework architecture is guaranteed to be realized by means of this experimental evaluation. It describes the main experiments conducted, the interpretations and performance evaluations of the prediction model, and how the machine learning models were developed using Random Forest, KNN, Navies Bayes, and SVM. The experiment's subsequent tasks are completed with Jupyter Notebook tools. All of the preparatory work done on the dataset, as well as a few of the more significant tasks completed, were discussed in chapter three. In order to arrive at the best model to meet the goal stated in chapter one, a synopsis of the main experiments conducted is presented in this section along with the recommended framework design. The essential data is fed into the Jupyter Notebook for the model building process once it has undergone the preparatory steps outlined in the preceding sections. To make the preprocessed data compatible with Jupyter Notebook, it is transformed to CSV format. Numerous experiments are carried out to determine the algorithms that yield prediction models with different sizes, precisions, and accuracies. Additionally, a number of activities pertaining to conducting and assessing model-building experiments, choosing the most suitable model, and offering justifications for the model of choice are presented in this part.

4.2 The Proposed Architecture

The proposed architecture in this research is shown in Figure 4.1 The procedures taken in the analysis and prediction of the mobile phone sales state from the Transsion Manufacturing mobile phone sales dataset are displayed in the suggested architecture. The six main stages of the proposed architecture are as follows: Business Understanding, which entails comprehending the problem to identify the pertinent attribute, and Data Understanding, which entails creating tasks for relevant attributes using sales prediction. Next are Data Preprocessing tasks, which involve cleaning up attributes found in the dataset. Data splitting: 80 percent of the total dataset used to train the model is used for training, and 20 percent is used for testing the model's performance. Classification uses four supervised machine learning techniques (Random Forest, KNN, Naïve Bayes and SVM) to

create the train model, which is then combined with the test dataset to produce the sales prediction output. In the end, the mobile brand ITEL or TECNO and the mobile type SMART or FEATURE are predicted by the mobile sales prediction model.

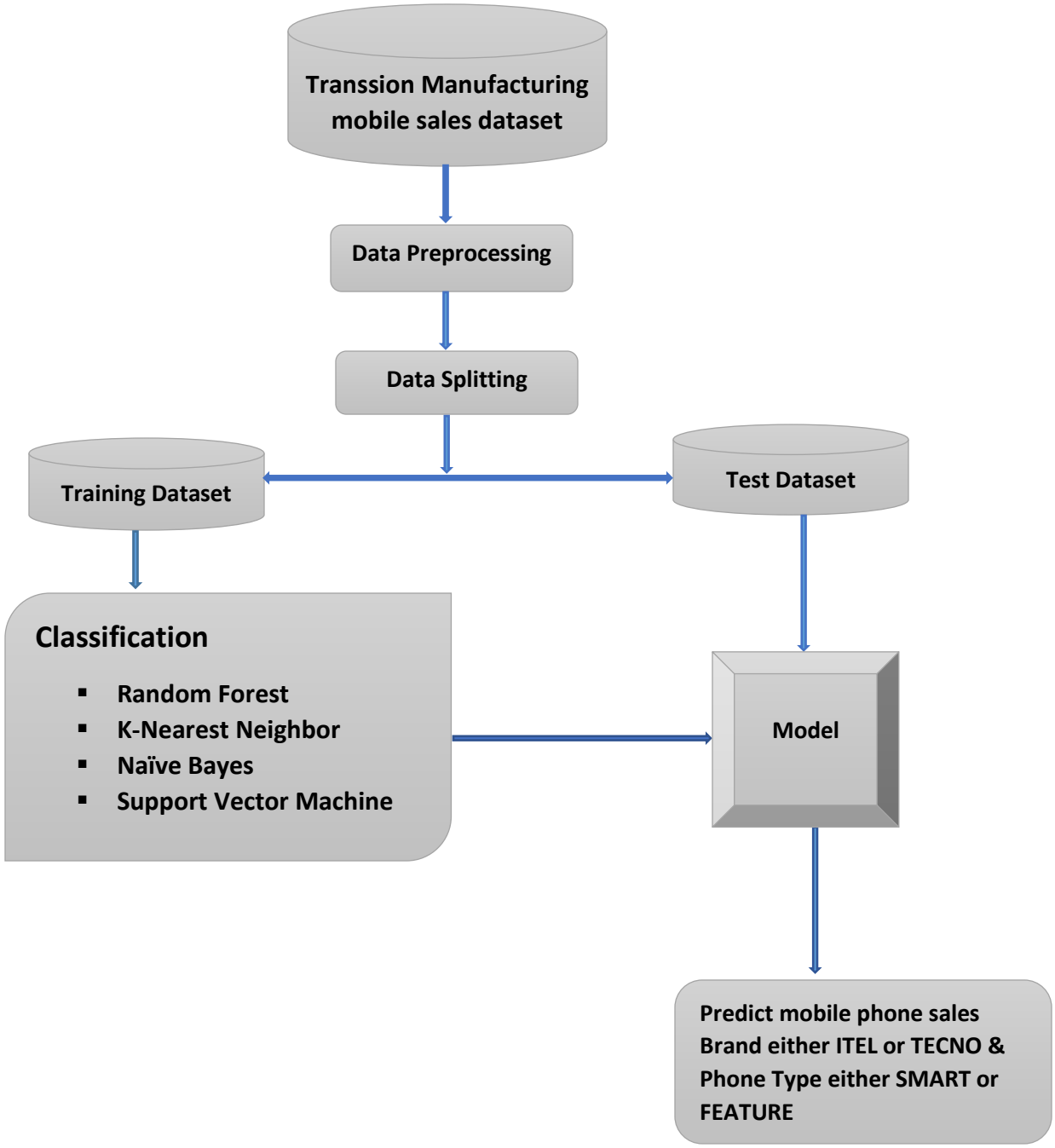


Figure 4.1: Proposed Architecture

4.2.1 Predictive Modelling

The main idea behind predictive modeling is creating a model with predictive capabilities. To create those predictions, such a model usually consists of a Machine Learning algorithm that acquires specific attributes from a training dataset. A group of mathematical methods or models that assist in determining a mathematical link between a goal or dependent variable and the predictor or independent variables are together referred to as predictive modeling. It assists in estimating the likelihood of a result when a series of independent variables are run through the model. Random Forest, KNN, Naïve Bayes and SVM models can all be applied to predictions [44].

The main objective of this study is to use computational methods to forecast sales by analyzing the sales pattern of Transsion manufacturing. In light of this, the study sought to determine which machine learning technique was most effective at forecasting mobile phone sales. Several machine learning methods were investigated during this work, and the most successful ones were applied to identify models in the data. In earlier related works, the effectiveness of each machine learning model was investigated and assessed. Subsequently, the most stable and successful algorithms were chosen based on their capacity to make accurate predictions. The machine learning methods that were chosen for this thesis include Random Forest, KNN, Naive Bayes and SVM. Each is chosen according to their merits and historical results from earlier studies. Various literatures have revealed that Random Forest, KNN, Naive Bayes and SVM are some of the frequently used classifier algorithms for prediction and classification. The study employed the aforementioned quartet of algorithms to construct four distinct models for the purpose of predicting and classifying mobile sales.

The following factors led to the selection of Naïve Bayes: it is simple to use; Naïve Bayes classifiers can be trained fast and the classification process is quicker than with other models. It is insensitive to unimportant information, and it can process a big and distinct amount of data [45].

The machine learning technique Random Forest Regression uses an ensemble of decision trees to ascertain the association between a dependent variable and one or more independent variables. It's a variant on the Arbitrary Woods calculation, which is often applied to orderly tasks but specifically meant for relapse problems.

Essentially, Random Forest Regression uses the combined knowledge of several decision trees to forecast regression problems with accuracy. It is famous for its robustness, ability to handle large and complicated datasets, and capacity to identify non-linear relationships between variables. The technique finds use in a wide range of domains, including marketing, finance, and healthcare, where precise regression forecasts are critical to sound decision making [7].

K-nearest neighbor (KNN) is chosen because, in a given training set, data is classified based only on its nearest neighbor (or neighbors). It is hard to envision a simpler technique than KNN. Learning doesn't necessitate assuming whatever about the concepts' properties and complex concepts can be taught through local approximation with straightforward methods [46].

Because SVM employs the kernel trick, which enables it to incorporate expert knowledge about the problem by engineering the kernel, and because it is based on the structural risk minimization principle, which aims to minimize an upper bound of generalization error, it is chosen for this study. While other models may tend to settle into a local optimal solution, the SVM model is a quadratic program that is linearly constrained, meaning that its solution is always globally optimal [47].

4.3 Dataset for Experiment

For this study, the data is collected from Transsion manufacturing SAP system and also the study specifies the main reasons (attributes) mobile phone sales, the data from Transsion manufacturing contains mobile phone BRAND, PHONE TYPE, MODEL, COLOR, MARKET TYPE and QTY attributes. The data contained 106,990 mobile sales records with 6 attributes one of the attributes is dependent field representing mobile BRAND, other dependent field represent PHONE TYPE; the data record for this research from Transsion manufacturing from 2017 G.C to 2023 G.C. The attributes, description and categorical values explored for the study are described.

4.3.1 Install and Import Important Library

Install basic necessary packages by pip install from Anaconda library in Jupyter Notebook and Import necessary libraries from different environments in Jupyter Notebook.

4.3.2 Read CSV Dataset and Checking Missing value

Read mobile phone CSV format sales dataset from Jupyter Notebook database as pandas and display the appearance of the dataset's head.

4.3.3 Remove Missing value

Import SimpleImputer package from sklearn that has uses for removing missing values and replace categorical feature missing value by “missing” and numerical feature missing value by “mean” of numeric value.

4.3.4 Splitting a Dataset in to dependent and independent variables

Splitting a dataset into dependent variable (y) and independent variable (x) is done for data analysis. That means mobile phone “BRAND” are dependent variables and remaining attributes are independent variables in case of experiment one. However, “PHONE TYPE” are dependent variables and remaining attributes are independent variables in case of experiment two.

4.3.5 Transforming Categorical feature in to numeric feature

Import One Hot Encoder and Column Transformer from Sklearn. To make the dataset suitable for machine learning we perform transforming categorical features into numeric features.

4.3.6 Creating Training and Test Dataset

To create training and test dataset, we Import ‘train test split’ library from Sklearn. Then, based on the 80:20 ratio, 80 percent of a dataset is used for training and 20 percent of dataset for a test.

4.4 Modeling using Random Forest

In this experiment, the Random Forest algorithm is used to develop a prediction model by Import Random forest classifier from sklearn. The relationship between a group of independent (explanatory) variables and a categorical dependent variable is investigated using random forest analysis. When the dependent variable just has two values either ITEL or TECNO and FEATURE or SMART phone.

4.4.1 Experiment one (Predict Mobile Brand ITEL or TECNO)

The dataset consists of 106,990 observations and 6 attributes with one dependent attribute/class (which is mobile phone “BRAND”) and the remaining predictor (independent) attributes. This experiment helps to predict the mobile sales of either ITEL Brand or TECNO Brand.

In this study an attempt is made to evaluate Random Forest predictive model of mobile phone sales into either ITEL Brand or TECNO Brand and print the accuracy, precision, recall and f1-score result. The overall performance of the model achieves 99.6 % accuracy.

Table 4.1: Confusion matrix of Random Forest experiment one

Random Forest (ITEL or TECNO)		**Predicted label**	
True label		**ITEL**	**TECNO**
	ITEL	**9405**	**81**
	TECNO	**0**	**11912**

From the Confusion matrix table, we can understand that True label are ITEL and Predicted label are ITEL (True Positive-TP) are 9405 class, True labels are TECNO and Predicted labels are TECNO (True Negative-TN) are 11912 class. Similarly, True labels are ITEL and the Predicted labels are TECNO (False Positive-FP) are 81 class, True labels are TECNO and Predicted labels are ITEL (False Negative-FN) are 0 class. From the experiment we can conclude that only 81 class are miss classified.

Based on the above information extracted from the confusion matrix, the model constructed by random forest registered 99.6% Accuracy, 99% precision, 100% recall and 95% f1-score.

4.4.2 Experiment two (Predict Mobile Phone type FEATURE phone or SMART phone)

The dataset consists of 106,990 observations and 6 attributes with one dependent attribute/class (which is mobile phone "PHONE TYPE") and the remaining predictor (independent) attributes. This experiment helps to predict the mobile sales of either FEATURE phone or SMART phone.

In this study an attempt is made to evaluate Random Forest predictive model of mobile phone sales into either SMART Phone or FEATURE phone and print the accuracy, precision, recall and f1-score result. The overall performance of the model achieves 96.8 % accuracy.

Table 4.2: Confusion matrix of Random Forest experiment two

Random Forest (FEATURE or SMART)		Predicted label	
True label	FEATURE	FEATURE	SMART
		10453	388
	SMART	291	10266

From the Confusion matrix table, we can understand that True labels are FEATURE and Predicted labels are FEATURE (True Positive-TP) are 10453 class, True labels are SMART and Predicted labels are SMART (True Negative-TN) are 10266 class. Similarly, True labels are FEATURE and the Predicted labels are SMART (False Positive-FP) are 388 class, True labels are SMART and Predicted labels are FEATURE (False Negative-FN) are 291 class. From the experiment we can conclude that only 679 class are miss classified.

Based on the above information extracted from the confusion matrix, the model constructed by random forest registered 96.8% Accuracy, 96% precision, 96% recall and 96% f1-score.

4.5 Modeling using KNN

This subsection deals with how the KNN prediction model is developed and how the information gained was calculated by Import KNN classifier from Sklearn.In this experiment a Prediction model building is done using KNN algorithm to predict mobile sales BRAND either ITEL or TECNO and FEATURE or SMART phone.

4.5.1 Experiment one (Predict Mobile Brand ITEL or TECNO)

The dataset consists of 106,990 observations and 6 attributes with one dependent attribute/class (which is mobile phone “BRAND”) and the remaining predictor (independent) attributes. This experiment helps to predict the mobile sales of either ITEL Brand or TECNO Brand.

In this study an attempt is made to evaluate KNN predictive model of mobile phone sales into either TECNO Brand or ITEL Brand and print the accuracy, precision, recall and f1-score result. The overall performance of the model achieves 98 % accuracy.

Table 4.3: Confusion matrix of KNN experiment one

KNN (ITEL or TECNO)		Predicted label	
		ITEL	TECNO
True label	ITEL	9202	114
	TECNO	8	12074

From Confusion matrix table we can understand that True labels are ITEL and Predicted labels are ITEL (True Positive-TP) are 9202 class, True labels are TECNO and Predicted labels are TECNO (True Negative-TN) are 12074 class. Similarly, True labels are ITEL and the Predicted labels are TECNO (False Positive-FP) are 114 class, True labels are TECNO and Predicted labels are ITEL (False Negative-FN) are 8 class. From the experiment we can conclude that only 122 class are miss classified.

Based on the above information extracted from the confusion matrix, the model constructed by KNN achieved 98% Accuracy, 98% precision, 99% recall and 98% f1-score.

4.5.2 Experiment two (Predict Mobile Phone type FEATURE phone or SMART phone)

The dataset consists of 106,990 observations and 6 attributes with one dependent attribute/class (which is mobile phone "PHONE TYPE") and the remaining predictor (independent) attributes. This experiment helps to predict the mobile phone sales either FEATURE phone or SMART phone. In this study an attempt is made to evaluate KNN predictive model of mobile sales into either SMART phone or FEATURE phone and print the accuracy, precision, recall and f1-score result. The overall performance of the model achieves 96 % accuracy.

Table 4.4: Confusion matrix of KNN experiment two

KNN (FEATURE or SMART)		Predicted label	
		FEATURE	SMART
True label	FEATURE	10402	439
	SMART	382	10175

From Confusion matrix table we can understand that True labels are FEATURE and Predicted labels are FEATURE (True Positive-TP) are 10402 class, True labels are SMART and Predicted labels are SMART (True Negative-TN) are 10175 class. Similarly, True labels are FEATURE and the Predicted labels are SMART (False Positive-FP) are 439 class, True labels are SMART and Predicted labels are FEATURE (False Negative-FN) are 382 class. From the experiment we can conclude that only 821 class are miss classified.

Based on the above information extracted from the confusion matrix, the model constructed by KNN registered 96% Accuracy, 95% precision, 96% recall and 95% f1-score.

4.6 Modeling using Naïve Bayes

This subsection deals with how the Naïve Bayes prediction model is developed and how the information gained was calculated. In this experiment a Prediction model building is done using Naïve Bayes algorithm to predict mobile sales BRAND either ITEL or TECNO and FEATURE or SMART. Import different types of Naïve Bayes classifiers from Sklearn and use multinomial classifiers which is best for this model.

4.6.1 Experiment one (Predict Mobile Brand ITEL or TECNO)

The dataset consists of 106,990 observations and 6 attributes with one dependent attribute/class (which is mobile phone "BRAND") and the remaining predictor (independent) attributes. This experiment helps to predict the mobile sales either ITEL Brand or TECNO Brand. In this study we evaluate Naïve Bayes predictive model of mobile phone sales into either TECNO Brand or ITEL Brand and print the accuracy, precision, recall and f1-score result. The overall performance of the model achieves 98 % accuracy.

Table 4.5: Confusion matrix of Naïve Bayes experiment one

Naïve Bayes (ITEL or TECNO)		**Predicted label**	
True label		**ITEL**	**TECNO**
	ITEL	**9157**	**159**
	TECNO	**33**	**12049**

From Confusion matrix table we can understand that True labels are ITEL and Predicted label are ITEL (True Positive-TP) are 9157 class, True labels are TECNO and Predicted labels are TECNO (True Negative-TN) are 12049 class. Similarly, True labels are ITEL and the Predicted labels are TECNO (False Positive-FP) are 159 class, True labels are TECNO and Predicted labels are ITEL (False Negative-FN) are 33 class. From the experiment we can conclude that only 192 class are miss classified.

Based on the above information extracted from the confusion matrix, the model constructed by Naive Bayes registered 98% Accuracy, 98% precision, 99% recall and 98% f1-score.

4.6.2 Experiment two (Predict Mobile Phone type FEATURE phone or SMART phone)

The dataset consists of 106,990 observations and 6 attributes with one dependent attribute/class (which is mobile phone "PHONE TYPE") and the remaining predictor (independent) attributes. This experiment helps to predict the mobile sales either FEATURE phone or SMART phone. In this study we attempt to evaluate Naïve Bayes predictive model of mobile phone sales into either SMART phone or FEATURE phone and print the accuracy, precision, recall and f1-score result. The overall performance of the model achieves 96 % accuracy.

Table 4.6: Confusion matrix of Naïve Bayes experiment two

Naïve Bayes (FEATURE or SMART)		**Predicted label**	
		FEATURE	**SMART**
True label	**FEATURE**	**10402**	**439**
	SMART	**382**	**10175**

From Confusion matrix table we can understand that True labels are FEATURE and Predicted labels are FEATURE (True Positive-TP) are 10402 class, True labels are SMART and Predicted labels are SMART (True Negative-TN) are 10175 class. Similarly, True labels are FEATURE and the Predicted labels are SMART (False Positive-FP) are 439 class, True labels are SMART and Predicted labels are FEATURE (False Negative-FN) are 382 class. From the experiment we can conclude that only 821 class are miss classified.

Based on the above information extracted from the confusion matrix, the model constructed by Naïve Bayes registered 96% Accuracy, 95% precision, 96% recall and 95% f1-score.

4.7 Modeling using SVM

The primary purpose of the supervised machine learning algorithm SVM (Support Vector Machine) is to categorize data into distinct classes which means either ITEL or TECNO and FEATURE or SMART. SVM uses a hyperplane, which functions as a decision boundary between the different classes, in contrast to other algorithms. Support vectors are those data points that are closest to the hyperplane.

4.7.1 Experiment one (Predict Mobile Brand ITEL or TECNO)

The dataset consists of 106,990 observations and 6 attributes with one dependent attribute/class (which is mobile phone "BRAND") and the remaining predictor (independent) attributes. This experiment helps to predict the mobile sales of either ITEL Brand or TECNO Brand. In this study an attempt is made to evaluate SVM predictive model of mobile phone sales into either TECNO Brand or ITEL Brand and print the accuracy, precision, recall and f1-score result. The overall performance of the model achieves 98 % accuracy.

Table 4.7: Confusion matrix of SVM experiment one

SVM (ITEL or TECNO)		**Predicted label**	
True label		**ITEL**	**TECNO**
	ITEL	**9157**	**159**
	TECNO	**33**	**12049**

From Confusion matrix table we can understand that True labels are ITEL and Predicted labels are ITEL (True Positive-TP) are 9157 class, True labels are TECNO and Predicted labels are TECNO (True Negative-TN) are 12049 class. Similarly, True labels are ITEL and the Predicted labels are TECNO (False Positive-FP) are 159 class, True labels are TECNO and Predicted labels are ITEL (False Negative-FN) are 33 class. From the experiment we can conclude that only 192 class are miss classified.

Based on the above information extracted from the confusion matrix, the model constructed by SVM registered 98% Accuracy, 98% precision, 99% recall and 98% f1-score.

4.7.2 Experiment two (Predict Mobile Phone type FEATURE phone or SMART phone)

The dataset consists of 106,990 observations and 6 attributes with one dependent attribute/class (which is mobile phone "PHONE TYPE") and the remaining predictor (independent) attributes. This experiment helps to predict the mobile sales of either FEATURE phone or SMART phone. In this study an attempt is made to evaluate SVM predictive model of mobile phone sales into either SMART phone or FEATURE phone and print the accuracy, precision, recall and f1-score result. The overall performance of the model achieves 95 % accuracy.

Table 4.8: Confusion matrix of SVM experiment two

SVM (FEATURE or SMART)		**Predicted label**	
True label		**FEATURE**	**SMART**
	FEATURE	**10234**	**607**
	SMART	**327**	**10230**

From Confusion matrix table we can understand that True labels are FEATURE and Predicted labels are FEATURE (True Positive-TP) are 10234 class, True labels are SMART and Predicted labels are SMART (True Negative-TN) are 10230 class. Similarly, True labels are FEATURE and the Predicted labels are SMART (False Positive-FP) are 607 class, True labels are SMART and Predicted labels are FEATURE (False Negative-FN) are 327 class. From the experiment we can conclude that only 821 class are miss classified.

Based on the above information extracted from the confusion matrix, the model constructed by SVM registered 95% Accuracy, 95% precision, 96% recall and 95% f1-score.

4.8 Comparison of Machine Learning Models

Table 4.9 presents an overview of the outcomes from all four models after they have been trained on about 85,592 occurrences. The results show that Random Forest produced the best accuracy.

However, KNN, SVM, and the Naive Bayes model all fared well in terms of accuracy and precision at the prediction level. The accuracy and precision of these machine learning models are used to evaluate (compare) them. The sales of mobile phones produced by Transsion Manufacturing were predicted using the four machine learning models (Random Forest, KNN, Naive Bayes, and SVM).

Table 4.9: Model comparison

	Random Forest		**KNN**		**Naïve Bayes**		**SVM**	
	Exp one (ITEL or TECNO)	**Exp two** (FEATURE or SMART)	**Exp one** (ITEL or TECNO)	**Exp two** (FEATURE or SMART	**Exp one** (ITEL or TECNO)	**Exp two** (FEATURE or SMART	**Exp one** (ITEL or TECNO)	**Exp two** (FEATURE or SMART)
Accuracy	99.6 %	96.8 %	98 %	96 %	98 %	96 %	98 %	95 %
Precision	99 %	96 %	98 %	95 %	98 %	95 %	98 %	95 %

The machine learning algorithm Random Forest achieved an astounding 99.6% accuracy in experiment one and 96.8% accuracy in experiment two, as indicated in the table above. The remaining three models also have excellent accuracy in experiment one and two, KNN score 98% and 96% accuracy in experiment one and two respectively, Naïve Bayes score 98% and 96% accuracy in experiment one and two respectively and SVM had accuracy rates of 98% and 95% in experiment one and two respectively. The study's findings are presented to senior Transsion sales managers and other domain specialists, who provide technical evaluations to verify the accuracy of the data and the study's conclusions.

4.7 Discussion of result

The primary goal of developing the prediction model, as previously said in chapter one, is to construct for each type of mobile phone sales (ITEL or TECNO and SMART or FEATURE) that would aid in forecasting the expected status of a new mobile phone sales in terms of these features.

The dataset is used to carefully test the study's experiment. The datasets that are tested are those that are covered in the section on data pretreatment. As previously mentioned, the Transsion Manufacturing mobile sales dataset has 106,990 records. Of these, 20% are designated as test samples, with the remaining 80% designated as training. A table arrangement called a confusion matrix, which shows a model's performance, is also included in the summary of the outcome. A standard confusion matrix is made up of rows and columns, where the number of instances in a True label is represented by each row, and the number of instances in a predicted label is represented by each column. The total number of true positives, false positives, false negatives, and true negatives is represented by a confusion matrix in predictive analytics.

In summary, this chapter covered the procedures for creating the four prediction models Random Forest, KNN, Naïve Bayes, and SVM Machine Learning algorithms as well as the methods for evaluating their performance. In the interim, a cross-tabulation was provided to show the model's output, a confusion matrix was shown to compute accuracy, precision, recall and F1-score.

The study attempted to answer the research questions stated in chapter one. The first research question is “Which attributes are more critical for Predicting Transsion Manufacturing mobile phone sales?” According to the results obtained, mobile “BRAND”,” COLOR” “MODEL” “MARKET TYPE” “QTY” and mobile “PHONE TYPE” are effective for predicting mobile phone sales in Transsion Manufacturing.

In addition, in relation to research question two, “Which classification algorithm is suitable for constructing a model that predicts mobile phone sales?” Random Forest emerged the most suitable for mobile phone sales prediction model Accuracy 99.6 % and 96.8 % in experiment one and experiment two respectively.

Finally, in relation to research question three “What is the performance of the model in predicting mobile phone sales?” Random Forest by 98.2%, KNN by 97%, Naïve Bayes by 97% and SVM by 96.5% mean accuracy of experiment one and two, were modeled to find out how machine learning can be utilized in making decision whether to extend mobile phone sales ITEL or TECNO and SMART or FEATURE.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Overview

Forecasting sales and demand has long been a key concern for the industrial sector. Having a precise estimate of the sales volume enables all supply chain participants to make appropriate plans and decisions. As a result, the supply chain operates in a more reliable, strong, efficient, and sustainable manner. The volume of data produced and kept by supply chain participants is also growing rapidly. The study conducted for this research has shown that businesses require product sales forecast systems to handle enormous amounts of data, and that the speed and accuracy of data processing technologies play a key role in corporate decision making. The machine learning techniques discussed in this article can offer a useful mechanism for data tweaking and decision-making. Businesses must arm themselves with specialized techniques in order to make predictions about the demand and sales of their products in order to maximize profit, considering the many forms of customer behavior. The full study's findings are outlined in this chapter. It is separated into two primary sections: recommendations and conclusion. It offers suggestions for future work.

5.2 Conclusions

Transsion Manufacturing provided the essential data for the experiment on a variety of Excel sheets, totaling 106,990 records in the dataset. Following the application of the required preprocessing operations to the dataset, 106,990 data were ready for the experiment. Algorithms like Random Forest, KNN, Naïve Bayes, and SVM were experimented with for developing classification and prediction models. Jupyter Notebooks using Python programming were utilized as the tools to replicate every aspect of the experiment. After using the confusion matrix, calculations for accuracy and sensitivity were made. All models (Random Forest, KNN, Naïve Bayes and SVM) yielded remarkable results when it comes to correctly classifying instances, Random Forest with accuracy of 99.6 %, 96.8 %, in the first and second experiments respectively, KNN with accuracy of 98 %, 96 %, in the first and second experiments respectively, Naïve Bayes with accuracy of 98 %, 96 %, in the first and second experiments respectively and SVM with accuracy of 98 %, 95 %, in the first and second experiments respectively.

Based on the above conclusion all the researcher questions are addressed as follows. Question one "Which attributes are more critical for Predicting Transsion Manufacturing mobile phone sales?" mobile "BRAND"," COLOR" "MODEL" "MARKET TYPE" "QTY" and mobile "PHONE TYPE" are found to be effective and relevant (important) predictors for the target class mobile Brand (ITEL or TECNO) and (FEATURE or SMART) for predicting mobile phone sales in Transsion Manufacturing.

In addition, in relation to research question two, "Which classification algorithm is suitable for constructing a model that predicts mobile phone sales?" Random Forest was the most suitable mobile sales prediction model with Accuracy 99.6 %, 96.8 % in Experiment one and two respectively.

In relation to research question three "What is the performance of the model in predicting mobile phone sales?" Random Forest with accuracy of 99.6 %, 96.8 %, in the first and second experiments respectively, KNN with accuracy of 98 %, 96 %, in the first and second experiments respectively, Naïve Bayes with accuracy of 98 %, 96 %, in the first and second experiments respectively and SVM with accuracy of 98 %, 95 %, in the first and second experiments respectively, were modeled to find out how machine learning can be utilized in making decision whether to mobile sales ITEL or TECNO and FEATURE or SMART.

Based on the experiment results, it can be inferred that machine learning techniques can be utilized successfully in the manufacturing sector to produce product sales prediction models that have a reasonable degree of accuracy. Transsion Manufacturing may thus utilize these machine learning models to predict market demand for specific mobile brands (ITEL or TECNO) and phone types (FEATURE or SMART) ahead of time. Transsion Manufacturing is anticipated to function extremely well in identifying and forecasting their mobile phone sales by using this approach, and to be able to provide the anticipated investment returns. A larger dataset with more and more diverse attributes involving even different kinds of mobile phone manufacturing companies could have produced a better modeling, though, given the quality and size of the dataset used, in addition to the Machine Learning tools and techniques, which are essential factors for the modeling performance. In order to help solve additional issue areas in the nation's manufacturing industries, it may have made it possible for the research to employ greater data with more qualities than those used in this study. The researcher believes that an improved model would have been produced if

additional mobile manufacturing company datasets with big data sizes and a variety of product types, as well as additional techniques, had been included in the model building experimentation. The experimentation was based on collected Transsion Manufacturing datasets with just 106,990 datasets. As a result, the recommendations in the following section are derived from the research's findings.

5.3 Recommendations

Although the primary goal of the research is academic, Transsion Manufacturing and other researchers with a similar interest will benefit greatly from it. Even if the study's findings are encouraging, there are still some unanswered questions that need to be answered in order to improve the inclusive approach and get it operational. In light of this, the researcher suggests the following topics for further investigation based on this study: -

- We suggest that future research integrate the prediction model for mobile phone sales with the production of mobile phones, based on the best model that was suggested in this study.
- In the future, we can expand our research by incorporating additional well-known mobile phone brands such as Samsung and Apple.
- Machine Learning algorithms can be used to limit or possibly completely extinguish the prediction of mobile phone sales in Transsion Manufacturing products. But there are generally other problems that exist in every Manufacturing industry. Therefore, interested researchers can look into different natural problems in the manufacturing industry from a Machine Learning perspective.
- Various Algorithms for Classification: Neural Network, Decision Tree, and Logistic Regression Learning may be applied to a manufacturing center in Ethiopia using a larger dataset Sample size and other features, like as the type of product and the economic condition of the clients, the government economic system, competition in the market, and unexpected occasions. to see if different researchers can come to different conclusions.
- The model needs to be slightly adjusted because different manufacturing companies implement different marketing strategies and customer handling policies. As a result, this mobile phone sales prediction model is specific to Transsion Manufacturing PLC, and we recommend that future research include sales data from other mobile phone manufacturing companies.

References

[1] Joonkoo Lee, Jong-Cheol Kim,Jinho Lim, "Globalization and Divergent Paths of Industrial Development: Mobile Phone Manufacturing in China, Japan,," *JOURNAL OF CONTEMPORARY,* vol. 46, no. 2, pp. 222-246, 2016.

[2] Xiang, Caifen, "4Ps Empirical Analysis on the Export of TRANSSION to Africa," *ATLANTIS PRESS,* vol. 335, 2019.

[3] Garud Akshada Anil*1, Chavan Ritambara Shankar*2, Bobade Prachi Santosh*3,, "SALES FORECASTING USING MACHINE LEARNING TECHNIQUES," *International Research Journal of Modernization in Engineering Technology and Science,* vol. 05, no. 03, 03/March-2023.

[4] Tesyon Korjo Hwase, Abdul Joseph Fofanah, "Machine Learning Model Approaches for Price Prediction in Coffee Market," *International Journal of Scientific Research in Science and Technology,* vol. 8, no. 6, pp. 10-48, 2021.

[5] P. Guru, J. Sathyapriya,K. V. R. Rajandran,J. Bhuvaneswari,C. Parimala, "Product Sales Forecasting and Prediction Using Machine Learning," *International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING,* pp. 355-366, 07/11/2023.

[6] Visakhapatnam, "DEMAND AND SALES FORECASTING USING MACHINE LEARNING," *Journal of Engineering Sciences ,* vol. 13, no. 09, pp. 408-415, 2022.

[7] Anitha, Ejjivarapu,Nagaraju,Geetha, "SALES PREDICTION USING MACHINE LEARNING TECHNIQUES," *International Research Journal of Modernization in Engineering Technology and Science,* vol. 05, no. 07, pp. 1089-1093, July-2023.

[8] Oryza Wisesa,Andi Adriansyah and Osamah Ibrahim Khalaf, "Prediction Analysis for Business To Business (B2B) Sales of Telecommunication Services using Machine Learning," *Majlesi Journal of Electrical Engineering,* vol. 14, no. 145-153, December 2020.

[9] Soham Patangia,Kevin Shah,Madhura Mokashi,Rachana Mohite, Gaurav Kolhe,Prajakta Rokade , "Sales Prediction of Market using Machine Learning," *International Journal of Engineering Research & Technology (IJERT),* vol. 9, no. 09, pp. 708-712, September-2020.

[10] Rao Faizan Ali,Amgad Muneer,Ahmed Almaghthawi, Amal Alghamdi and , Suliman Mohamed Fati, "big mart sales prediction using different machine learning techniques," *International Journal of Artificial Intelligence,* vol. 12, no. 2, pp. 874-883, June 2023.

[11] B. BETEMARIAM, "QUALITY IMPROVEMENT USING SPC TOOLS IN MOBILE," ST. MARY'S UNIVERSITY, Addis Ababa, June 2021.

[12] K. Alireza, "Using Machine Learning and Big data in Sales Forecasting for Production," ostfold University College, 2020.

[13] Prajwal Amrutkar, Shubhangi Mahadik, "Sales Prediction Using Machine Learning Techniques," *International Journal of Research Publication and Reviews,* vol. 3, no. 8, pp. 1887-1890, August 2022.

[14] O. Kenneth, "A Comparative Analysis of Four Machine Learning Algorithms to Predict," Department of Data Analytics Dublin Business School, Dublin, 2022.

[15] Saira Malik, Muhibullah Khan,Muhammad Kamran Abid and Naeem Aslam, "Sales Forecasting Using Machine Learning Algorithm in the Retail Sector," *Journal of Computing & Biomedical Informatics,* vol. 06, no. 02, 2024.

[16] Bhardwaj, Kumar , "Rise of Data Mining: Current and Future Application Areas," *International Journal of Computer Science,* vol. 5, no. 8, pp. 256-260, 2011.

[17] "ZenTut," Data Mining Techniques. Data mining, [Online]. Available: http://www.zentut.com/data-mining. [Accessed 11 February 2024].

[18] S. Mark, "A Survey of Machine Learning Algorithms and Their Application," San Jose State University, San Jose, California, 2018.

[19] V. BUYAR, "A FRAMEWORK FOR MODELING SALES PREDICTION USING BIG DATA," Southern Connecticut State University, New Haven, Connecticut, May 2019.

[20] Mummidi, Sai Nikhil Boyapat ,Ramesh, "Predicting sales using Machine Learning Techniques," Blekinge Institute of Technology, Karlskrona, Sweden, 2020.

[21] Farquad, Khalid, "Comparative Analysis of Support Vector Machine," in *14th International Conference on Information Technology*, 2015.

[22] Oja, Jorma Laaksonen , Erkki, "Classification with Learning k-Nearest Neighbors," Helsinki University of Technology Laboratory of Computer and Information Science, Helsinki, 1996 IEEE..

[23] Jiangtao Ren, Sau Dan Lee,Xianlu Chen, "Naive Bayes Classification of Uncertain Data," in *Ninth IEEE International Conference on Data Mining*, 2009.

[24] Li, Yuguang Huang ,Lei, "Naïve Bayes Classification Algorithm Based on Small Sample Set," in *Beijing University of Posts and Telecommunications, Proceedings*, Beijing, 2011.

[25] R. Saxena, "Introduction to k Nearest Neighbor Classification and Condensed Nearest Neighbour Data Reduction," *Learning journal monthly blog,* December 23, 2016.

[26] Jie Yang, d Sakgasit Ramingwong, "Analysis of Sales Influencing Factors and Prediction of Sales in Supermarket based on Machine Learning," *Data Science and Engineering (DSE) Record,* vol. 3, no. 1, pp. 67-77.

[27] P. Schuller, "A machine learning approach to promotional sales forecasting," Mestrado Integrado em Engenharia e Gestão Industrial, 2018.

[28] Bresam, Hussam Mezher Merdas, "FOOD SALES PREDICTION USING MACHINE LEARNING TECHNIQUES," Council of the College of Computer Science & Information, 2023.

[29] S. Maharaj, "www.analyticsvidhya.com/blog/2021/05," Machine Learning Model Evaluation, 2021. [Online].

[30] Sunita, Yashaswi,Smita Chavan, "Commerce & Management predictive model Insurance Industry," *National Monthly Refereed Journal of Research,* 2013.

[31] Renukaprasad, Vignesh Bengaluru, "Decision Support System for a Restaurant to Forecast Sales using Machine Learning Techniques," Dublin Business School, 2020.

[32] Cheriyan , Ibrahim ,Mohanan and Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," in *Comput. Electron. Commun. Eng*, 2018.

[33] Arif , Sany ,Nahin , Shahariar and Rabby, "Comparison Study :Product Demand Forecasting with Machine Learning for Shop," 2019.

[34] Chand N, Adhikari ,Garg ,Datt ,Das ,Deshpande and Misra , "Ensemble methodology for demand forecasting," 2017.

[35] Kilimci , Akyuz , Uysal , Akyokus , Uysal , Atak Bulbul , Ekmis and Silva , "An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain Complexity," 2019.

[36] D. E. University, "COMPARATIVE STUDY ON RETAIL SALES FORECASTING BETWEEN SINGLE AND COMBINATION METHODS," Dokuz Eylul University , 2017.

[37] Gallagher, Madden and Arcy, "A Bayesian classification approach to improving performance for a real-world sales forecasting application," in *IEEE*, 2015.

[38] MUHAMMAD ADNAN,SHAZIA SAQIB,TAHIR ALYAS and YOUSAF SAEED, "Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning," *IEEE,* vol. 8, pp. 13-23, July 2, 2020.

[39] M. Mary, "Overview of Experimental Research," Center for INOVATION IN RESERCH ON TEACHING, April 2019.

[40] B. Will, 6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples), 5 Jan 2019. [Online]. Available: https://towardsdatascience.com/. [Accessed 7 May 2024].

[41] Pavlyshenko, Bohdan , "Machine-Learning Models for Sales Time Series Forecasting," 2019.

[42] Kedar Potdar, Taher and Chinmay, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *International Journal of Computer Applications,* vol. 175.4, pp. 7-9, 2017.

[43] Smith and Draper , Applied Regression Analysis, John Wiley & Sons, 2014.

[44] Sebastian and Raschka , "Predictive modeling, supervised machine learning, and pattern," IEEE, 2014.

[45] AidaKrichene, "Using a naive Bayesian classifier methodology for loan risk assessment," *Journal of Economics,Finance and Administrative Science,* vol. 22, no. 42, pp. 3-24, 2017.

[46] Satish, Ram Babu and Rama, "Improved of K-Nearest Neighbor Techniques in Credit Scoring," *International Journal for Development of Computer Science & Technology,* vol. 1, no. 2, March 2013.

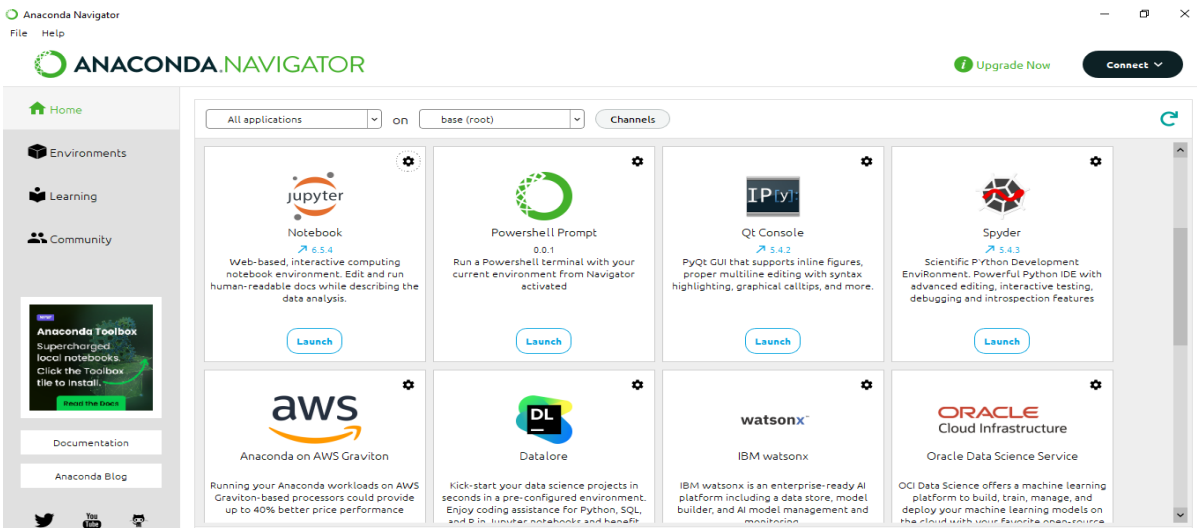
[47] Jesper and Groot , "Credit risk modeling using a weighted support vector machine," Utrecht University, Master Thesis, September 23, 2016.

Appendix

Appendix 1-A: The snapshot Transsion Manufacturing original data set Dataset

BRAND	MODEL	COLOR	PHONE TYPE	MARKET TYPE	QTY
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60
ITEL	it 2180	Dark Blue	FEATURE	LOCAL	60

Appendix 1-B: The snapshot Install ANACONDA.NAVIGATOR



Appendix 1-C: The snapshot Install Necessary package from Anaconda library

Appendix 1-D: The snapshot Install basic necessary packages

```
!pip install pandas
!pip install numpy
!pip install matplotlib
!pip install scikit-learn
```

Appendix 1-E: The snapshot Import necessary libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
```

Appendix 1-F: The snapshot Read CSV dataset

```
df=phone_Sales=pd.read_csv("Salse Dataset mobile Transsion Manufacturing.csv")
```

Appendix 1-G: The snapshot Remove Missing value

```
cat_imputer=SimpleImputer(strategy="constant",fill_value="missing")
num_imputer=SimpleImputer(strategy="mean")
cat_feature=["BRAND","MODEL","COLOR","MARKET TYPE","PHONE TYPE"]
num_feature=["QTY"]
```

```
imputer=ColumnTransformer([("cat_imputer",cat_imputer,cat_feature),
                           ("num_imputer",num_imputer,num_feature)])
filled_df=imputer.fit_transform(df)
filled_df
```

```
array([['ITEL', 'it 2180 ', 'Dark Blue', 'LOCAL', 'FEATURE', 60.0],
       ['ITEL', 'it 2180 ', 'Dark Blue', 'LOCAL', 'FEATURE', 60.0],
       ['ITEL', 'it 2180 ', 'Dark Blue', 'LOCAL', 'FEATURE', 60.0],
       ...,
       ['ITEL', 'IT2160', '               Dark Blue', 'LOCAL',
        'FEATURE', 80.0],
       ['ITEL', 'IT2160', '               Dark Blue', 'LOCAL',
        'FEATURE', 80.0],
       ['ITEL', 'IT2160', '               Dark Blue', 'LOCAL',
        'FEATURE', 80.0]], dtype=object)
```

Appendix 1-H: The snapshot Splitting a dataset

```
x=phone_sales_filled.drop("BRAND",axis=1)
y=phone_sales_filled["BRAND"]
```

Appendix 1-I: The snapshot Transform Categorical feature

```
<106989x513 sparse matrix of type '<class 'numpy.float64'>'
        with 534945 stored elements in Compressed Sparse Row format>
```

```
categorical_features=["PHONE TYPE","MODEL","COLOR","MARKET TYPE"]
one_hot=OneHotEncoder()
transformer=ColumnTransformer([("one_hot",
                                one_hot,
                                categorical_features)],
                                remainder="passthrough")

transformed_x=transformer.fit_transform(phone_sales_filledx)
transformed_x
```

```
categori_features=["BRAND"]
one_hot=OneHotEncoder()
transformer=ColumnTransformer([("one_hot",
                                one_hot,
                                categori_features)],
                                remainder="passthrough")

transformed_y=transformer.fit_transform(yphone_sales_filled)
transformed_y
```

```
array([[1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       ...,
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.]])
```

Appendix 1-J: The snapshot of Create Training and Test dataset.

```
from sklearn.model_selection import train_test_split
```

```
x_train,x_test, y_train,y_test=train_test_split(transformed_x,y, test_size=0.2)
```

Appendix 1-K: The snapshot of modeling using Random Forest

```
from sklearn.ensemble import RandomForestClassifier
np.random.seed(42)
```

```
clf=RandomForestClassifier(n_estimators=100)
clf.fit(x_train,y_train)
```

```
y_preds=clf.predict(x_test)

from sklearn.metrics import classification_report, confusion_matrix

print(classification_report(y_test,y_preds, zero_division=1))
print(confusion_matrix(y_test,y_preds))
```

```
              precision    recall  f1-score   support

        ITEL       1.00      0.99      1.00      9486
       TECNO       0.99      1.00      1.00     11912

    accuracy                           1.00     21398
   macro avg       1.00      1.00      1.00     21398
weighted avg       1.00      1.00      1.00     21398

[[ 9405    81]
 [    0 11912]]
```

```
              precision    recall  f1-score   support

     FEATURE       0.97      0.96      0.97     10841
       SMART       0.96      0.97      0.97     10557

    accuracy                           0.97     21398
   macro avg       0.97      0.97      0.97     21398
weighted avg       0.97      0.97      0.97     21398

[[10453   388]
 [  291 10266]]
```

Appendix 1-L: The snapshot of modeling using KNN

```
from sklearn.neighbors import KNeighborsClassifier
```

```
y_preds=clf.predict(x_test)

from sklearn.metrics import classification_report, confusion_matrix

print(classification_report(y_test,y_preds, zero_division=1))
print(confusion_matrix(y_test,y_preds))
```

```
              precision    recall  f1-score   support

        ITEL       1.00      0.99      0.99      9316
       TECNO       0.99      1.00      0.99     12082

    accuracy                           0.99     21398
   macro avg       0.99      0.99      0.99     21398
weighted avg       0.99      0.99      0.99     21398

[[ 9202   114]
 [    8 12074]]
```

```
              precision    recall  f1-score   support

     FEATURE       0.96      0.96      0.96     10841
       SMART       0.96      0.96      0.96     10557

    accuracy                           0.96     21398
   macro avg       0.96      0.96      0.96     21398
weighted avg       0.96      0.96      0.96     21398

[[10402   439]
 [  382 10175]]
```

Appendix 1-M: The snapshot of modeling using Naïve Bayes

```
from sklearn.naive_bayes import BernoulliNB
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import MultinomialNB
```

```
y_preds=clf.predict(x_test)

from sklearn.metrics import classification_report, confusion_matrix

print(classification_report(y_test,y_preds, zero_division=1))
print(confusion_matrix(y_test,y_preds))
```

```
              precision    recall  f1-score   support

        ITEL       1.00      0.98      0.99      9316
       TECNO       0.99      1.00      0.99     12082

    accuracy                           0.99     21398
   macro avg       0.99      0.99      0.99     21398
weighted avg       0.99      0.99      0.99     21398

[[ 9157   159]
 [   33 12049]]
```

```
y_preds=clf.predict(x_test)

from sklearn.metrics import classification_report, confusion_matrix

print(classification_report(y_test,y_preds, zero_division=1))
print(confusion_matrix(y_test,y_preds))
```

```
              precision    recall  f1-score   support

     FEATURE       0.97      0.94      0.96     10841
       SMART       0.94      0.97      0.96     10557

    accuracy                           0.96     21398
   macro avg       0.96      0.96      0.96     21398
weighted avg       0.96      0.96      0.96     21398

[[10234   607]
 [  327 10230]]
```

Appendix 1-N: The snapshot of modeling using SVM

```
from sklearn import svm
```

```
clf=classifier= svm.SVC(kernel='linear')
clf.fit(x_train,y_train)
```

```
SVC(kernel='linear')
```

```
y_preds=clf.predict(x_test)

from sklearn.metrics import classification_report, confusion_matrix

print(classification_report(y_test,y_preds, zero_division=1))
print(confusion_matrix(y_test,y_preds))
```

```
              precision    recall  f1-score   support

        ITEL       1.00      0.98      0.99      9316
       TECNO       0.99      1.00      0.99     12082

    accuracy                           0.99     21398
   macro avg       0.99      0.99      0.99     21398
weighted avg       0.99      0.99      0.99     21398

[[ 9157   159]
 [   33 12049]]
```

```
y_preds=clf.predict(x_test)

from sklearn.metrics import classification_report, confusion_matrix

print(classification_report(y_test,y_preds, zero_division=1))
print(confusion_matrix(y_test,y_preds))
```

```
              precision    recall  f1-score   support

     FEATURE       0.97      0.94      0.96     10841
       SMART       0.94      0.97      0.96     10557

    accuracy                           0.96     21398
   macro avg       0.96      0.96      0.96     21398
weighted avg       0.96      0.96      0.96     21398

[[10234   607]
 [  327 10230]]
```