B&W LogoSEC7001

Data Analytics and Technologies

Master’s Project

Prepared by Bandana Adhikari M.Sc. (Hons)

**Utilising Data Mining Techniques to Evaluate Customer Buying**

**Patterns of the Retail Industry**

School of Arts and Creative Technologies

Supervisor: Pradeep Hewage

*A report submitted in the partial fulfilment for the MSc degree in Data Analytics and Technologies-extended*

University Of Bolton April 2023

**Acknowledgements**

For this project, I have put so much effort in bringing it together. However, it would not have been possible without the help of my supervisor and module tutor. Therefore, I would like to thank all of them sincerely.

I am highly obliged to my supervisor, Mr Pradeep Hewage, for his precious guidance and support throughout the project. His Supervision was so important for me to give life to this project as he was showing me the direction from the start. Likewise, I am thankful towards my module tutor Mr Thaier Hamid, for his immense support and guidance who created a friendly environment for me. The mentioned members have contributed a lot in guiding me throughout this project. I present my sincere gratitude towards them.

Bandana Adhikari (Student Id. 2118001)

**Abstract**

In this competitive era, every business, supermarket, or company is trying hard to survive and gain profitable market values from their products. From that understanding, customer buying patterns and customer satisfaction is an essential factor in the business sector because customers are the ones who use the services and products. Every company with great marketing strategies can improve user experiences, gain loyal customers, and increase sales revenue. Analysing the exact accurate buying pattern helps entrepreneurs make marketing plans and strategies, understanding customers and their purchasing habits, such as what they will buy next. To obtain that customer purchased items or their requirements should be gathered from various sources, then different data mining algorithms should be implemented on those data. This study aims to analyse customer buying patterns using deep learning and machine learning algorithm. The experiment is conducted on the historical transaction data of the grocery store where Deep learning and Machine learning models were built using different algorithms like Logistic Regression, Decision Tree, Random Forest, and Long Short-Term Memory. Different evaluation metrics are used to evaluate the performance of each model. The experiment result illustrates how the customer will purchase the item and predict which items will likely be bought by customers. The data analysis on the transaction’s dataset enables an understanding of which items are sold out most in a timeframe. Overall, the project analysed what products are likely to be purchased and what items a customer will purchase in the future.

**Keywords- Machine learning, deep Learning, Analysis, Long Short-Term Memory (LSTM), Random Forest Algorithm, Decision Tree, Logistic Regression**

**Contents**

[Chapter 1- Introduction 1](#_Toc133787094)

[1.1 Introduction 1](#_Toc133787095)

[1.2 Background 1](#_Toc133787096)

[1.3 Statement of the Problem 3](#_Toc133787097)

[1.4 Aim/s-Objectives or Purpose of Study 3](#_Toc133787098)

[1.5 Research Questions and /or Hypotheses 4](#_Toc133787099)

[1.6 Significance of Study 4](#_Toc133787100)

[1.7 Scope and Limitation of Study 5](#_Toc133787101)

[1.8 Thesis Organisation 5](#_Toc133787102)

[Chapter 2- Literature Review 7](#_Toc133787103)

[2.1 Introduction 7](#_Toc133787104)

[2.2 Literature Review References 7](#_Toc133787105)

[2.3 Review of Related Works Using Data Mining Techniques 1](#_Toc133787106)

[2.4 Deep Learning and Machine Learning 3](#_Toc133787107)

[2.4 Conclusion 5](#_Toc133787108)

[Chapter 3 - Methodology 6](#_Toc133787109)

[3.1 Working Mechanism of Selected Algorithm 7](#_Toc133787110)

[3.1.1 Random Forest 7](#_Toc133787111)

[3.1.2 Logistic Regression 8](#_Toc133787112)

[3.1.3 Decision Tree 9](#_Toc133787113)

[3.1.4 LSTM 10](#_Toc133787114)

[3.2 Tools Used 11](#_Toc133787115)

[3.3 Data Collection 11](#_Toc133787116)

[3.4 Data Pre-processing 12](#_Toc133787117)

[3.4.1 Data Cleaning 12](#_Toc133787118)

[3.5 Model Train and Evaluation 13](#_Toc133787119)

[3.6 Ethical Consideration 14](#_Toc133787120)

[3.7 Conclusion 14](#_Toc133787121)

[Chapter 4 - Implementation 15](#_Toc133787122)

[4.1 Introduction 15](#_Toc133787123)

[4.2 Libraries Used for Deep Learning 15](#_Toc133787124)

[4.3 Libraries Used for Machine Learning 16](#_Toc133787125)

[4.4 Analysis of Data 17](#_Toc133787126)

[4.5 Data Pre-processing for Machine Learning 21](#_Toc133787127)

[4.5.1 Data Cleaning 21](#_Toc133787128)

[4.6 Data Pre-processing for Deep Learning 24](#_Toc133787129)

[4.6.1 Data Cleaning 24](#_Toc133787130)

[4.7 Implementation of Selected Algorithm 26](#_Toc133787131)

[4.7.1 Logistic Regression and Result 26](#_Toc133787132)

[4.7.2 Decision Tree and Result 27](#_Toc133787133)

[4.7.3 Random Forest Algorithm and Result 27](#_Toc133787134)

[4.7.4 LSTM and Result 28](#_Toc133787135)

[4.8 Result evaluation of three different Machine learning Models 30](#_Toc133787136)

[4.8.1 accuracy measurement of LR 30](#_Toc133787137)

[4.8.2 accuracy measurement of DT 30](#_Toc133787138)

[4.8.3 accuracy measurement of RF 31](#_Toc133787139)

[4.9 Comparison between Models 33](#_Toc133787140)

[Chapter 5 - Limitation and Future Work 35](#_Toc133787141)

[5.1 Limitation 35](#_Toc133787142)

[5.2 Future Work 35](#_Toc133787143)

[Chapter 6- Conclusion 37](#_Toc133787144)

[References 40](#_Toc133787145)

[Bibliography 45](#_Toc133787146)

[Appendix-A 46](#_Toc133787147)

List of Tables

Table 2.1: References used in Literature Review…………………………………8

Table 4.1: Model Accuracy before use of different methods to increase accuracy……………………………………………………………………………….48

Table 4.2: Model Accuracy after the Use of different approaches improve accuracy……………………………………………………………………………….49

Table 4.3: Model Accuracy of DL and ML…………………………….……………50

Table 6.1: Research Question Answers……………………………………………53

List of Figures

Figure 1.1: Working Mechanism of Data Mining Technique 2

Figure 3.1:Flowchart showing overall process 22

Figure 3.2: Working Mechanism of Random Forest 23

Figure 3.3: Working Mechanism of Decision Tree 25

Figure 3.4: Working mechanism of LSTM 26

Figure 4.1: Library used for LSTM 31

Figure 4.2: Library used for Machine Learning 32

Figure 4.3: Visualisation of the number of items repeated 34

Figure 4.4: Count the number of items purchased by a member 34

Figure 4.5: Most sold-out product in the month of April 35

Figure 4.6: Displays in which month the customer buys the items 36

Figure 4.7: Display the items purchased in a month 37

Figure 4.8: Converting the dataset into variable length data 38

Figure 4.9: Null value check-in dataset 38

Figure 4.10: Duplicate value check in the dataset 38

Figure 4.11: Choosing the top5 most purchased items 38

Figure 4.12: Categorical value in numerical form 39

Figure 4.13: Checking the length of the label and prediction 40

Figure 4.14: Tokenization result 41

Figure 4.15: Before and After Padding Result 41

Figure 4.16: Predicted Result from Logistic Regression 42

Figure 4.17: Predicted Result from Decision Tree 43

Figure 4.18: Predicted Result from Random Forest Algorithm 44

Figure 4.19: Model Compilation 44

Figure 4.20: Decrease in a loss in the course of training 45

Figure 4.21: Predicted items of LSTM model 45

Figure 4.22: Accuracy before and after hyperparameters of DT 47

Figure 4.23: Accuracy before and after hyperparameters of RF 48

Figure A1: Dataset 62

Figure A2:Data Pre-processing for Machine Learning 62

Figure A3:Logistic Regression Model 64

Figure A4: Accuracy measurement of Logistic Regression 64

Figure A5: Decision Tree Model 65

Figure A6: Accuracy measurement of Decision Tree 65

Figure A7: Hyperparameters tuning and accuracy measure 66

Figure A8: Random Forest Model 67

Figure A9: Accuracy measurement of RF 67

Figure A10: Ensemble Method and Accuracy measure 68

Figure A11: Data Cleaning for Deep Learning 69

Figure A12: LSTM model fit and compilation 70

Figure A13: Prediction of new items 71

Figure A14: Gantt Chart 71

List of abbreviations

DT Decision Tree

LR Logistic Regression

RF Random Forest

LSTM Long Short-Term Memory

TP True Positive

FP False Positive

FN False Negative

SML Supervised Machine learning

CSV Comma-Separated Values

IoT Internet of Thing

# Chapter 1- Introduction

## Introduction

With the evolution and competition in the retail industry, retailers should have a good and deep understanding of customer buying patterns and preferences. The analysis of customer buying patterns can gain valuable insight, this is possible with the use of different data mining techniques. Data mining is extracting valuable insights from large datasets using machine learning and statistical algorithm (Raorane & Kulkarni, 2011). In this report, I use data mining techniques to evaluate customer buying patterns in the retail industry using transaction data. Various data mining techniques are implemented in the retail dataset and which is fitted one in for this dataset. Furthermore, this report includes the methodology of these techniques and implementation.

## Background

Different analysing techniques have been implemented with the changing marketing strategies and customer buying behaviour. The study of customer buying behaviour has been one of the famous research areas in marketing. In the early 20th century, retail pioneers used various methods such as, storing sales records conducting different surveys and observing customers so that valuable insights into customer behaviour could be gained from it (Tang et al., 2022). These steps have faced several challenges, including the quality of data, privacy concerns, lack of standardisation and data availability.

With the development in computer technology and the increase in large-scale retail datasets, data mining techniques like Market Basket Analysis, frequency analysis and cross-tabulation were developed. These data mining techniques helped to solve the issue of traditional techniques used in analysing customer buying patterns. Furthermore, the rise in cloud computing and data mining strategies has made it possible to process large-scale real-time data. Data mining is analysing and retrieving meaningful information from big and complex data. Mining from a large group of data is a complicated process, there are different techniques and algorithms which should be implemented according to the required future outcomes. Many projects had been carried out previously where Ilung Pranatac and Geoff Skinner analysed the buying behaviour of Portuguese wholesalers using six different products with the implementation of the K-Mean algorithm. By the annual expenditure of $90,000, $50,000 and $20,000, they divided the customers into clusters. The project was focused on the selection of the best fresh food and they concluded that all three-customer buying pattern was the same (Ponyiam and Arch-int, 2018).

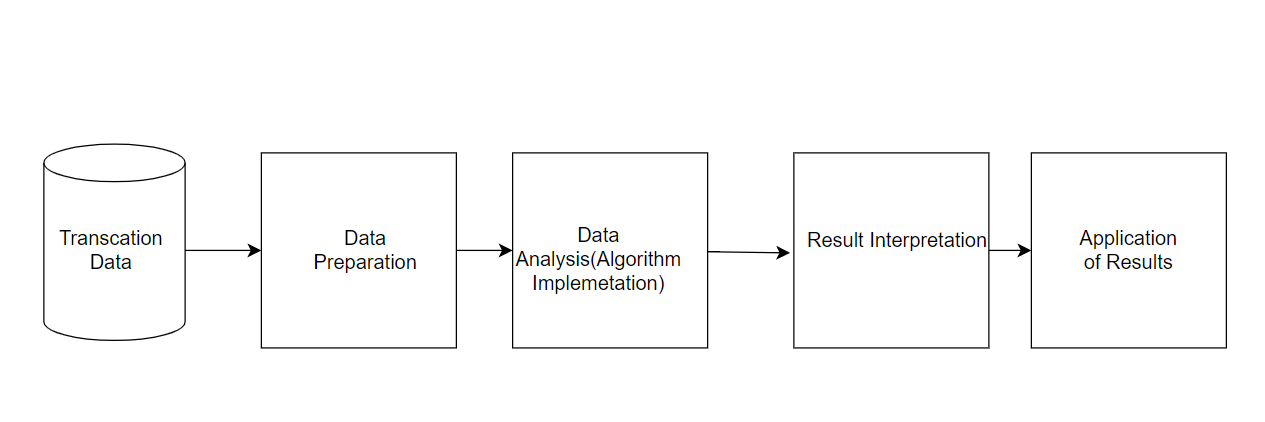


Figure 1.1: Working Mechanism of Data Mining Technique

Data mining technique plays an important role in analysing customer buying patterns for better decision-making, to find valuable insight from uncovering patterns. The working mechanism of the technique is shown in above figure 1.1 where the first stage is the collection of relevant customer data such as transaction data, and behavioural data collected from different sources. Secondly, data is prepared and cleaned in a suitable format for further analysis. Thirdly different mining algorithms are implemented to uncover patterns and relationships in those data. The result from the analysis is presented in visual formats like tables and graphs.

Mining is processing statics or machine learning to find hidden knowledge. It is achieved by creating a model of supervised or unsupervised learning. It can be achieved by supervised machine learning, which requires label data to learn the relationship between the output variable and the input feature. Also, label data is used for data training for the better performance. There are two different supervised learning which is classification and regression. Classification is a type of supervised learning where output is categorical, and for regression, output variable is continuous.

Another is unsupervised machine learning, where labelled data are absent is called Unsupervised Learning. Here model attempts to find the pattern without labelled output guidance which requires highly sophisticated data analysis (Ponyiam and Arch-int, 2018). It includes a clustering algorithm where models group the data based on their attributes or features. The similarity is calculated by using Euclidean distance or any other distance metrics. There is no specific output variable for unsupervised learning as the output is a set of unexpected data.

## Statement of the Problem

Customer buying patterns have become one of the most integral parts of the Retail industry. Despite investigating different resources in this industry, they cannot fulfil the customer demand (Ponyiam & Arch-int, 2018). There should be a deeper understanding of customer preferences and demand to address this problem, we can get valuable insight from data which include items, customer IDs, prices, and so on, which improve customer experience. So, this paper intends to implement different data mining techniques to analyse customer buying patterns in the retail industry.

## Aim/s-Objectives or Purpose of Study

Due to changing demand of customers, the global market is facing difficulties in fulfilling their customer requirements. So, this project aims to analyse customer buying patterns to help determine which products are most likely to be purchased so that businesses can understand and focus on specific products to increase the chances of upselling and cross-selling.

Following are the objectives which are considered to reach the aim addressed above which are: -

* To determine the importance of analysing customer buying nature
* To understand customer needs and preferences every month
* To determine which algorithm is suitable for this dataset.
* To predict the new products according to the history of customer buying behaviour.

## Research Questions and /or Hypotheses

This study shows the related question for this project. So, the following are the question to be solved.

* Why is it important to analyse customer buying patterns?
* What are the customer preferences and needs in a month?
* Which is the suitable algorithm for this dataset?
* What are the suggestion and recommendations for the products based on the results?

## Significance of Study

From Kaggle, the groceries dataset, an existing dataset, is used in this project. From this most common and frequently purchased items can be identified. The dataset contains member number, date and item description. This dataset is analysed using Excel.

Due to the changing demand of customers, it is very challenging for any supermarket to fulfil their requirements. For that, any organisation should collect daily, weekly or monthly transaction data to formulate different mining techniques. So, in this project, analyses of historical transaction data are carried out at a grocery store which helps predict items' future predictions. The significance of this study is to help an organisation analyse which items are purchased frequently together and which items should be kept together.

All items in a retail store can be related to each other, where they might be weak or strong. In some cases, one sale of one item can relate to the sale of another item. So, this technique should be implemented in a retail store or supermarket. Because of these, retailers can improve marketing strategies and know effective product placement, cross-selling and up-selling strategies.

## Scope and Limitation of Study

Every day tremendous amount of transaction data is generated from retail stores. So, analysing large-scale datasets is quite challenging. In this research, deep learning and machine learning algorithms are implemented, which are used in the large dataset to better understand buying behaviour. Mainly, this project’s scope will help in analysing what products a customer will purchase in future and how the items will be purchased. This will help to identify consumer likes and dislikes, which allows marketers to improve marketing strategies and plans. This project focuses on providing customers with new items and help retailers to make profitable markets. This allows customers to be become acquainted with new products, which improves their quality of life.

Because the increase in production of items is increasing rapidly with that, customer requirements are changing. For that, this research work might not be practical as this research uses historical data, which might need to be more to predict customer buying patterns. It works only on transaction data and finding the exact datasets is quite challenging.

## Thesis Organisation

This report is documented according to the following chapter: -

The second chapter is a literature review that looks at the past works implemented by different authors and researchers in analysing and predicting customer buying patterns using various machine learning and deep learning algorithms.

The third chapter includes the machine learning and deep learning algorithm, the ones who are in this project which are Random Forest, Decision Tree, Logistic Regression and LSTM algorithm. Furthermore, it looks at how the dataset is collected from where and what are the data pre-processing steps such as data cleaning, encoding and so on.

Chapter four is the implementation part of the thesis, which discusses the library used to build the model, data analysis, and comparison of results of each model is carried out. It includes the tools used, such as Google Colab Environment, Python programming language, Hive and PowerBI. Also, the evaluation of the machine learning algorithm is determined using different evaluation metrics.

Finally, chapter five provides the overall summary of the implementation of the project. It includes and recommends the areas for improvement so there will be a better analysis of customer buying patterns in future work. Also, it points out the limitation related to the project, which can be further improved in future. And finally provides the summary of the project along with the complete concept of how the project works.

# Chapter 2- Literature Review

## 2.1 Introduction

This chapter will provide the existing works carried out by other researchers in the field of mining techniques, customer behaviour analysis and uses of the different kinds of algorithms. Here the different technique is highlighted, which was already implemented by other authors to understand customer buying behaviour. Various machine learning was implemented, such as KNN, Random Forest, Clustering and Apriori algorithm, to predict and analyse the customer buying patterns. In the 1950s, researchers from the marketing field began to focus on customer buying patterns which became the central focus. The importance of understanding customer buying patterns continues to grow due to the changing demands for products, so retailers and marketers need to use different data analytics tools.

The rise of internet technology created opportunities for marketers to store customer behaviour online, and offline, use different analytics tools, and gain helpful sight from it which help to improve customer experience and develop targeted strategies. This term helps many retailers and marketers to enhance their business. With the advancement some changes were faced like data quality, data privacy and data integration. Keeping up with the changing behaviour of customers was very changeling for a retailer. For this, they must keep an eye on and monitor every dataset generated and analysed, But this can be overcome with the use of advanced data analytics tools.

Overall, this chapter discusses what kind of algorithm, what kind of dataset, how the algorithm was implemented, and the result obtained. In addition, this chapter discusses each paper's limitations and includes recommendations for future improvement.

## 2.2 Literature Review References

Table 2.1 References Used in Literature Review

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Reference*** | ***Relevance*** | ***Methodology*** | ***Strengths*** | ***Limitations*** | ***Contribution*** |
| Ayu, S.K., Surjandari, I. and Zulkarnain, Z. (2018) “Mining Association rules in Seasonal transaction data,” 2018 5th International Conference on Information Science and Control Engineering (ICISCE) [Preprint]. | 4 | Experimental | Focuses on analysing sale according to the type of season and proposed a new algorithm to handle seasonal data. | Did not compare performance of the algorithm used with other algorithms. | Provides the patterns of sales for each season which shows that this algorithm is effective for seasonal transaction data. |
| Bhaskar Velpula, V., Pakanati, S. and M, S. (2010) “Analyzing target customer behaviour by mining the E-COMMERCE data,” International Journal on Information Sciences and Computing, 4(1), pp. 27–30. | 4 | Experimental | Highlights the critical factors which affect the customer behaviour. | Not mentioned future work. | Able to find which gender customer spends more time on e-commerce sites. |
| Ding-An Chiang et al. (2009) “The cyclic model analysis on sequential patterns,” IEEE Transactions on Knowledge and Data Engineering, 21(11), pp. 1617–1628. | 4 | Experimental | Highlights the future work with what kind of technique to be used. |  | Propose a new model to identify and analyse complex systems.  Highlights customer buying nature being based on the time interval Use different real-world dataset. |
| Diwandari, S. and Zaky, U. (2021) “Analysis of Customer Purchase Behavior Using Association rules in e-shop,” 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) [Preprint]. | 4 | Experimental | Detail use of attributes for analysis and methodology, which is used in association rule mining. | Only focuses on only one factor and did not account for other factors that influence customer behaviour. | Demonstrates the effectiveness of the proposed approach in real-world dataset.  Highlights three high-level association rules from each algorithm, |
| Dlamini, M.G. et al. (2015) “Extracting interesting patterns from e-commerce databases to ensure customer loyalty,” 2015 IEEE 12th International Conference on Networking, Sensing and Control [Preprint]. | 4 | Experimental | Focuses on data mining techniques to ensure customer loyalty. Mention the use of other another algorithm to compare the result. | The dataset used here is only focused on e-commerce. | Presents an immediate solution to improve the marketing strategies and highlights the importance of analysing e-commerce sequential pattern. |
| Gupta, A.K. and Gupta, C. (1970) Analyzing customer behaviour using data mining techniques: Optimizing relationships with customer: Semantic scholar, Management Insight. | 3 | Case study | Highlights about the result derived and this result can be used in different developing countries. | No detailed description about which data mining techniques are used. | Identifies the major problems of customer behaviour which impact the sales and production. |
| Isa, N., Neddy, S.K. and Mohamed, N. (2021) “Association rule mining using FP-growth algorithm to prevent Maverick buying,” 2021 IEEE 11th IEEE Symposium on Computer Applications &amp; Industrial Electronics (ISCAIE) [Preprint]. | 5 | Experimental | Provides detailed explanation of the proposed approach and real-world dataset with the use of the real-world dataset. | No comparison with another algorithm. | Identifies the association rule between items purchased by customers helped in preventing Maverick’s buying behaviour. Uses two different values for support and confidence to find the quality of discovered association rule. |
| Kaur, M. and Kang, S. (2016) “Market basket analysis: Identify the changing trends of Market Data Using Association Rule Mining,” Procedia Computer Science, 85, pp. 78–85. |  | Experimental | Helps in finding patterns in large about of data and provides a detailed explanation of the methodology. | Poorly presented report and does not consider external factors that affects retail store data. | Identifies backtracking in data and provides a better method to find outliers. |
| Khade, A.A. (2016) “Performing customer behaviour analysis using big data analytics,” Procedia Computer Science, 79, pp. 986–992. | 3 | Experimental | Highlights the importance of analysing customer buying behaviour in the field of marketing.  Mentioned the future work. | Not mentioned the source of the dataset and no detailed explanation of the methodology. | New powerful technologies were introduced instead of traditional decision trees algorithms. |
| Maheswari, K. and Priya, P.P. (2017) “Predicting customer behavior in online shopping using SVM Classifier,” 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS) [Preprint]. | 4 | Experimental | Gives plenty of information about the factors which affects the buying behaviour. | Not mentioned the source of data. | Able to find which age group were more attracted towards online shopping. |
| Moens, S., Aksehirli, E. and Goethals, B. (2013) “Frequent itemset mining for Big Data,” 2013 IEEE International Conference on Big Data [Preprint]. | 5 | Experimental | Detail work on comparing different algorithms and which one was most effective in all of them. | No elimination of duplicate data, which took more space for processing. | Developed a method that is new in Association rule mining. |
| Mujianto, A.H. et al. (2019) “Consumer Customs analysis using the Association rule and Apriori Algorithm for determining sales strategies in retail central,” E3S Web of Conferences, 125, p. 23003. | 3 | Experimental | Highlights the problems in sales such as unsold products, products that are not in demand.  Gives good insights about the algorithm used and the results obtained. | Extremely complex. | Use an average value of support and confidence, based on the result retail owners where recommended. |
| Nasti, S.J., Asgar, M. and Butt, M.A. (2017) “Analysis of customer behaviour using modern data mining techniques,” International Journal of Computer Sciences and Engineering, 5(12), pp. 64–66. | 4 | Experimental | Able to find out if the rule generated in this paper can be used by retailers to man-the-spot discount. | Methods involved in implementation is not mentioned properly. | Able to find the condition, age and gender of buying products. |
| Nithin, S.S. et al. (2022) “Retail demand forecasting using CNN-LSTM model,” 2022 International Conference on Electronics and Renewable Systems (ICEARS) [Preprint]. | 3 | Experimental | Highlights the use of different parameters which can be considered by deep learning. Provides visualisation of model predictions. |  | Able to find the best-fit model for the dataset used here. |
| Ponyiam, P. and Arch-int, S. (2018) “Customer behavior analysis using data mining techniques,” 2018 International Seminar on Application for Technology of Information and Communication [Preprint]. | 5 | Experimental | Highlights that there are weaknesses in creating association rules with a basic threshold of support and confidence. So, author suggests using a new method as it reveals association rule with high performance. | Does not provide a detailed explanation of specific data mining techniques used. | Provided solutions for a better understanding of customer behaviour. |
| Raorane, A. and Kulkarni, R.V. (2011) “Data Mining Techniques: A source for consumer behaviour analysis,” International Journal of Database Management Systems, 3(3), pp. 45–56. | 3 | Case Study | Provides recommendations on where this technique can be used. Discuss the usefulness of a data mining system in a business. | No mention of the future work. | Able to find there are certain buying habits of customer. |
| Schwenke, C., Vasyutynskyy, V. and Kabitzsch, K. (2010) “Simulation and analysis of buying behavior in supermarkets,” 2010 IEEE 15th Conference on Emerging Technologies &amp; Factory Automation (ETFA 2010) [Preprint]. | 4 | Experimental | Uses a dataset which includes customer demographics and product purchases which allows for in detailed analysis. | Data generated from simulation requires validation. | Highlights simulation structure to capture real time event data so that it can be used to test data mining algorithm. |
| Shrirame, V. et al. (2020) “Consumer behavior analytics using machine learning algorithms,” 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) [Preprint]. | 5 | Experimental | Detail structured explanation of methodology and visualization of results. | Focused on single algorithm for analysis. no comparison among different machine learning algorithm. | Identifies the factors that affects any customer buying behaviour,  Also, able to understand neutral review and potential of machine learning in business. |
| Singh, S.P. et al. (2018) “Data Mining: Consumer Behavior Analysis,” 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information &amp; Communication Technology (RTEICT) [Preprint]. | 4 | Experimental | Provides the detail explanation about the importance in analysing customer behaviour and gives the detail description of different data mining technique. | Not enough information about dataset. | Supports the knowledge discovered from different mining technique |
| Tai, Y., Sun, Z. and Yao, Z. (2021) “Content-based recommendation using machine learning,” 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP) [Preprint]. | 4 | Experimental | Mentioned about future work and many datasets to give more reliable result. | Short report. | Use different algorithm and predict three different categories.  Able to analyse the combination of algorithm can give good accuracy. |
| Valecha, H. et al. (2018) “Prediction of consumer behaviour using random forest algorithm,” 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) [Preprint]. | 4 | Experimental | Detailed work on the performance of different algorithm and shows which algorithm provides best result |  | Highlights different factors might affect customer buying behaviour based in different region. |
| Weijs, D. and Caron, E. (2022) “Customer journey analytics: A model for creating diagnostic insights with process mining,” Proceedings of the 17th International Conference on Software Technologies [Preprint]. | 5 | Experimental | Detail discussion and Suggestion to use different data mining technique. | Correctness of information depends on performance of association algorithm and the accuracy of clustering. Data used here was from limited time frame so it cannot suggest items in different time frame.  Difficult to bug the system. | Looks to improve proposed system with the implementation of different algorithm. This system proposed can generate a summarized result of association rule where marketing people can understand and check the least frequent itemsets and association |

## 2.3 Review of Related Works using Data Mining Techniques

In this paper, the author focused to solve the problem of whether the buying nature of customers was cyclic or not and the duration for items to be in sequential form. So, the author proposed a new set of algorithms which was a mixture of data mining skills collected from a company. This research can be improved by applying fuzzy techniques with some statistics tools whereas the CMA algorithm can be implemented (Ding-An Chiang et al., 2009)

(Liu et al., 2018) implemented LSTM recurrent neural network to build a prediction model and to analyse the historical data of stock. The experimental result showed the model can do a better prediction even with low accuracy. They mentioned more feature extraction for the training of the model as a future work.

In this paper, the author focuses to find the identify products which interest customers in electronic shopping where web mining was used. Using these different categories of customers were formed which helps people in business to make the right strategy. Doing this helps in the business growth and focuses on specific selling products. Data were collected from the previous online session and shopping logs. Apriori Algorithm and FP-Growth were implemented. In accordance with the author future work for this paper was comparing large groups of customers (Diwandari and Zaky, 2021).

(Valecha et al., 2018) this research paper goal was to examine the relationship between customer interest in buying certain products and the factors that affect customer buying behaviour which was an environmental, organizational, individual and interpersonal factor. A Random Forest algorithm was used for the prediction of customers buying a product. The dataset was collected from Kaggle where the datasets were divided into subcategories. Python was used as a programming language.

(Schwenke et al., 2010) in this research, the author developed a simulation of buying behaviour in supermarkets which can be used to investigate data mining methods. This paper describes how a retail store, and its customers can be modelled so that test data can be simulated. The simulation structure was two parts customers and products. The future work for this paper can be an extension of the model for more details and development of a data mining algorithm for real-world event.

(Sharma and Kumar Soni, 2020) focused on implementing Artificial Intelligence in different companies to classify potential buyers for companies’ services and buyers. In this paper, different algorithm was implemented like Support Vector Machine, Random Forest, Logistic Regression and K Nearest Neighbour. These algorithms then were fed onto the customer dataset. This paper claims that Random Forest Algorithm gives more accuracy with classification and categorical problems and K-Nearest Neighbour performance was least in comparison to four other algorithms. The data used in this paper was collected for one year through UCI Repository. The future work for this paper was to introduce a hybrid model using a machine learning algorithm like Neural Network.

(Dlamini et al., 2015), here the main aim of this paper was to analyse customer behaviour so that it will help an organization to conduct profitable business and ensure customer loyalty by using data mining techniques. The decision tree algorithm was used for the classification and the prediction. Data were collected from various e-commerce data warehouses which were further integrated by using Microsoft Excel. After the implementation of one of the dataset authors concluded that there was a need for the improvement of sales in products. The author concluded that if a product was online for more than 550 days then the item will not be bought by any customers. Also, the association rule indicates the number of times customers visited the online store and the products together.

(Valecha et al., 2018) this research employed a random forest algorithm to examine what parameters affect the buying pattern of customers. The author concluded that random forest performed well that any other machine learning algorithm with 94% of accuracy. The researcher mentioned that the behaviour of customers depends on individual, environmental and interpersonal factors.

(Zhang et al., 2022), the author implemented different deep learning models for the prediction of customer behaviour. This project focused on improving and exploring DNN models and proposed rDNN and KmDNN models. Then these three-model evaluations and comparisons were carried out using AUC and F values.

(Ramaswamy & DeClerck, 2018) explored Deep learning and Natural Language Processing for the analysis of contextual information. The author implemented an algorithm of the car review dataset and concluded deep learning aspect is most applicable to customer perception.

(Lico et al., 2021) implemented DBSCAN, K-Means, K-Medoids and Agglomerative algorithm to find the customer with similar buying patterns of annual transactions dataset of department stores in the year 2020. The author mentioned clustering provides a very effective technique in Customer Relationship Management. Therefore, they concluded that K-Medoids performed well in comparison to another algorithm.

## 2.4 Deep Learning and Machine Learning

For analysing the customer buying pattern there are different data mining strategies. As suggested ML and DL algorithms can be implemented to evaluate customer buying patterns. Deep learning and machine learning are powerful tools which are used for prediction. Furthermore, there is a brief description of these two tools as mentioned below.

Deep learning lately has evolved to be one of the best choices of science to learn about large real-world data. It is the subfield of artificial intelligence and machine learning which is based on an artificial neural network with multiple layers of interconnected nodes (Wason, 2018). Deep learning is used in various applications for e-commerce, image, and speech recognition. The main advantage is to handle high dimensions and complex input and be able to learn useful features for that data automatically. There are several types which are Feedforward Neural Networks, Convolutional Neural Networks, Autoencoders, Generative Adversarial Networks, Recurrent Neural Networks and Long Short-term Memory. There are different types of Deep learning techniques. But LSTM is the most suitable algorithm for the analysis of sequential data.

1. LSTM: Long Short-Term Memory is a type of recurrent neural network which is used in deep learning. This algorithm is useful in the processing and analysis of sequential data for prediction. It works by using memory cells and a system of gates which allows to learn patterns and long-term dependencies in sequential data. It is designed to work with sequential data and handles a sequence of variable length which makes it the perfect fit to analyse customer buying patterns as there are sequences of items purchased over time.

Machine Learning

With the evolution, different techniques and tools have been formed to accomplish different tasks more efficiently. So, with the creativity and creation of machines by humans, machines made life easier in which machine learning is one of the terms. In some situations where viewing the data cannot give the insights or pattern of those data then machine learning is applied. So, machine learning is a study which gives computer the ability to learn without explicit programming whose main purpose is to learn from the data (Mahesh, 2020). Machine learning algorithms is categorized into 4 types which are shown below:-

* Supervised learning
* Semi-supervised learning
* Unsupervised learning
* Reinforcement learning

There are different types of machine learning algorithms but here are the three selected algorithms: -

1.Random Forest algorithm is a supervised learning algorithm which helps in classification. This algorithm is used to handle the large number of input features and is robust to noisy data which is very important for the analysis of customer buying behaviour. Random forest is known for its good accuracy for predicting customer buying behaviour. One of the projects revealed that its accuracy is much better than KNN which is of 94% (Valecha et al., 2018) during the prediction of customer behaviour. Another study concluded that random forest performs best in comparison to other algorithms during the buying behaviour analysis of customers.

2.Decision Tree: A decision Tree is a classification technique which handles alphabetical, numeric, and nominal data which is very important in evaluating customer buying patterns (Somvanshi et al., 2016). It is a classification approach which uses a tree like model to classify the given data into different classes based on predefined attributes. It can be used for both supervised and unsupervised learning tasks which is suitable for both numerical and categorical data. This algorithm represents the relationship between the target and input variables using a hierarchy of branches and nodes. Each node of this algorithm represents a decision for a specific variable and the branches represent the outcomes which are possible for that decision. The algorithm splits the data recursively into subsets based on the values of variables until it reaches the final prediction. One of the research projects shows that a decision tree can predict most of the customers behaviour correctly.

3.Logistic Regression: Logistic Regression is a predictive analysis method which is used to find the relationship between the target variables and predictors. It uses a logistic or sigmoid function for the linear combination of input variables and their targeted weights. The output of this algorithm is a probability score between 0 and 1. It is mostly used in marketing, finance, health care and social sciences. A researcher revealed that the accuracy of this algorithm is about 91% while analysing the customer buying behaviour (Sharma et al., 2022).

## 2.4 Conclusion

Overall, different projects have mentioned different kinds of machine learning algorithms and deep learning on different datasets from online sites or collected from different retail stores, warehouses from this above study, analysing and understanding consumer buying behaviour plays an important role for any business to run. According to the dataset used different models predict different results for the customer in correspondence to their purchasing habit. But in this project LSTM, Random Forest, Decision Tree and Logistic Regression algorithm are most suitable in predicting items. Therefore, these are the selected algorithm for this project.

# Chapter 3 - Methodology

In research project methodology is a framework which includes different techniques which are, algorithms, used to conduct research, analyse and collect data to achieve a goal. Methodology varies depending on the research goal or question to be achieved or solved. As per section two Logistic Regression, Random Forest, long short-term Memory, and Decision tree are techniques for implementation.

Here, the research aims to analyse customer buying behaviour along with prediction. So, in this section, we are going to talk about the method that is going to be used to achieve the desired goal. The research methodology used in this project is qualitative and quantitative. These types of methodologies analyse both numerical and text data.

As a part of implementation section groceries data set was collected from Kaggle.

To achieve the aim for this project objective following are considered: -

* To determine the importance of analysing customer buying nature
* To understand customer needs and preferences every month
* To determine which algorithm is suitable for this dataset.
* To predict the new products according to the history of customer buying behaviour.

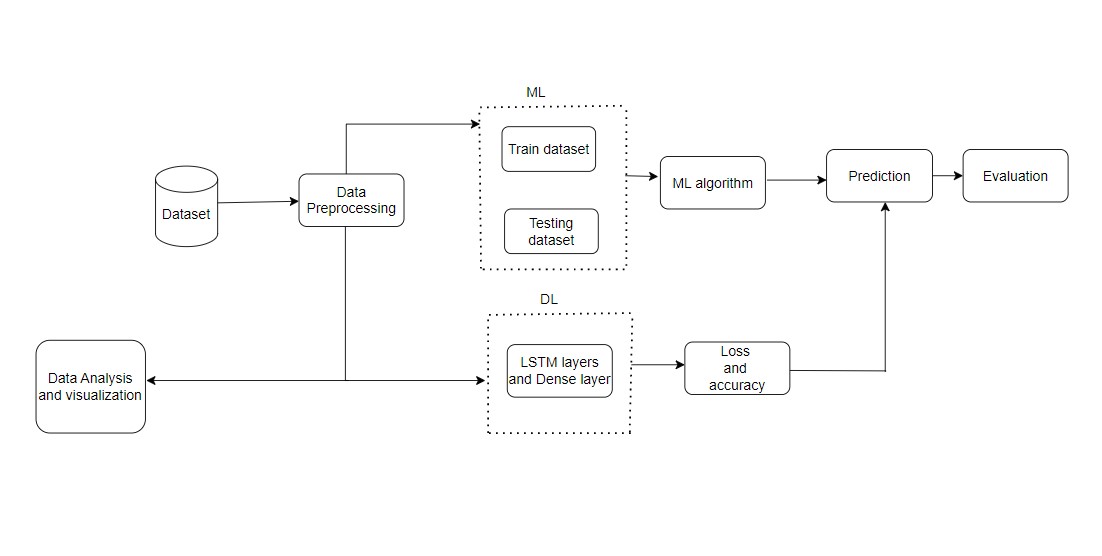


Figure 3.1: Flowchart showing overall process.

Figure 3.1 discusses about the overflow of the project and each part of the figure is discussed in detail in following section.

## 3.1 Working Mechanism of Selected Algorithm

This step involves the correct selection of algorithms which will fit my dataset. It is an important step in building a successful machine learning algorithm which depends on several factors such as data size, problem type to solve, and the resources available. From the above study Random Forest, LSTM, Decision Tree and Logistic Regression are suitable algorithms for the evaluation or analysis of customer buying pattern.

### 3.1.1 Random Forest

It is a popular supervised machine learning algorithm based on decision tree used for both regression and classification task.

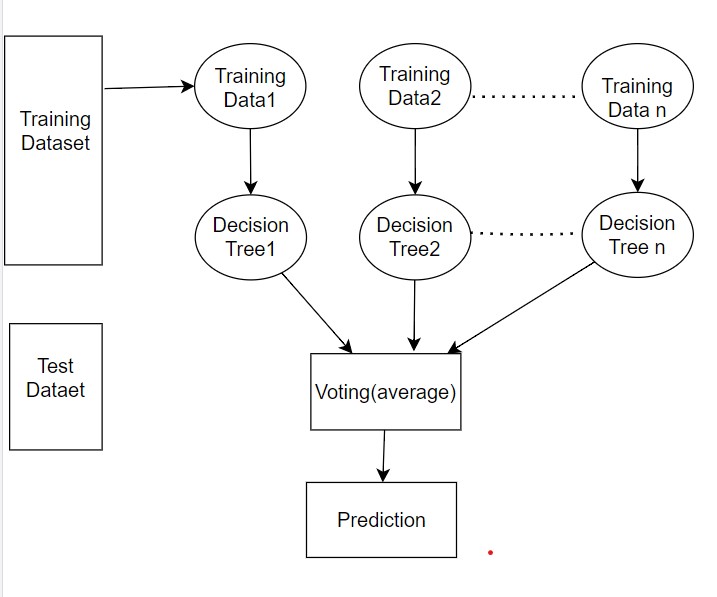


Figure 3.2: Working Mechanism of Random Forest

It is a popular supervised machine learning algorithm based on a decision tree used for both regression and classification tasks. The figure above is the working mechanism of the Random Forest Algorithm Gupta (2023). First random forest generates multiple decision trees for each training dataset where each decision tree is created by taking a random sample of data. A recursive process is used to generate the decision tree where each node algorithm chooses the best features to split the data using Gini impurity or information gain. Until the tree reaches a certain depth this process continues where the output of each node is regression prediction. Lastly, all prediction of each node of the trees is determined by aggregating the prediction. This algorithm has good performance and can maintain accuracy and missing values but there is some limitation meaning it can be prone to overfitting because the number of trees is large. So, ensemble method can be used where it combines multiple decision trees to improve accuracy. It combines the predictions of multiple models which helps to improve the performance of the model.

Equation of Random Forest algorithm:

F(x) = 1/n∑ \_{i=1}^n f\_i(x) (1)

Where F(x) is the prediction for input x, f\_i(x) is the prediction of i-th tree in the forest and n is the total number of trees in the forest.

### 3.1.2 Logistic Regression

It is a machine learning classification algorithm which is used to predict the probability of classes based on some dependent variables. The output of logistic regression is 0 or 1. A sigmoid function is used for the transformation of the linear combination of input features into 0 and 1. This algorithm adjusts the weight and biases of the model using optimization and cost function algorithm during the training process. The cost function calculates the difference between actual binary output and predicted probability. The next step is to make predictions on new data by passing the input features through the model. A Logistic function used to calculate the probability of the output variable. The predicted value is rounded to the nearest value of 1 or 0 to make the final classification decision (Swaminathan, 2019).

(Brownlee, 2020) equation of Logistic Regression is shown below: -

Y = e^(b0 +b1\*x)/ (1+e^(b0 +b1\*x)) (2)

where b0 is intercept term

b1 is coefficient for single input x

### 3.1.3 Decision Tree

The goal of the decision tree is to create a model which learns the decision rule generated from features of the dataset and predicts the value of target variables. It is a tree-like structured classifier which consists of internal nodes, branches, and leaf nodes. The internal node is the features, the branches are the decision rules, and the leaf node are the outcomes (Raj, 2021).

A decision tree algorithm starts from the node for predicting the class of the dataset. Then it compares the real dataset attributes with the root attributes and based on that it follows the branch and go to next node. The algorithm compares the sub-nodes with the attribute values. This process continues until it reaches to leaf node as shown in the figure below (Raj, 2021).

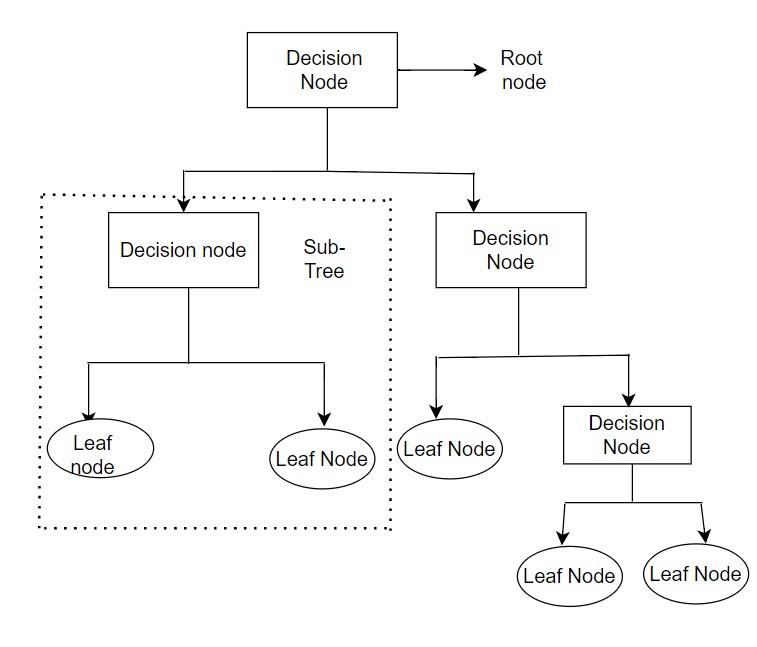


Figure 3.3: Working Mechanism of Decision Tree.

### 3.1.4 LSTM

LSTM is capable of processing sequential data and is capable of learning long-term dependencies. There are three different gate presents in the memory cell of LSTM which are the input gate, forget gate and output gate. The input gate controls what information is added, forget gate controls what information is removed, and the output gate contains what information is output from the memory cell. (Says, 2023)

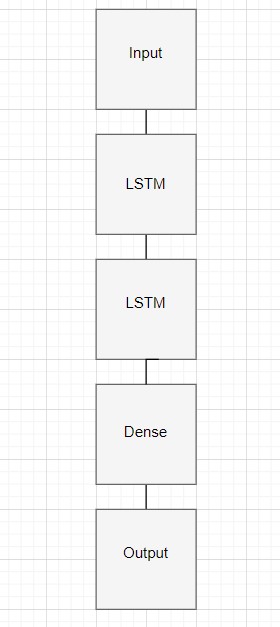


Figure 3.4: Working Mechanism of LSTM

Following are the steps taken place for the LSTM as shown in figure above (says:, 2023)

Step1: After data is pre-processed, it is an input for the LSTM model. I define the architecture of the neural network model.

Step2: Couple of new layers are created to deal with the pre-processed sequential data. I added an embedding layer first.

Step3: next step involves adding the layer to model long short-term memory layer.

Step4: compiling the model using an optimizer which suits the LSTM task also.

Step5: after the compilation we fit the model same way step 2, 3. The model is fit for 100 epochs and loss is calculated using categorical\_crossentropy.

Step6: Now I use the model to make prediction on new items.

## 3.2 Tools Used

1.Excel: It is a powerful spreadsheet which allows to analyse, manipulate, and organize data. It provides various tools and function to perform different complex calculations, build data charts and graphs for visualizations. It offers different features like formatting, sorting filtering so that user can analyse the data effectively.

2.Google Collaboratory: In this paper I used this tool to write and execute python code which runs directly on the browser. As this project uses python programming language and this tool here create an environment for user to run python code without requiring any installation and setup.

3.Hive: Hive is an open-source software which is maintained by Apache Software Foundation. It is a warehouse platform which offers for the querying and analysis to extract meaningful insights from large dataset. Here a pre-configures Virtual Box needs to be setup the Hive queries can be written, and data can be explored.

4.PowerBI: PowerBI is an analytical tool which allows user to visualize the data and creates an interactive report. With powerBI user can connect with different data sources like Excel, databases create reports, then can clean the dataset, and can create report using different graphs, chart, pie charts, scatterplot just by drag and drop interface.

## 3.3 Data Collection

This dataset is quantitative which is downloaded from Kaggle, and this dataset was implemented by different authors to analyse customer buying patterns using different data mining algorithms. The data contains three different attributes which are Member\_number, Date, itemDescription with 38765 values. The data was stored in .csv format where Member\_number is integer, Date and itemDescription are string. This data is an input for the system. This data contains the item sets which are purchased by the customer in a single transaction.

## 3.4 Data pre-processing

Data pre-processing includes the steps to prepare and clean the raw before using it for any analysis or machine learning process. The reason behind pre-processing data involves getting a better result with clean data, avoiding any sort of error while fitting the data, and during the implementation of an algorithm. Besides that, good and clean data ensures the validity and accuracy of the result that has been obtained throughout the process. After the collection of the raw data quality of the data is checked before it is fed to any algorithm. As real-world data is incomplete, inconsistency and noisy. So, to transform those data into meaningful following pre-processing steps are carried out.

### 3.4.1 Data Cleaning

1. Missing Data: It is common that a dataset with missing values. In this transaction dataset, there might be a null value so it is very important to remove those value before it is fed into the model so that an accurate model can perform more accurately. This can be done by either deleting a column or replacing it with 0, mean, median or mode. But in this collected dataset the attributes do not have any missing values.
2. Removing Duplicates Data: Handling and removing duplicate data is another important step so that more accurate results and analyses can be done, but in this paper, I did not remove any duplicate data because a different customers can buy the same item and the same customer can buy the same items multiple times.
3. Lowercasing: The text data of this dataset is converted into lowercase as a part of Data Cleaning which makes it easier to process and analysis.
4. Conversion of Data: In this dataset, same customer bought different items and the items has been stored in each column so I stored the data in a variable length format so that it can be further used for processing.
5. Feature Selection: This projects main aim is to find out what a customer will buy in the future so know to distinguish the customer I have selected the member number and items of the customer as features to predict the items.
6. Encoding: The machine learning algorithm requires numerical input, so the encoding step is carried out. Encoding helps to convert categorical data into numerical form. This step helps to increase the accuracy of the machine-learning model. There are different types of encoding processes such as one hot encoding, label encoding, count encoding and target encoding. This choice of techniques for encoding depends on the requirements and the type of machine learning model.
7. Text Normalization: This step involves the process of removing special characters, converts all the characters to lowercase, punctuation marks. It helps to standardize the text and makes it easier to analyse and process the data.
8. Tokenization: Tokenization is a step used in natural language processing which involves breaking the text into smaller chunks which are represented numerically. So, this step is used in the deep learning algorithm (LSTM) in this project.
9. Padding: Padding is one of the pre-processing steps which ensures all the input sequences of the dataset have the same length. Some of the input sequences have different lengths which is challenging to process using neural networks because of which this step is required for the input to have a fixed size.

## 3.5 Model train and Evaluation

After the pre-processing and algorithm selection step next step is to train the model. Model training depends on the machine learning algorithm which is to be used. The pre-processed data is split into train and test ratios depending on the size of the dataset which is normally 70/30, 75/25 or 80/20 for the machine learning model. The smaller ratio of data is used for testing the performance of the model and the larger portion of data is for training the model.

Once the model is trained the performance of the model is evaluated in a separated test dataset using different evaluation metrics like F1-score, precision and recall analysing whether it is performing well or not. Here I used accuracy score, precision, recall and F1 score to evaluate the machine learning model.

Accuracy\_score: It is a function of sklearn library to calculate the accuracy score of classification model.

Precision: The proportion of true positive (TP) out of all positive predictions as shown in 3.

precision = TP/(TP+ FP) (3)

Recall: The proportion of true positive among all actual positive results as shown in 4.

Recall = TP/(TP+FN) (4)

F1 score: It is the harmonic mean of precision as shown in 5.

F1 = 2\*(precision\*recall)/(precision+recall) (5)

## 3.6 Ethical Consideration

The dataset used here consists of only three attributes Member\_number, Date and item description. Although the member\_number is one of the unique identifiable keys this dataset is already a published dataset. The author or the person who has collected this dataset must have kept all the ethical consideration like the privacy of a customer, consent of the data, transparency, and fairness.

As this research is performed by using already published dataset so there is no ethical consideration applied to this project.

## 3.7 Conclusion

In conclusion, from this chapter I am able to find out the best methodologies for analysing the dataset. This chapter provides a comprehensive overview of these methodology with figure, equation, and diagram. It discussed about different pre-processing steps to be carried out on the implementation section. Also, ethical consideration and tools used has been considered.

# Chapter 4 - Implementation

## 4.1 Introduction

Implementation of the project means the process of executing plans to action so that we can reach the goal with the completion of the project objective. There are several steps to effectively execute the research process which involves setting a project plan, able to find the requirements to execute the plans by forming a design and deploying the process based on the plans of the project. Furthermore, getting the result and evaluating the performance and results.

This section covers the implementation of the algorithm as mentioned in chapter three above also about ways in analysing customer buying patterns. The process to reach the goal cannot be completed in a single step. Data Cleaning and pre-processing play an important role in machine learning and deep learning so that the model can give its best result. Firstly, in this project data is downloaded in CSV format which is a raw data. For further pre-processing, analysing, encoding, and creating model every step which is required to reach the project aim is carried out in this section.

## 4.2 Libraries used for Deep learning.

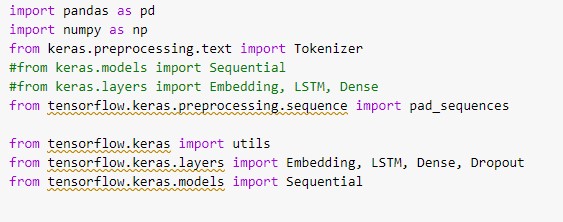


Figure 4.1: Library used for LSTM

Figure above shows the library used for implementation of Deep Learning algorithm. Here is the description of each library as shown below: -

Pandas: it is a powerful library in python for working with data such as manipulation, visualization, and analysis

NumPy: It is python library for mathematical/ numerical operation and data analysis.

Tokenizer: It is a class of Keras used for text pre-processing like conversion of text into numerical form.

Pad\_sequences: It is function for dealing with the variable length used in python with the Keras deep learning library.

Utils: It provides a utility function to deal with one hot encoding and data pre-processing.

Embedding: It is used to map words to dense vector of real number usually in neural network.

LSTM: It is used for processing the sequential data.

Dense: Used to add fully connected layer to neural network also takes care of bias and weight of connected layer

Dropout: This layer prevents the overfitting and encourages network to generalize new data

Sequential: This class provide a way to create linear stack of neural network to define and train deep learning model easily.

## 4.3 Libraries used for Machine Learning



Figure 4.2: Library used for Machine Learning

Figure above shows the library used for implementation of Machine Learning algorithms. Here is the description of each library as shown below: -  
Label Encoder: - It is used to convert categorical data to numerical data for modelling and analysis.

Gaussian: - To run Gaussian Naive Bayes classifier algorithm.

train\_test\_split: - To split dataset into training and testing dataset.

DecisionTreeClassifier: To build decision tree model.

Accuracy\_score: - This is used to calculate the accuracy of the predicted values.  
GridSearchCV: - This is a function used for hyperparameters tuning.

VotingClassifier: - It is used to combine the prediction of different machine learning algorithm.

## 4.4 Analysis of Data

Before processing the data, it is important to understand, analyse the raw data so that we can get valuable insights from those data. This valuable information can help in further pre-processing and distinguish the important features of the data. The analysis of data is carried out using different big data technologies such as Hive, Spark, Hadoop, MongoDB. These big data tools are used for the processing, storing, analysing large dataset to derive a valuable insight from it.

In this project, the analysis of Data I have used Hive Query language and PowerBi for visualization and query analysis.

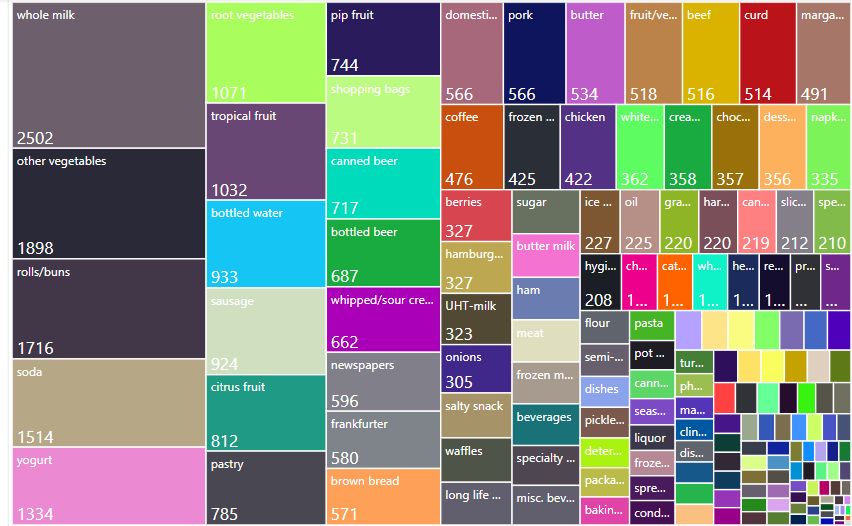


Figure 4.3: Visualization of the number of items repeated

The tree map in figure shows the number of times each items bought. This diagram provides the clear understanding of the dataset and able to analyse the visualized result. It shows that whole is purchased the most followed by other vegetables, roll/buns, soda and so on.

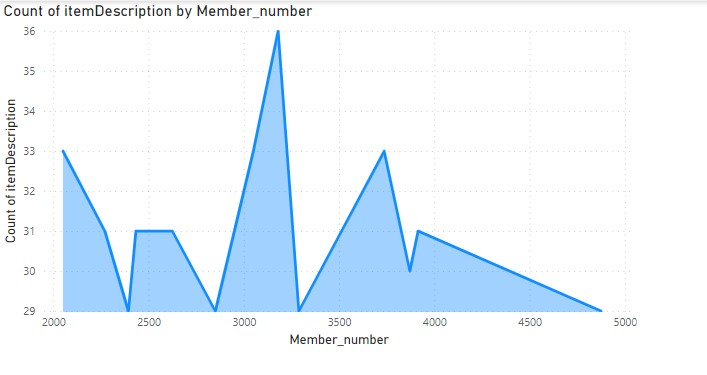


Figure 4.4: Count the number of items purchased by a member.

This chart above shows the information related to item description and member. Here the visualization provides the number the count of items purchased by member number. It can be observed that 3000-member id customer purchased the most items and visited the supermarket the most.

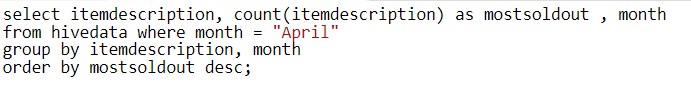




Figure 4.5: Most sold out product in a month of April

The above analysis is performed using Hive query Language which shows the top most sold out item in April month. It can be observed that whole milk, other vegetables is the most sold-out product on month April. In this way items purchased mostly can be analysed for each month and focus on the most sold-out product to increase the customer satisfaction and profit in the market

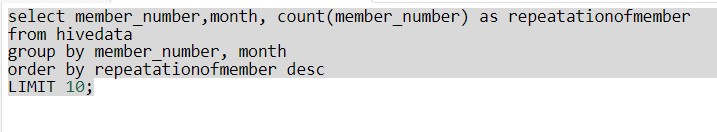




Figure 4.6: Displays in which month the customer buys the items.

The above analysis shows the result based on member number and month. It can be observed that result above shows a customer with 2524 member\_number visited mostly compared to other customer. This customer visited the store for 18 times in month october.

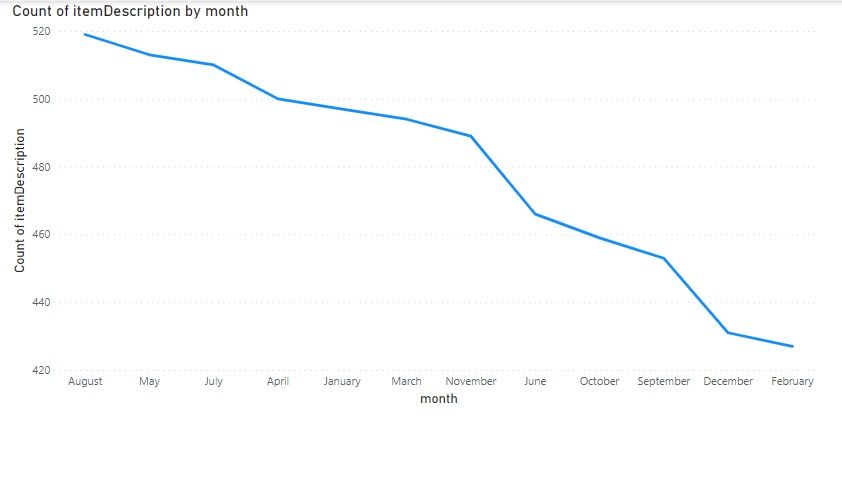


Figure 4.7: Displays the items purchased in a month

Chart above shows the result based on the count of items by month. It can be observed that most of the items are purchased on the month of August and the least purchased items was on month February. This way supermarket can increase the production rate on the month of August and decrease in month of February.

## 4.5 Data pre-processing for Machine Learning

### 4.5.1 Data cleaning

Following are the data pre-processing steps which is carried out before feeding it into model. First step is to group the itemDescription according to the Member\_number. As same number members bought different items in same month so this process is implemented so that there will not be duplication of member number. The main aim of this paper is to analyse customer buying behaviour for that I choose two attribute Customer\_id and items. So, Date attribute is removed and Member\_number is renamed to Customer\_ID and itemDescription is rename to items.

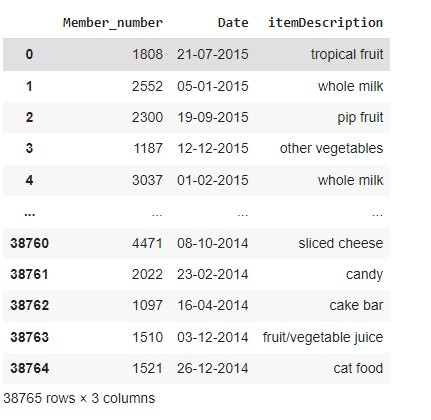
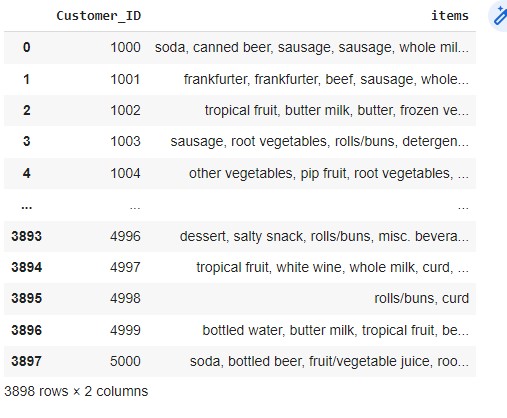
 

Figure 4.8: Converting the dataset into variable length data

After that item the total number of data is 3,898. Then null value is checked.

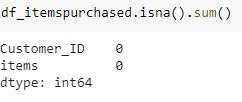


Figure 4.9: Null value check in the dataset

All the item attribute value is converted into lower case.

1.Duplicate values:

A member can purchase same items number of times so there is no duplicate value is not removed.

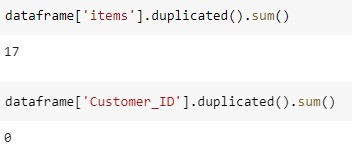


Figure 4.10: Duplicate values check in the dataset

2.Data Filtering:

Here I have taken the most occurred top 5 items and removed other values of item’s column from the dataset.



Figure 4.11: Choosing top 5 most purchased item

3.Encoding:

The items attributes of this dataset are in categorical form so encoding step should be carried out. Since machine learning models are based on mathematical equation due to which we cannot feed categorical data into those mathematical equations. But in this paper MultiLabelBinarizer is used which is a class in scikit-learn which is type of encoding process. As the data used here is in a variable length format which means there are multiclass target variables. The value of each element is either 0 or 1 if the sample or attribute is present in the sample then it is 1 otherwise it is 0.

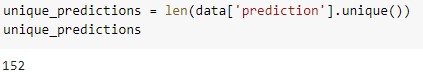


Figure 4.12: Categorical value into numerical form

4.Splitting dataset into independent and dependent variable

The algorithm for this project is supervised learning which requires label dataset. There is absence of label dataset because of which a prediction column is created which store the last value of items column from each row.   
The prediction column store categorical data so the values are replaced with the integer value of item stored and if the item is not present in that column prediction, then it is replaced by NaN.

Null values are replaced after this step because prediction column have 152 and total unique items present is 167

   
 Figure 4.13: Checking the length of label and prediction.

Prediction column is targeting variable and other column except prediction column is features.

5.Train Model

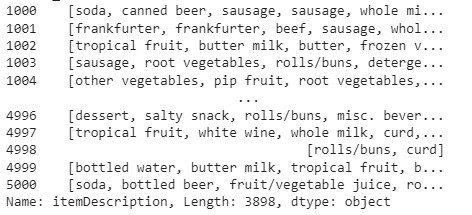
Train\_test\_split function from sklearn.model\_selection for splitting the dataset into training and testing dataset where X\_test, y\_test is the subset of orginal dataset used to evaluate the performance of model and X\_train, y\_train is the subset original dataset. Here the dataset is split into 80/20 ratio.

## 4.6 Data pre-processing for Deep learning

### 4.6.1 Data Cleaning

1. Text Normalization

The dataset consists of items bought by member id for over past one year which is downloaded from Kaggle. Then Pandas library is used to read the dataset. After the dataset is loaded the item is grouped by each member. id.  
The itemDescription attribute of the dataset consist of value with two word like tropical fruit so for the further processing I need to clean the dataset. So, all the punctuation marks are removed and space between the word of single item is removed by underscore like tropical\_fruit as shown in figures below.



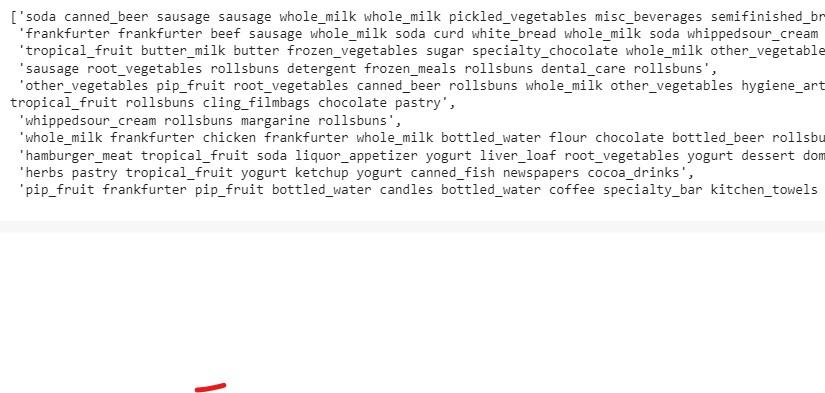


Figure 4.13: Data before and Data after cleaning

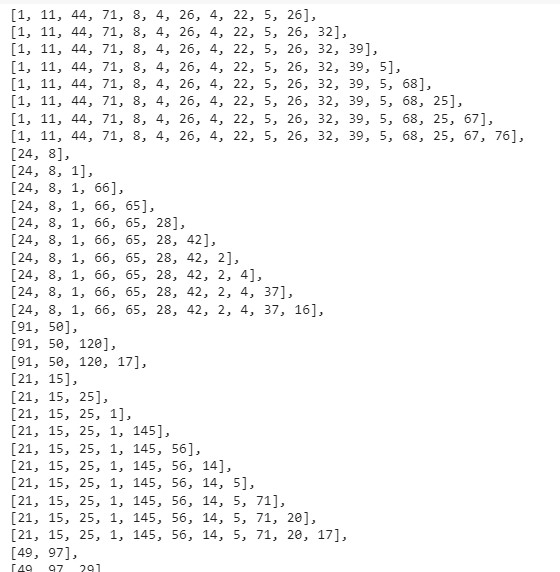
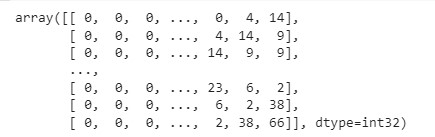
2.Tokenization  
This step is carried out because tokenization is an important step which is needed for further processing. It is another important step to prepare text data for LSTM model. This step breaks down the text data to smaller chunks (tokens) then encode the text data into numerical form, which is easier for model to understand, and this model requires numerical input data. Keras have a Tokenizer class used for text tokenization, so I imported that library. Now the items are tokenized means each word of items is converted into representative number, sequences of tokens will be created. For every item a sequence is formed, tokenizer is used as show in figure below. These sequences are what we will use to train the LSTM model.



Figure 4.14: Tokenization result

3.Padding

Here, the sequences are formed having different variable length. So, before model training each sequences need to have same length. For that Keras has a pad\_sequences built-in method which we can use to do padding.

Before padding result After padding result

Figure 4.15: Before and After Padding Result

4.Splitting dataset

Then splitting of dataset into target variable and predictors is done where last word of each sequence will be labels and predictors is every word except last word

## 4.7 Implementation of Selected Algorithm

### 4.7.1 Logistic Regression and result

After the data is pre-processed as defined in above section,  
the pre-processed data is split into testing and training sets. Logistic Regression class in sklearn.linear\_model model is used to perform Logistic Regression. A new object lr is created of Logistic regression class with random\_state and max\_iter parameters. Then the logistic regression model is train with the training dataset. Once the model is trained it can be used for the prediction of items are using the testing dataset. So, after the prediction step is implemented all the result of prediction is stored in y\_pred variable. Then I used argsort() method to find top three items.

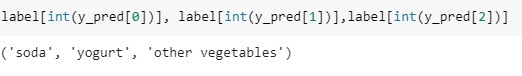


Figure 4.16: Predicted Result from Logistic Regression

The figure above shows that soda and yogurt are the items which will be purchased by customer in future.

### 4.7.2 Decision Tree and result

After data is pre-processed and is split into testing and training dataset the algorithm is trained on training dataset. Decision Tree class in in sklearn.linear\_model model is used to perform Decision Tree Classifier. Each node represents feature, and each branch is the representation of rule and leaf node represents the numeric value. Here Decision Tree classifier object is defined with maximum depth of 3 which is hyperparameter that prevents overfitting by controlling the depth of tree. The decision classifier is trained on the training dataset where dataset is divided into 80/20 in data pre-processing section using fit method. This method uses an algorithm which recursively split the node of decision tree based on feature. Gini impurity index is use by scikit learn to choose the feature. Items are predicted using the testing dataset on another step by using predict() method. The X\_test variable is feature matrix of testing dataset where predict() uses trained decision tree model which classifies the sample of testing dataset. The output in y\_pred variable is vector of predicted labels for each sample in testing dataset.

The result of the predicted value is stored in y\_pred variable. Then I use argsort() method to find top two items.

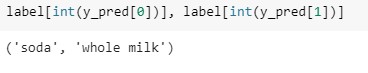


Figure 4.17: Predicted Result from Decision Tree

The figure above shows that soda and whole milk can be purchased by a customer in future.

### 4.7.3 Random Forest Algorithm and result

After the data is pre-processed as mentioned above, RandomForestClassifier class from the scikit-learn library for Random Forest algorithm. Then the algorithm is applied on the training dataset. Like other machine learning algorithm fit method is used to train random forest model on training dataset. During this process each decision tree is built using randomly sampled subset of training data and features. This step is repeated until the specified tree is formed. Once the model is trained, items are predicted using the testing dataset on another step by using predict() method. The prediction is made by using majority vote of each individual tree prediction. The predicted value is stored in y\_pred variable. The result gives the top two predicted values to recommend the items.

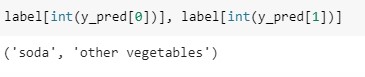


Figure 4.18: Predicted Result from Random Forest Algorithm

The figure shows that soda and other vegetables are the top 2 items which can be purchased.

### 4.7.4 LSTM and Result

Long Short-Term Memory is a powerful tool to process sequential data which is very important in analysing customer buying pattern and helps to identify the patterns and trends of customer purchasing behaviour. The data used in this research is sequential data, so LSTM model is implemented here. After the data is pre-processed, data is fed into the LSTM model. There is input layer, LSTM layer and output layer.

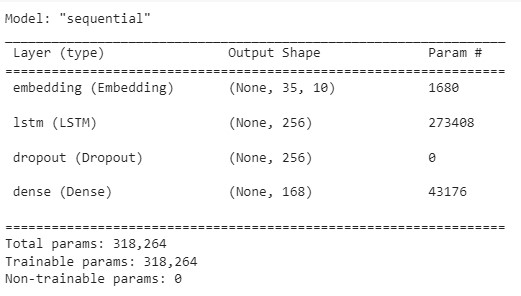


Figure 4.19: Model Compilation

After that model is compiled with categorical cross entropy as an item is to be predicted total itemsets. Then data is fitted into the model, where the model is trained for 100 epochs. The loss of the model is only 1.9283 in average when the model is trained for 100 epochs. Loss is the measure of how well the model is able to predict the model. The lower the value of loss it indicates more accuracy. The accuracy is 0.5157 which means model us predicting output correctly.

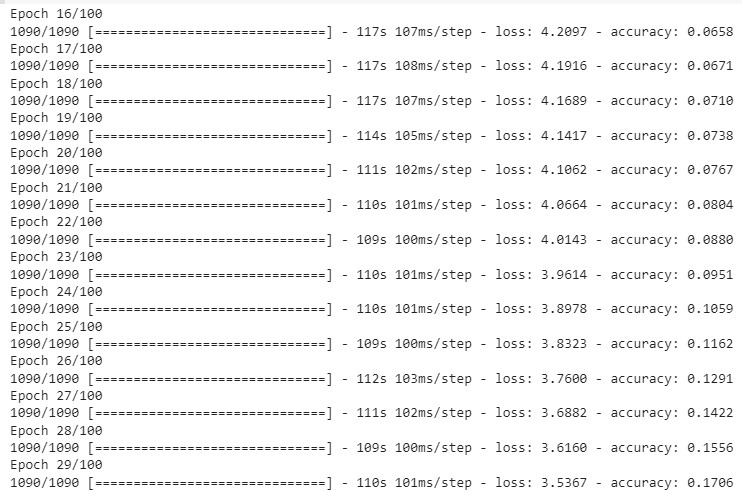


Figure 4.20: Decrease in loss in course of training.

The algorithm can predict the items which will be purchased next if particular items are in the basket.





Figure 4.21: Predicted items of LSTM model

The figure above shows that if a customer buys herbs, pastry, tropical fruit, yogurt, newspapers, cocoa drinks, canned fish the s/he can buy whole milk as predicted by the algorithm.

## 4.8 Result evaluation of three different Machine learning models

### 4.8.1 accuracy measurement of LR

For the accuracy of the LR predicted value accuracy function of sklearn.metrics module is used which is accuracy\_score. The comparision between predicted values and test data is performed. The accuracy score ranges from 0 to 1 where 0 indicates no accuracy and 1 indicates perfect accuracy. The accuracy of Logistic Regression algorithm in this dataset is 73%. Here I tried using different methods like hyperparameters tuning, regularization, ensemble methods but it did not work.

### 4.8.2 accuracy measurement of DT

For the accuracy of the DT predicted value accuracy function of sklearn.metrics module is used which is accuracy\_score. The comparison between predicted values and test data is performed. The accuracy score ranges from 0 to 1 where 0 indicates no accuracy and 1 indicates perfect accuracy.  
The accuracy of Decision Tree in this dataset is 59% which is quite low. There are different types of methods to increase the accuracy of Decision Tree. But here hyperparameters tuning only worked. This method is applied where the accuracy of machine learning needs to be increased. It is applicable for regression and classification. So here I used hyperparameters tuning to increase the accuracy. The accuracy score value is increased by 14%.

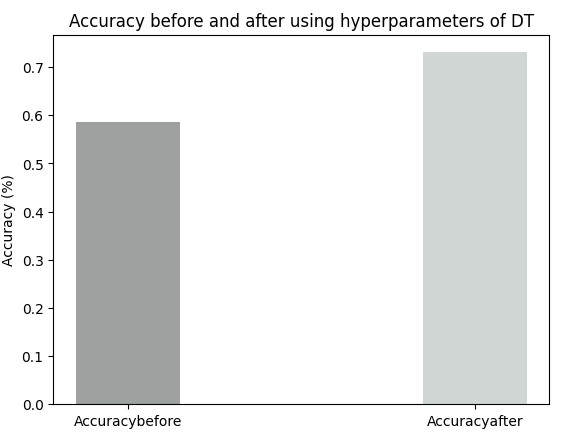


Figure 4.22: Accuracy before and after hyperparameters of DT

### 4.8.3 accuracy measurement of RF

For the accuracy of the LR predicted value accuracy function of sklearn.metrics module is used which is accuracy\_score. The comparision between predicted values and test data is performed. The accuracy score ranges from 0 to 1 where 0 indicates no accuracy and 1 indicates perfect accuracy.  
The accuracy of Logistic Regression algorithm in this dataset is 68%. I used different method to increase its accuracy. But for this algorithm ensemble method where I used rf, lr and nb and the accuracy increased by 6%.

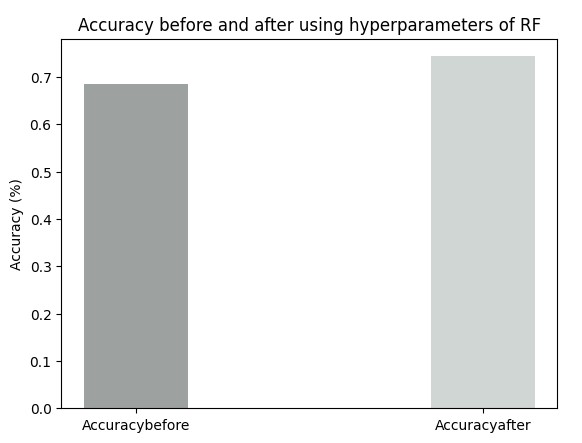


Figure 4.23: Accuracy before and after hyperparameters of RF

The table below shows the overall accuracy evaluation using different methods for SML algorithm.

Table 4.1: Model Accuracy before use of methods to increase accuracy

|  |  |  |  |
| --- | --- | --- | --- |
| Evaluation Metrics | RF | DT | LR |
| Accuracy score | 68.58 | 58.54 | 73.74 |
| Precision | 68.72 | 42.70 | 78.54 |
| Recall | 68.58 | 58.54 | 73.74 |
| F1 score | 68.57 | 48.72 | 73.33 |

Table 4.2: Model Accuracy after use of methods to increase accuracy

|  |  |  |  |
| --- | --- | --- | --- |
| Evaluate Metrics | RF | DT | LR |
| Accuracy score | 74.58 | 73.07 | 74.15 |
| Precision | 78.16 | 78.26 | 79.04 |
| Recall | 74.35 | 72.63 | 74.15 |
| F1 score | 73.72 | 73.07 | 73.77 |

## 4.9 Comparison between Models

From above analysis it can be concluded that different classification algorithm like LR, DT, RF can be used by retailer store to analyse customer buying pattern for item prediction. In addition, Deep learning algorithm like LSTM can be used for item prediction that a customer can purchase.

Based on the above accuracy measurement it showed logistic regression which is machine learning algorithm performed well for this dataset. Meanwhile after using different method to increase the accuracy Random Forest and Decision Tree also performed well. Meanwhile the LSTM model performed well for this dataset for the item prediction, but accuracy measure of the result predicted by LSTM is not measured using evaluation metrics. The table below shows the model evaluation using different evaluation metrics.

Table 4.3: Model Accuracy of DL and ML.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Evaluate Metrics | RF | DT | LR | LSTM |
| Accuracy score | 74.58 | 73.07 | 74.15 | N/A |
| Precision | 78.16 | 78.26 | 79.04 | N/A |
| Recall | 74.35 | 72.63 | 74.15 | N/A |
| F1 score | 73.72 | 73.07 | 73.77 | N/A |

# Chapter 5 - Limitation and Future Work

## 5.1 Limitation

Analysing customer buying pattern using historical data provides valuable insight into customer buying pattern but there is always some limitation. Firstly, the data used here are purchased items by a customer in past. analysing historical data is not able to capture the emerging trends and might not fulfil the changing demand of the customer because of which its result might lead to outdated and low accuracy results. There are some variables like social trends and economy which may impact customer buying nature but these variables are not captured in the historical data.

## 5.2 Future Work

Customer satisfaction is the motto of every company, supermarket, store and shopping centre. Providing customer satisfied products is a challenging task for the day-to-day changing demand. It requires businesses to understand their customer, collect and analyse data (purchasing patterns, review) to gain valuable insight into what a customer wants. In this research historical data is used for the prediction of items which can be purchased but in the future real-time data should be collected and analysed. Collecting real-time data from different tools like sensors, IoT devices and other devices help the business to provide up-to-date information about customer buying nature. This allows up-to-date prediction which leads business to full their customer requirements on time. Also, data can be collected from diverse sources like social media and marketing data with the addition of different variables like seasonal factors for the improvement of the accuracy of the model. Also, the data fed into the model must be regularly maintained and updated.

The choice of model is playing an improvement role to provide good results for the analysis of buying patterns. So, in the future more sophisticated models like deep ensemble methods or deep neural networks can be developed. By testing this type of model, researchers can find the best-performing model to solve the addressed problem. The study here focuses only on the grocery dataset the model must be tested on online shopping, fashion retail stores or in different contexts. Doing this can identify the potential of the model. Different feedbacks can be collected from the customer about the prediction made by models so that those feedbacks can be used for the improvement of the model and helps in increasing the accuracy of the model. In addition, evaluation of Deep learning model LSTM against the other machine learning algorithm can be carried out.

# Chapter 6- Conclusion

For the prediction of retail store products is a challenging task as it depends on various factors. By predicting customer buying patterns, it helps to understand what kinds of items a customer purchased and which items can be sold most.

To the recap this research aims to analyse the customer buying patterns to find out which items will be purchased next which helps supermarkets to fulfil the customer requirements demands to address this goal there are four objectives which are: -

To determine the importance of analysing customer buying nature.

To understand customer needs and preferences every month

To determine which algorithm is suitable for this dataset

To predict new products according to the historical transaction dataset.

All the above objectives are achieved at the end of the project. There are four research question which answer are presented in tabular format below.

Table 6.1: Research question answers

|  |  |
| --- | --- |
| Research question | Answers |
| Why is it important to analyse customer buying nature? | The answer for this question is from chapter one and literature review which shows why is it necessary to analyse customer buying nature. From the study it can be observed that it is high time, customer needs should be addressed so that any company or organization can fulfil their need. For that analysing customer transaction data or retails real time is an important factor for any organization or business |
| What are the customer preferences and needs in a month? | From the above study, importance in analysis is identified then answer for RQ2 is from implementation and methodology. Where the answer is identified about the most frequently purchased items in a month using Hive. As an example, from this dataset I am able to derive that on the month of April whole milk was the most sold item. This way the mostly purchased items for the following months can be known. Basically, I am able to find the customer needs and their preferences in particular time frame. |
| Which is the suitable algorithm for this dataset? | The model used here has obtained result based on historical grocery dataset. The dataset is pre-processed before fed into the model. Various algorithm has been implemented and compared and best fit model is obtained. The models RF, DT, LR, LSTM are implemented here. The choice of appropriate algorithm depends on the available data, special aim of the analysis also different other factors. So different algorithm has been implemented and tested so that it performs well, and one can achieve good results. Here I managed to find the accuracy of each model implemented so I concluded that Logistic Regression performed well which accuracy was 73% in predicting items to be purchased this answer is for the third research question. |
| * What are the suggestion and recommendation of the products based on results? | Based on the result of analysis it appears that these models were successful for analysing customer buying patterns on historical customer transaction data and able to predict which item will be purchased next. The practical application of the model is for retail store, supermarket for the analysing customer data and give recommendation of products. |

Overall, from this project I have been able to find the items which can be bought in future, the aim of the project using 4 different models. The study showed that every model performed well for the dataset used here but Logistic Regression performed well in comparison to other. For more effective and good results, in future real time datasets can be used for better results.

# **References**

Ayu, S.K., Surjandari, I. and Zulkarnain, Z. (2018) “Mining Association rules in Seasonal transaction data,” 2018 5th International Conference on Information Science and Control Engineering (ICISCE) [Preprint]. Available at: <https://doi.org/10.1109/icisce.2018.00074>.

Bhaskar Velpula, V., Pakanati, S. and M, S. (2010) “Analyzing target customer behaviour by mining the E-COMMERCE data,” International Journal on Information Sciences and Computing, 4(1), pp. 27–30. Available at: https://doi.org/10.18000/ijisac.50062.

Brownlee, J. (2020) Logistic regression for machine learning, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/logistic-regression-for-machine-learning/ (Accessed: April 21, 2023).

Ding-An Chiang et al. (2009) “The cyclic model analysis on sequential patterns,” IEEE Transactions on Knowledge and Data Engineering, 21(11), pp. 1617–1628. Available at: <https://doi.org/10.1109/tkde.2009.36>.

Diwandari, S. and Zaky, U. (2021) “Analysis of Customer Purchase Behavior Using Association rules in e-shop,” 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) [Preprint]. Available at: <https://doi.org/10.1109/icitisee53823.2021.9655892>.

Dlamini, M.G. et al. (2015) “Extracting interesting patterns from e-commerce databases to ensure customer loyalty,” 2015 IEEE 12th International Conference on Networking, Sensing and Control [Preprint]. Available at: <https://doi.org/10.1109/icnsc.2015.7116067>.

Gupta, A.K. and Gupta, C. (1970) Analyzing customer behavior using data mining techniques: Optimizing relationships with customer: Semantic scholar, Management Insight. Available at: https://www.semanticscholar.org/paper/ANALYZING-CUSTOMER-BEHAVIOR-USING-DATA-MINING-WITH-Gupta-Gupta/8d34a4028af65f7639f5dfca27213e9f5a936818 (Accessed: February 16, 2023).

Gupta, B. (2023) Random Forest algorithm in machine learning, Scaler Topics. Scaler Topics. Available at: https://www.scaler.com/topics/machine-learning/random-forest-algorithm/ (Accessed: April 20, 2023).

Isa, N., Neddy, S.K. and Mohamed, N. (2021) “Association rule mining using FP-growth algorithm to prevent Maverick buying,” 2021 IEEE 11th IEEE Symposium on Computer Applications &amp; Industrial Electronics (ISCAIE) [Preprint]. Available at: <https://doi.org/10.1109/iscaie51753.2021.9431821>.

Krishnamurthy, M. et al. (2013) “Prediction of customer buying nature from frequent itemsets generation using Quine-Mccluskey method,” IET Chennai Fourth International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2013) [Preprint]. Available at: <https://doi.org/10.1049/ic.2013.0345>.

Kaur, M. and Kang, S. (2016) “Market basket analysis: Identify the changing trends of Market Data Using Association Rule Mining,” Procedia Computer Science, 85, pp. 78–85. Available at: https://doi.org/10.1016/j.procs.2016.05.180.

Khade, A.A. (2016) “Performing customer behavior analysis using big data analytics,” Procedia Computer Science, 79, pp. 986–992. Available at: <https://doi.org/10.1016/j.procs.2016.03.125>.

Lico, L., Enesi, I. and Cico, B. (2021) “Analyzing performance of clustering algorithms on a real retail dataset,” 2021 International Conference on Information Technologies (InfoTech) [Preprint]. Available at: https://doi.org/10.1109/infotech52438.2021.9548359.

Liu, S., Liao, G. and Ding, Y. (2018) “Stock transaction prediction modeling and analysis based on LSTM,” 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA) [Preprint]. Available at: https://doi.org/10.1109/iciea.2018.8398183.

Maheswari, K. and Priya, P.P. (2017) “Predicting customer behavior in online shopping using SVM Classifier,” 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS) [Preprint]. Available at: https://doi.org/10.1109/itcosp.2017.8303085.

Moens, S., Aksehirli, E. and Goethals, B. (2013) “Frequent itemset mining for Big Data,” 2013 IEEE International Conference on Big Data [Preprint]. Available at: https://doi.org/10.1109/bigdata.2013.6691742.

Mujianto, A.H. et al. (2019) “Consumer Customs analysis using the Association rule and Apriori Algorithm for determining sales strategies in retail central,” E3S Web of Conferences, 125, p. 23003. Available at: https://doi.org/10.1051/e3sconf/201912523003.

Nasti, S.J., Asgar, M. and Butt, M.A. (2017) “Analysis of customer behaviour using modern data mining techniques,” International Journal of Computer Sciences and Engineering, 5(12), pp. 64–66. Available at: https://doi.org/10.26438/ijcse/v5i12.6466.

Nithin, S.S. et al. (2022) “Retail demand forecasting using CNN-LSTM model,” 2022 International Conference on Electronics and Renewable Systems (ICEARS)[Preprint]. Available at: https://doi.org/10.1109/icears53579.2022.9752283.

Ponyiam, P. and Arch-int, S. (2018) “Customer behaviour analysis using data mining techniques,” 2018 International Seminar on Application for Technology of Information and Communication [Preprint]. Available at: https://doi.org/10.1109/isemantic.2018.8549803.

Raj, A. (2021) An exhaustive guide to decision tree classification in python 3.X, Medium. Towards Data Science. Available at: https://towardsdatascience.com/an-exhaustive-guide-to-classification-using-decision-trees-8d472e77223f (Accessed: April 21, 2023).

Raorane, A. and Kulkarni, R.V. (2011) “Data Mining Techniques: A source for consumer behavior analysis,” International Journal of Database Management Systems, 3(3), pp. 45–56. Available at: https://doi.org/10.5121/ijdms.2011.3304.

Ramaswamy, S. and DeClerck, N. (2018) “Customer perception analysis using deep learning and NLP,” Procedia Computer Science, 140, pp. 170–178. Available at: https://doi.org/10.1016/j.procs.2018.10.326.

Schwenke, C., Vasyutynskyy, V. and Kabitzsch, K. (2010) “Simulation and analysis of buying behavior in supermarkets,” 2010 IEEE 15th Conference on Emerging Technologies &amp; Factory Automation (ETFA 2010) [Preprint]. Available at: https://doi.org/10.1109/etfa.2010.5641255.

Sharma, A. et al. (2022) “Machine learning approach: Consumer buying behavior analysis,” 2022 IEEE Pune Section International Conference (PuneCon) [Preprint]. Available at: https://doi.org/10.1109/punecon55413.2022.10014928.

Shrirame, V. et al. (2020) “Consumer behavior analytics using machine learning algorithms,” 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) [Preprint]. Available at: https://doi.org/10.1109/conecct50063.2020.9198562.

Singh, S.P. et al. (2018) “Data Mining: Consumer Behavior Analysis,” 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information &amp; Communication Technology (RTEICT) [Preprint]. Available at: <https://doi.org/10.1109/rteict42901.2018.9012300>.

Somvanshi, M. et al. (2016) “A review of machine learning techniques using decision tree and support vector machine,” 2016 International Conference on Computing Communication Control and automation (ICCUBEA) [Preprint]. Available at: https://doi.org/10.1109/iccubea.2016.7860040.

Tai, Y., Sun, Z. and Yao, Z. (2021) “Content-based recommendation using machine learning,” 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP) [Preprint]. Available at: https://doi.org/10.1109/mlsp52302.2021.9596525.

Valecha, H. et al. (2018) “Prediction of consumer behaviour using random forest algorithm,” 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) [Preprint]. Available at: https://doi.org/10.1109/upcon.2018.8597070.

Weijs, D. and Caron, E. (2022) “Customer journey analytics: A model for creating diagnostic insights with process mining,” Proceedings of the 17th International Conference on Software Technologies [Preprint]. Available at: <https://doi.org/10.5220/0011263900003266>.

Zhang, Y., Wang, A. and Hu, W. (2022) “Deep learning-based Consumer Behavior Analysis and Application Research,” Wireless Communications and Mobile Computing, 2022, pp. 1–7. Available at: https://doi.org/10.1155/2022/4268982.

# **Bibliography**

Panjaitan, Suprianto, et al. “Implementation of Apriori Algorithm for Analysis of Consumer Purchase Patterns.” Journal of Physics: Conference Series, vol. 1255, no. 1, Aug. 2019, p. 012057. DOI.org (Crossref), https://doi.org/10.1088/1742-6596/1255/1/012057.

Wason, Ritika. “Deep Learning: Evolution and Expansion.” Cognitive Systems Research, vol. 52, Dec. 2018, pp. 701–08. DOI.org (Crossref), <https://doi.org/10.1016/j.cogsys.2018.08.023>.

Sheshasaayee, Ananthi, et al. “Implementation of Clustering Technique Based RFM Analysis for Customer Behaviour in Online Transactions.” 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, 2018, pp. 1166–70. DOI.org (Crossref), <https://doi.org/10.1109/ICOEI.2018.8553873>.

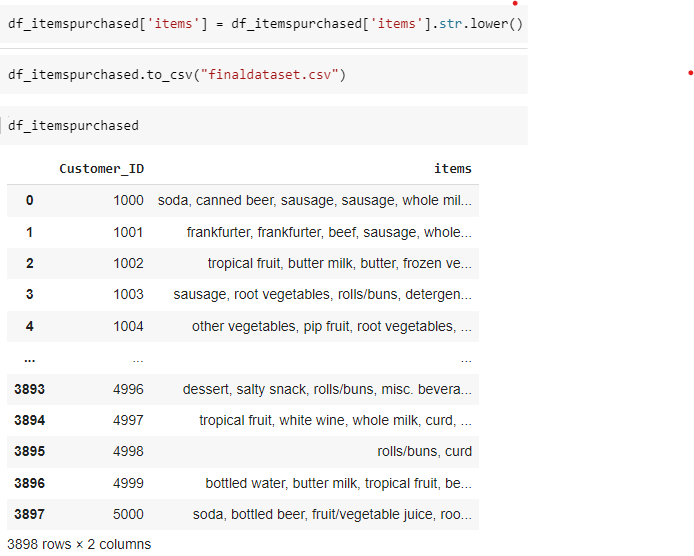
Manohar, E., et al. “A Collective Data Mining Approach to Predict Customer Behaviour.” 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE, 2021, pp. 1310–16. DOI.org (Crossref), <https://doi.org/10.1109/ICICV50876.2021.9388558>.

Gopal, Akshay C., et al. “Customer Behavior Analysis Using Unsupervised Clustering and Profiling: A Machine Learning Approach.” 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, 2022, pp. 2075–78. DOI.org (Crossref), https://doi.org/10.1109/ICACITE53722.2022.9823646.

# **Appendix-A**



Figure A1: Dataset



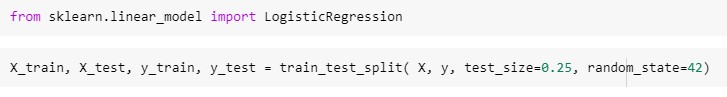
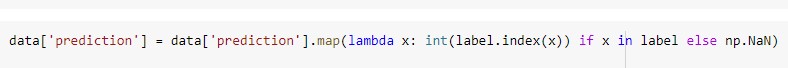
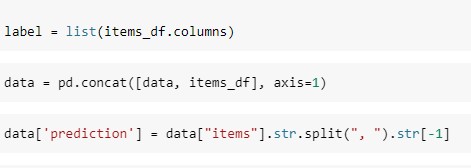
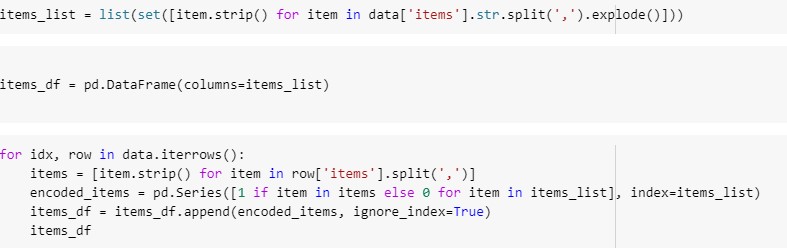
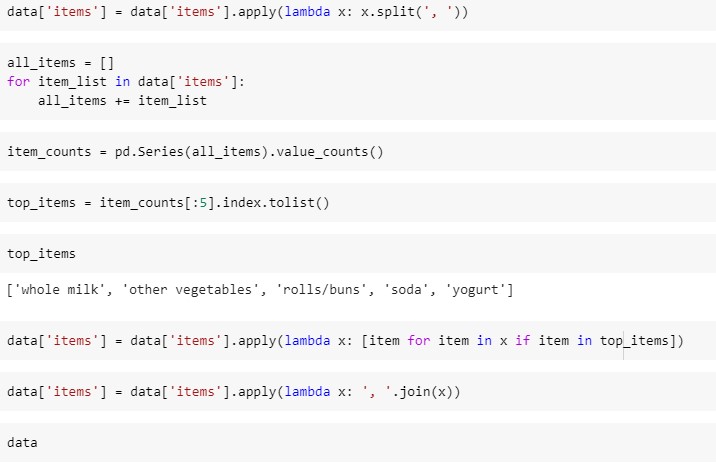


Figure A2: Data Pre-processing for Machine Learning

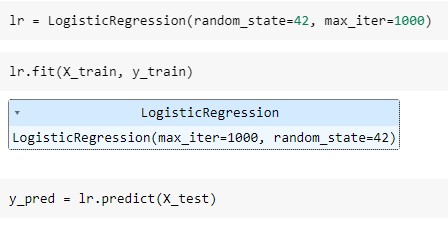


Figure A3: Logistic Regression Model



Figure A4: Accuracy measurement of Logistic Regression

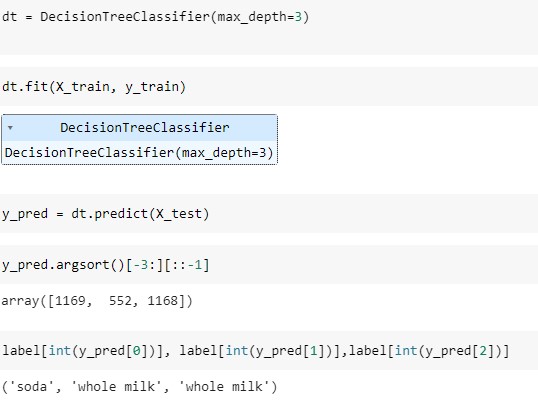


Figure A5: Decision Tree Model

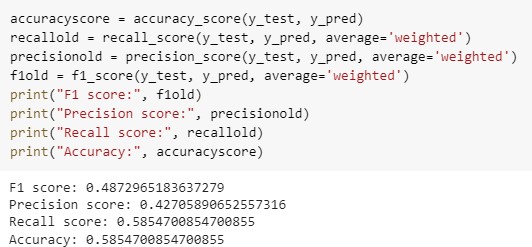


Figure A6: Accuracy measurement of Decision Tree



Figure A7: Hyperparameters tuning and accuracy measure

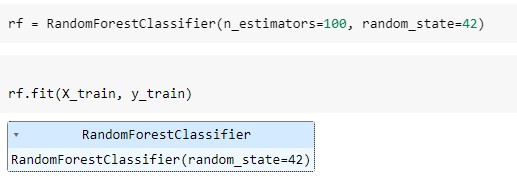


Figure A8: Random Forest Model

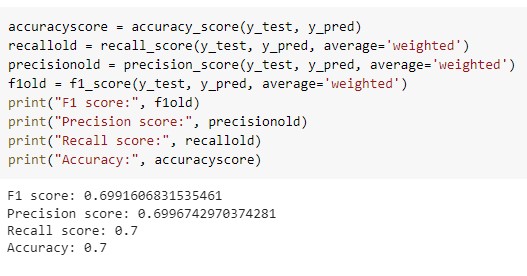


Figure A9: Accuracy measurement of RF

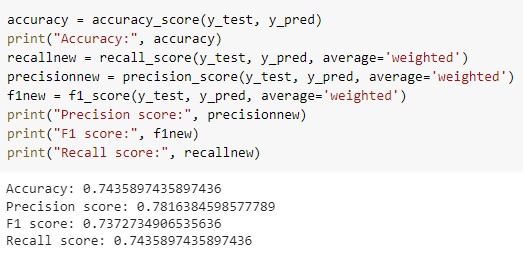
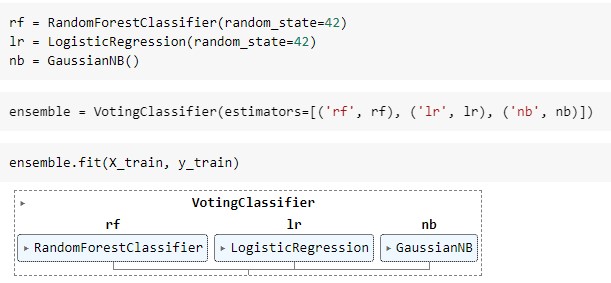
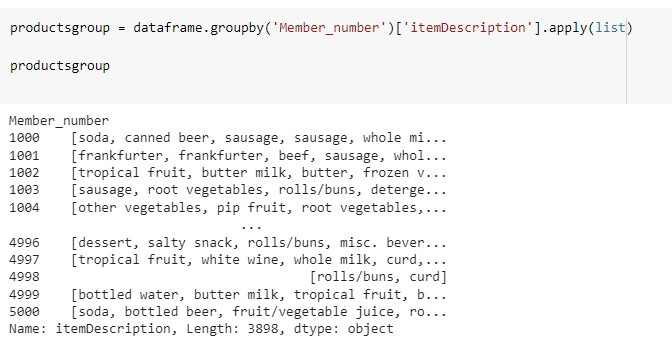
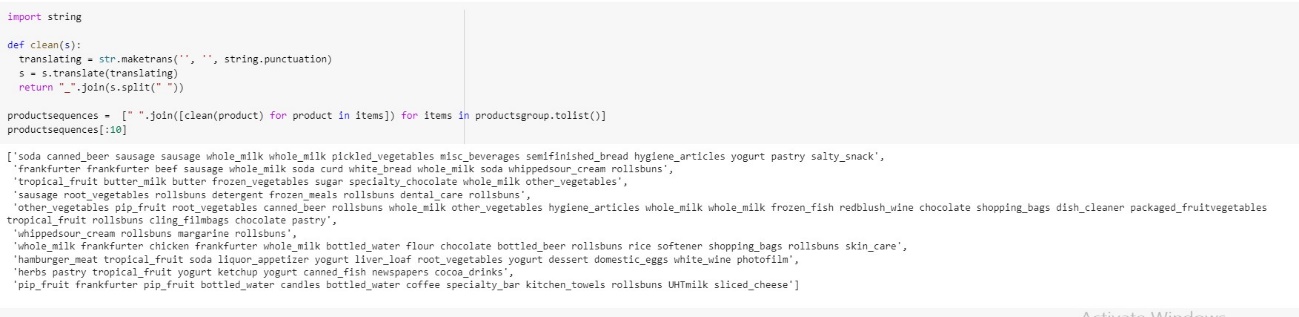


Figure A10: Ensemble Method and Accuracy measure







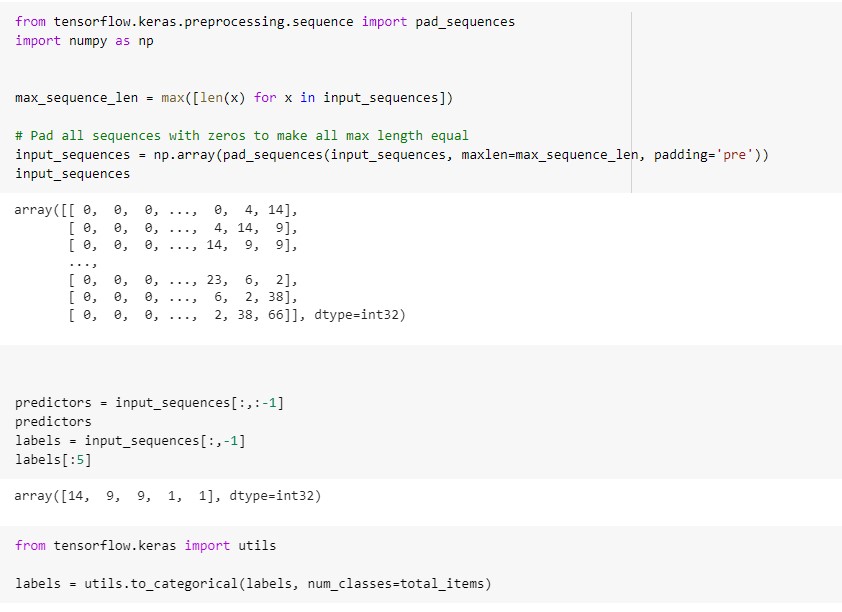


Figure A11: Data Cleaning for Deep Learning

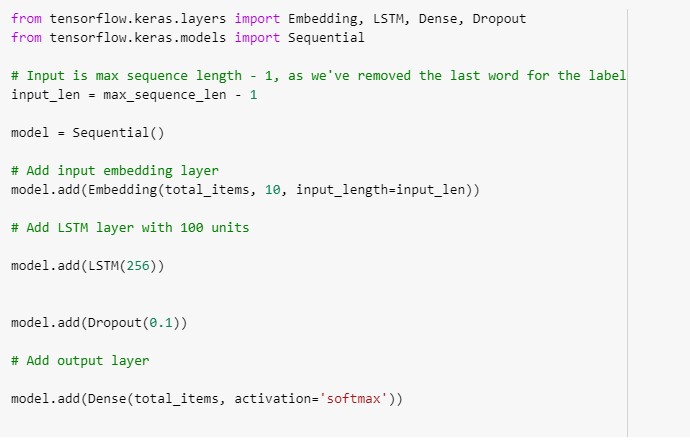






Figure A12: LSTM model fit and compilation

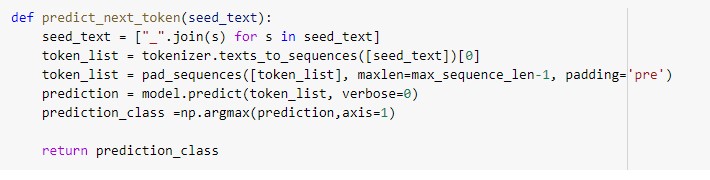




Figure A13: Prediction of new items

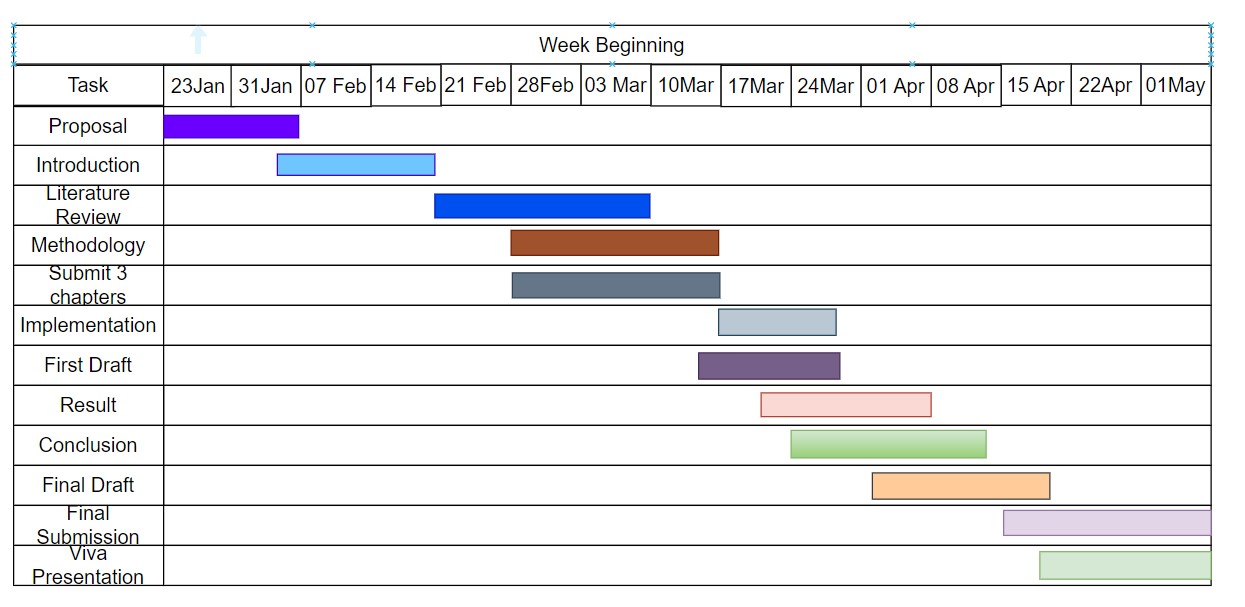


Figure A14: Gantt Chart