ACKNOWLEDGEMENT

Alhamdulillah, praises and thanks to Allah, the Almighty, for His showers of blessings, I can complete this research successfully within the time duration given. He has provided me strength and encouragement during the difficult moments in finishing this research. I am eternally grateful for His endless love, mercy, and grace.

First and foremost, I would like to express the deepest appreciation to my supervisor, Miss Nor Hasnul Azirah Binti Abdul Hamid, for her continual supervision and understanding. She has provided me with valuable feedback and suggestions that have been a substantial aid in this accomplishment. Without her guidance, this project would not have been possible.

Further, I extend my gratitude for my final year project lecturer, Madam Ummu Fatihah Binti Mohd Bahrin, for her careful instructions and guidance from the beginning until the end of my final year project journey. Her positive encouragement while ensuring that my progress is always on track are something I really appreciate. It has been a great pleasure and honor to have her as my lecturer.

My special thanks go to my examiner, Dr Gloria Jennis Tan, for her constructive criticism and advice in helping me to improve my research on this project. Thanks also should go to all the owners of my reference materials who have provided numerous information and knowledge for my research.

I wish to express my sincere appreciation to my beloved parents for their incredible love and support. Their constant prayers for me throughout my journey have allowed me to complete this project successfully. I also would like to give my gratitude to my friends for their direct or indirect assistance, as well as the emotional support they have given, which has motivated me to continue with my research. It is a blessing to have them as friends.

May Allah shower the above mentions individuals with success and honor in their life.

ABSTRACT

Sales Prediction in Media Platform Advertising Expenditure Using Linear Regression is a critical research endeavor addressing the challenges associated with forecasting sales in the context of media platform advertising expenditure. The study focuses on the application of the Linear Regression algorithm to predict sales outcomes based on advertising spending patterns. Accurate sales predictions are essential for businesses to optimize their advertising strategies and maximize return on investment. The primary goal of this research is to develop a robust prediction system capable of aiding businesses in the media industry by providing insights into potential sales outcomes. Early and accurate sales predictions can significantly impact decision-making processes, allowing organizations to allocate resources effectively and enhance overall marketing strategies. The methodology employed in this project involves the implementation of the Linear Regression algorithm, a statistical modeling technique that analyzes the relationship between advertising expenditure and sales. The algorithm calculates a linear equation that represents the best-fit line through the data points, enabling the prediction of future sales based on advertising investment. Throughout the research, three key objectives were achieved: identifying the requirements for sales prediction in media platform advertising using Linear Regression, developing a prototype of the sales prediction system based on Linear Regression, and evaluating the accuracy of the algorithm in predicting sales outcomes. The project progressed through various phases, including Preliminary Study, System Design, Implementation and Development, and Testing and Evaluation. The developed sales prediction system demonstrated a commendable accuracy level of 99%. To further enhance the system's performance, future work could focus on refining data preprocessing techniques and incorporating a comprehensive database. In conclusion, this research contributes to the advancement of sales prediction methodologies in the media industry, offering businesses a valuable tool to optimize their advertising strategies and improve decision-making processes. The success of the project underscores the significance of leveraging machine learning algorithms, specifically Linear Regression, for accurate and timely sales predictions in the dynamic landscape of media platform advertising.

TABLE OF CONTENT

CONTENTS **PAGES**

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

This chapter will provide a summary of the study as well as an explanation of the purpose of this research. The reader will be enabled in digging further into the research and will be better able to comprehend both the objective of the investigation and its scope. This chapter provides the study's context, problem statement, objective, scope, significance, anticipated result, and conclusion.

1.1 BACKGROUND STUDY

This study looks at how the relationship between advertising expenditure and sales in media platforms may be analyzed with the use of linear regression. When it comes to prediction for sales, this sort of research may be helpful to companies since it takes into thought the amount of money invested in advertising. This study can identify critical elements that drive sales and optimize advertising strategies to maximize return on investment by using linear regression, which is a statistical technique.

Advertising expenditures should be based on various kinds of considerations, including business goals, target market, industry, level of competition, and marketing approach. Typically, businesses devote a portion of their revenue to advertising, but the recommended proportion differs by industry. Probabilities and trends in the future are what predictive analytics is all about. Predictive analytics is being used by companies to enhance their product catalogs, pricing strategies, and customer service (Puri et al., 2013). In the field of sales management, intelligent forecasting may play a crucial role. In order to estimate the rise of retail sales, intelligent forecasting analyses public relations data to determine which forms of advertising will be most successful. Though the process of forecasting tends to be complex, it is a straightforward technique to determine its accuracy (Pinki et al., 2018). We can predict how many sales it will have in the near future and what variables impact those sales with the use of machine learning (Martin et al., 2020). Also, forecasting is an important aspect of many business organizations today (Nguyen et al., 2013). Media platforms now provide a way for businesses to reach a wider audience and compete in the global marketplace. Businesses today need to know how to implement media platforms in a way that is consistent with their business plan (Mangold et al., 2009) because the tools and methods for interacting with customers have undergone significant transformations since the advent of social media platforms.

Linear regression is often used to predict sales for media platform advertising expenditure sales. This model plots two variables—advertising spending and sales—on a straight line to identify their connection. Advertising budget and sales are linear in the linear regression model. Advertising spending increases sales proportionally. The algorithm predicts sales based on the predicted advertising budget and past data. Thus, the linear regression model requires data on advertising budget and sales over time. A scatter plot and line of best fit are created from the data. By entering estimated advertising costs, the linear regression model can forecast sales.

Data mining is a kind of corporate analytics that goes beyond traditional approaches like counting, describing data, and reporting. It also, introduces data visualization, which is often the initial step into more complex analytics, but its primary emphasis is on more sophisticated methods of analyzing data. In particular, it encompasses automated decision-making aided by statistical and machine-learning techniques. The ability to predict outcomes is usually crucial, particularly at the personal level. Data mining's implementation speed has increased in the big data age. Due to their robustness and automation, data mining techniques can handle massive datasets and get value from them (Shmueli et al.,2017). Researcher found ensemble model performs better in comparison with other data mining techniques by exploring the application details of both paid and defaulter (Zurada, 2002). Also, analyzed the performance of 15 different classification methods with 23 different features of applicant data and found that linear regression was used to formulate as the final model (Turkson et al. ,2016) and found out taking a greater number of attributes will result in the model learning better using logistic regression techniques by statistically analyzing the distribution of features and prediction of the mode (Vaidya, 2017).

Thus, the proposed project in this final year project is to predict the sales for media platforms advertising expenditure using a Linear Regression algorithm that could be used to assist the business in prediction the sales based on advertising expenditures for media platforms which are television, radio, and newspapers.

1.2 PROBLEM STATEMENT

Ideally, the best advertising budget takes into consideration both the amount spent and its efficiency. When compared to a greater advertisement expenditure with poor execution, a well-planned advertising campaign that targets the correct audience and leverages the proper channels may be more successful. However, there are many times in the industry facing such problem.

Firstly, when deciding which advertising platforms will generate the most revenue and how much to invest in that channel to reach the target, the company looks at past trends in estimating advertising spend. Advertisers may not be represented accurately in this trend study of their advertising performance, which might result in wasteful expenditure and a poor return on investment. Since marketers may not be aware of changes in consumer behavior or trends within the industry, inaccurate trend analysis may also lead to lost growth opportunities and increased competition. Additionally, inaccurate trend analysis can make it difficult for advertisers to optimize their spending and allocate resources effectively, which can ultimately impact their bottom line. Trend analysis can have some potential disadvantages as a tool for making investment decisions. One of these disadvantages is that the accuracy of the analysis depends on the quality of the data being used. If the data is incomplete, inaccurate, or otherwise flawed, the analysis may be misleading or inaccurate (Adam, 2023).

Despite the fact that media platform advertising spending is becoming increasingly significant to businesses, there is still a sizable knowledge gap regarding how to predict sales based on advertising spending accurately. Executives are concerned with their budget justification for a social media plan, when there is a lack of supporting materials to confirm the effectiveness of the social media platform (in example the conversion rate, the relation between buyer-seller relationship, and increase in sales), and the rate of return on investment (ROI) that they can earn from this plan (Blanchard, 2011). Because of this lack of understanding, marketing funds may be allocated inefficiently, resulting in a poor return on investment (ROI). In order for companies to make informed judgments about their advertising budgets, further research is needed to determine the most efficient means of sales prediction for media platform advertising expenditure. There is presently no industry-wide standard for measuring the efficacy of advertising on media platforms due to a lack of standardized measurement and analysis. Despite holding companies' focus on attention metrics, there is still no industry measurement standard (Olivia, 2023). It may be challenging for businesses to assess the effectiveness of different kinds of media or to make accurate comparisons to their competitors' performance. Thus, selecting suitable variables and models is difficult. Finding the right factors to analyze and the right models to use may help with sales forecasting for media platform advertising budgets. It may be difficult for businesses to determine which factors are most relevant or which models to use for generating reliable sales forecasts. Organizations may spend millions of dollars on collecting and analyzing information with various data analysis tools, but many fall flat when it comes to actually using that data in actionable, profitable ways (Bernardita, 2023).

In the field of advertising, the complex web of interactions and interdependencies that exists across various channels and platforms is what gives conception to what is generally referred to as cross-channel impacts. The modern business environment is characterized by the fact that businesses are simultaneously advertising their products or services across a multiplicity of different platforms. The success of one channel may have a significant effect on the success of another channel. The management of many interactions is executed with decreasing levels of control, which poses a problem for marketers who are attempting to navigate this complicated and fragmented channel landscape. Because of this difficulty, multi-channel and omni-channel retailing strategies have shifted their attention to it (Lemon, 2016). Marketing through multiple channels has consistently outperformed expectations, with a notable 10 percent actual performance compared to the predicted 8 percent (Econsultancy, 2018). This highlights the significance of effectively coordinating marketing efforts across various digital and online channels, which further emphasizes the importance of effectively overseeing marketing efforts across these channels. Therefore, businesses are becoming more aware of the advantages of taking an integrated strategy to marketing efforts. This approach involves using complementary capabilities to compensate for weaknesses in current channels, and applying features that are compatible with one another to encourage interaction across many channels. For optimal development and execution of marketing strategies, an in-depth understanding of cross-channel dynamics is even more crucial, given the unique capabilities that are specific to each channel (Vroomen, 2007).

In conclusion, advertising expenditure management and sales forecasting are complicated. Trend analysis errors may waste money and miss growth possibilities, damaging the bottom line. Businesses struggle to evaluate media platform advertising and allocate resources due to the absence of standardization. Lack of supporting records makes it difficult for executives to explain social media initiatives, which may lead to poor budget allocations. Marketing's complex cross-channel effects make managing various touch points across platforms much more difficult. Companies recognize that efficient advertising strategies need an integrated strategy, complementary capabilities, and an in-depth understanding of cross-channel dynamics. Further study is needed to build industry-wide standards and reliable predicted sales models so businesses can make intelligent choices and optimize their advertising expenditure.

1.3 OBJECTIVES

The purpose of this proposal is to develop a system that forecasts sales in media platforms' advertising expenditure. These are three objectives of this project:

1. To study linear regression algorithm in sales prediction for media platforms' advertising expenditure.
2. To develop a prototype of sales prediction for media platforms' advertising expenditures using the linear regression algorithm.
3. To evaluate accuracy of linear regression algorithm in predicting the sales for the media platform's advertising expenditure.

1.4 SCOPE OF RESEARCH

This proposed project intends to assist the business organizations such as advertisers in using linear regression to predict sales for media platforms' advertising expenditure. The system developed from this project will be implemented for any business that advertises on media platforms and wants to improve its marketing effectiveness. This could include businesses of any size and industry that use advertising to promote their products or services, such as e-commerce companies, B2B businesses, brick-and-mortar retailers, and more. The data requirements that will be needed for this research is the past data of sales in and advertising expenditure for media platforms such as social media, radio, television and type of influencer collaboration. The result from the system will display the outcome on the sales based on the advertisement expenditure for each media platform. In addition, the system will be implemented throughout all business industries and made accessible to authorized users.

For the purpose of this study, datasets have been obtained from the online source, Kaggle. The dataset has 4573 samples based on the media platform, which is amount of advertising expenditure on television, amount of advertising expenditure radio, amount of advertising expenditure social media, type of influencer collaboration. The goal of this dataset is to predict the sales based on the advertisement expenditure for specific media platforms.

1.5 SIGNIFICANCE OF THE STUDY

This study addresses important problems in the field of sales prediction and could benefit the business industry, considering that the number of sales and advertisement expenditure keeps increasing every year. Better cost management in businesses can optimize their advertising budgets and allocate resources more effectively, reducing waste and improving return on investment (ROI). Thus, the development of a prototype that is derivable from this research might result in accuracy in predicting the sales for media platforms advertising expenditure. Increased accuracy in the sales forecast may help executives make choices across departments, given the large improvements in the forecast at the overall sales level (Cui et al.,2018). Using past sales data and advertising expenditures, linear regression models may assist companies in making precise prediction of sales. Sales prediction also can help businesses plan for future expansion and development. Businesses can make better decisions regarding employment, product development, and future investments if they are aware of their expected revenues and profits. Other than that, this proposed research also can benefit from identifying the most effective advertising channels and tactics. By analyzing historical data on advertising expenditures and sales, businesses can identify which media platforms and types of advertising have the highest correlation with sales. This information can then be used to inform future advertising decisions and campaigns. Last but not least, the simplicity of linear regression is one of its primary benefits. It is a simple and uncomplicated method that can be utilized by businesses of any size and level of data analysis expertise. This analysis is useful for a small firm since it reveals which elements are most significant, which can be ignored, and how they interact with one another. This analysis also is a strong statistical tool that enables a company to investigate the connection between two or more variables of interest (Leon, 2019). The results of linear regression models are simple to interpret and communicate to key stakeholders.

1.6 OVERVIEW THE OF RESEARCH FRAMEWORK

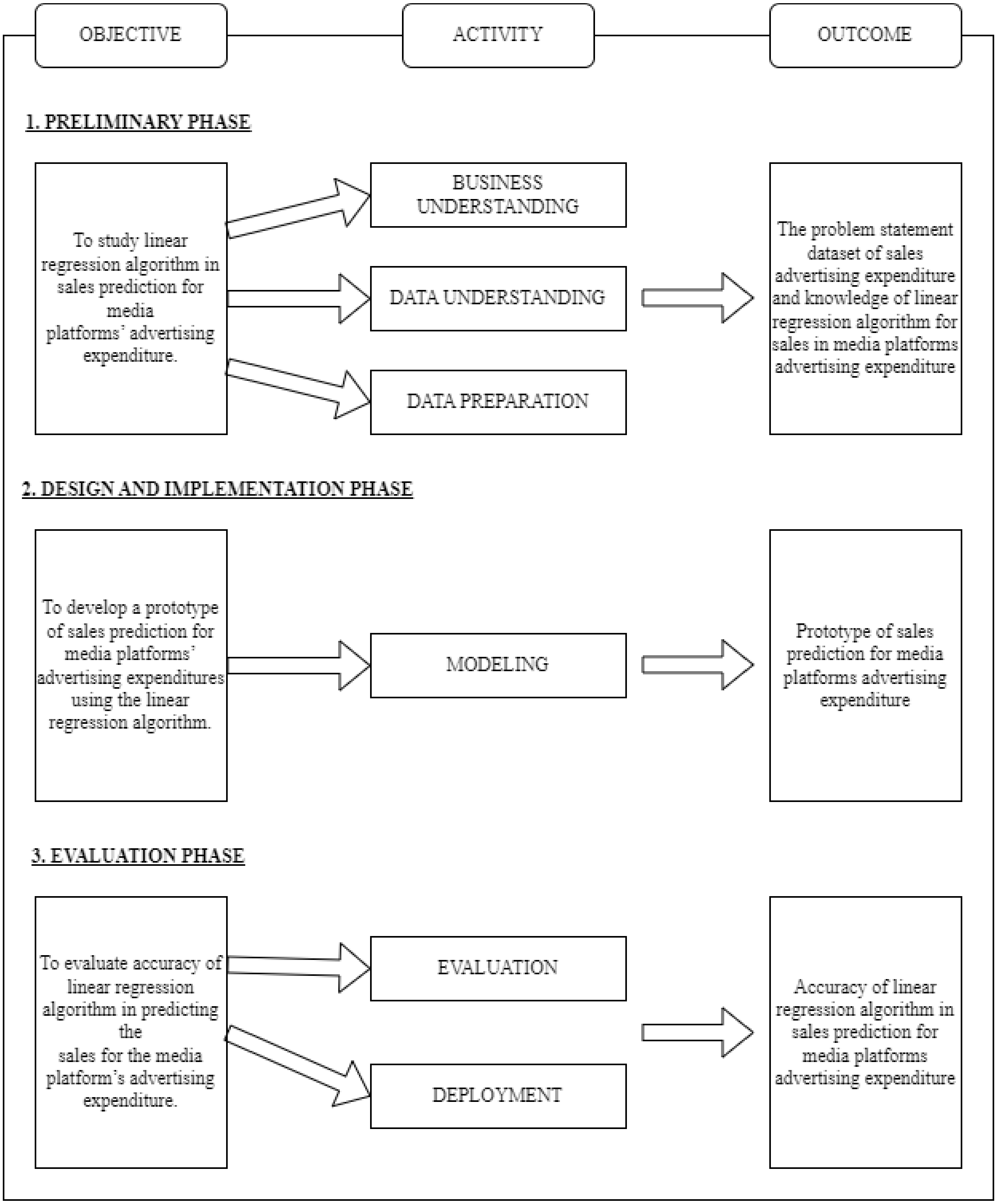

OBJECTIVE
ACTIVITY
OUTCOME
1. PRELIMINARY PHASE
To study linear regression algorithm in sales prediction for media platforms' advertising expenditure.
BUSINESS UNDERSTANDING
DATA UNDERSTANDING
DATA PREPARATION
The problem statement dataset of sales advertising expenditure and knowledge of linear regression algorithm for sales in media platforms advertising expenditure
2. DESIGN AND IMPLEMENTATION PHASE
To develop a prototype of sales prediction for media platforms' advertising expenditures using the linear regression algorithm.
MODELING
Prototype of sales prediction for media platforms advertising expenditure
3. EVALUATION PHASE
To evaluate accuracy of linear regression algorithm in predicting the sales for the media platform's advertising expenditure.
EVALUATION
DEPLOYMENT
Accuracy of linear regression algorithm in sales prediction for media platforms advertising expenditure

Figure 1.0 Overview of the research framework

1.7 CONCLUSION

Based on past research conducted on sales prediction for media platforms advertising expenditure using linear regression, it was discovered that this strategy works quite effectively in predicting the sales (Mitra et al.,2022). In order to determine trends, we compared sales figures to advertising budgets over time. Linear regression was used to determine which factors had the most bearing on sales. Overall, the results of the study indicate that linear regression is a useful tool for predicting sales in media platforms' advertising expenditure and that this information may be utilized by companies to plan advertising expenditure accordingly (Singh et al., 2020). However, it should be remembered that the quality of the data utilized and the assumptions used during the research will determine the accuracy of the projections. Therefore, organizations should adopt a holistic approach to sales prediction, considering a variety of approaches and considerations before making any judgements.

CHAPTER 2: LITERATURE REVIEW

This chapter will discuss the significance of literature and research on sales prediction for media platform advertising expenditure systems. It will cover the introduction of sales prediction for media platform advertising expenditure in the first section. Then, the second section will discuss the forecasting modeling area, followed by the description and explanation of the algorithm used in this proposed project, in the third section. This chapter will also discuss the implementation of prediction systems in various areas and similar works of the proposed project.

2.1 MEDIA PLATFORM FOR ADVERTISING

For the purpose of achieving marketing goals over an extended period of time, various kinds of media including television, radio, and newspapers have been used. Every medium presents its own one-of-a-kind set of benefits and communicates with a distinct subset of the population.

Firstly, advertising on television enables businesses to communicate with a wide range of audiences through audio-visual content. It has the capacity to captivate audiences and leave an everlasting mark. In order to maximize their reach, advertisements on television might be targeted to certain demographics and shown during popular programming. According to research that was published by Statista in 2020, the total amount spent on television advertising in the United States amounted to $68.5 billion, showing its continuous significance in the landscape of advertising (Michel, 2023). Even though early television advertising contributed a significant amount of impact to advertising, television is still obtaining greater advantages to company advertising nowadays. The fact that television advertising is still considered a major medium for reaching a large audience is the primary factor that contributes to its continued selection as the media platform for advertising in this research. Television advertising continues to develop, even though conventional TV viewing has become more dispersed as a result of the development of streaming services and on-demand programming. The fact that expenditure on television advertising in the United States is expected to reach $62.35 billion in 2023, as stated by eMarketer (eMarketer, 2021) shows that it will continue to be an important kind of marketing. Advertisers may now reach audiences beyond those that can be reached via conventional TV channels by placing commercials on streaming platforms, video-on-demand services, and targeted advertising on smart TVs. These options were previously unavailable.

Radio advertisements are commercials that play on radio stations. It is useful for reaching a targeted demographic or interest group and may help you connect with people in your immediate area. Over 90% of Americans listen to radio at least once every week, as stated by the Radio Advertising Bureau. The Radio Advertising Bureau (RAB) also claims that airing radio commercials many times throughout the day to the same demographic is very efficient. With the growing popularity of internet radio and audio streaming services, radio commercials have gone digital. Marketers may reach certain demographic subsets or geographical areas of listeners via these platforms' tailored advertising options. Revenues from digital audio advertising are projected to rise (Deloitte, 2021). This increase is likely to be driven by programmatic audio advertising and podcast advertising. The paper highlights the usefulness of digital audio platforms in addressing certain demographics, as well as their rising popularity.

Finally, it should be acknowledged that social media advertisements, which were previously limited to digital platforms, are currently expanding their scope to include a wider demographic. In contrast to the decreased overall implementation of standard methods of advertising, social media platforms provide a constantly developing environment for targeted advertising (Pew Research Centre, 2020). A significant majority of Americans, specifically 14%, indicated that they regularly obtained their news exclusively from digital sources in 2020. Digital-only channels consist of social media platforms and the applications that are related to them. This emphasizes the ongoing significance of social media in satisfying the informational needs of diverse audiences (Pew Research Centre, 2020).

While there are now more places to place advertisements thanks to the rise of digital media, certain markets and demographics prefer the tried-and-true methods used by traditional media. Considerations including demographics, campaign goals, available funds, and potential exposure all play a role in settling on a certain medium. Broader and more efficient reach may frequently be achieved by integrated advertising strategies that use both conventional and digital media. Due to their large user populations, powerful targeting options, and quantifiable outcomes, social media platforms, search engines, and other digital channels are becoming more popular for advertising. Advertisers often utilize these mediums to communicate with targeted audiences, stimulate user interaction, and measure campaign efficacy.

2.1.1 SALES IN MEDIA PLATFORM ADVERTISING EXPENDITURE

Businesses increasingly rely on media platforms to reach their target audiences and promote their products or services in the current digital era. Media platform advertising expenditure refers to the quantity of money spent on advertising across multiple digital channels, such as social media, search engines, online video platforms, and mobile applications. Over the past decade, this form of advertising has acquired immense popularity and experienced significant growth. Several factors contribute to the development of media platform advertising. Firstly, the pervasive adoption of digital technologies has altered the manner in which individuals consume content, resulting in a significant transition toward online platforms. Therefore, businesses have recognized the need to establish a robust online presence in order to remain competitive and effectively engage their target customers.

In the constantly evolving field of advertising, businesses are gradually tapping into the potential of digital platforms such as television, radio, and social media to establish connections with their targeted consumers. These channels have unique benefits, allowing organizations to create customized messaging using demographic knowledge, interests, and behavioral patterns. Such reliability improves budget allocation, ensuring that marketing activities have the greatest possible impact in the highly competitive digital environment.

An important trend influencing advertising on media platforms is the collaboration with influencers, which includes high-profile and large-scale influencers to smaller and more focused influencers. This collaborative strategy extends through several channels, including television, radio, and social media. Businesses aim to enhance their brand messaging and encourage more audience engagement by using the convincing reach of people. This approach demonstrates a significant impact in effectively targeting specific audiences and maintaining real connections with customers across many media channels. Understanding the changing patterns of advertising spending and the collaborations between influencers and media platforms becomes essential for companies aiming to succeed in the digital age. It is important to stay updated and well-informed by using trustworthy sources, research on the industry, and market research to make educated decisions and effectively adapt to evolving trends.

Factors like the economy, new technologies, shifts in consumer behavior, and developing advertising tactics may cause changes in both the total amount of sales and the patterns in advertising spending across media platforms. The best way to stay abreast of this rapidly evolving field is to regularly consult reliable resources, such as industry studies, market research journals, and advertising and marketing news outlets.

2.2 DATA MINING

Data mining is an iterative process, the progression of which is determined by the finding of new information using either automated or human approaches. In the context of exploratory analysis, when there are no predetermined preconceptions about what would constitute an "interesting" output, data mining is at its most beneficial. The process of mining huge amounts of data with the purpose of discovering new, useful, and nontrivial information is known as "data mining." It is the collaborative attempt of human experts to describe problems and objectives using computer search capabilities (Wiley, 2011). In the hands of decision-makers, data mining technology is now a hot favorite since it may offer important hidden commercial and scientific "intelligence" from enormous amounts of historical data (Wiley, 2011). The following illustrative instances illustrate some of the numerous areas in which it has been shown to be very effective and beneficial for business organizations. Many of these commercial data mining applications aim to address an urgent issue or investigate a new business opportunity in order to develop a long-lasting competitive edge. In the retailing industry, data mining can be used to predict accurate sales volumes to determine correct inventory levels, identify sales relationships between different products, forecast consumption levels of different product types to optimize logistics and maximize sales, and discover purchasing patterns.

Prediction, and description are the two most important uses of data mining. Prediction is the process of utilizing information from one or more fields within a dataset to make assumptions on the most likely or predicted values of another variable. In contrast, the descriptive analysis looks for human-understandable patterns in the data. On the predictive end of the spectrum, the objective of data mining is to generate a model, expressed as executable code, that can be used for classification, prediction, estimation, and other similar tasks. The objective of descriptive analysis, on the other hand, is to acquire a comprehension of the analyzed system by identifying patterns and relationships in large data sets. The relative relevance of prediction and description can differ substantially across data mining applications.

2.2.1 PREDICTIVE ANALYTICS AND MODELING

Predictive Analytics

Data mining and predictive modeling come together in predictive analytics data mining; an area of research that may predict future events based on past data. Predictive analytics is the practice of mining existing data for insights and developing new models to predict outcomes.

Following are explanations of methods under predictive analytics data mining:

Table 2.1 Explanations of methods

Method	Explanation
Classification	Classification is a predictive algorithm that assigns data instances to predefined classes or categories based on their features. It aims to learn a decision boundary that separates different classes in the data space. Commonly used classification algorithms include Decision Trees, Random Forests, Support Vector Machines (SVM), and Naive Bayes. Classification algorithms aim to solve problems such as spam detection, sentiment analysis, customer churn prediction, and disease diagnosis. They learn from labeled training data to build a model that can classify new, unseen data (Hastie et al., 2009)
Regression	Regression is a predictive algorithm used to estimate or predict a continuous target variable based on one or more input variables. It models the relationship between the input variables and the target variable by fitting a mathematical function to the data. Linear regression, polynomial regression, and ridge regression are common regression techniques. Regression algorithms are used to solve problems such as sales forecasting, housing price prediction, and demand estimation. They learn from historical data with known target values to create a model that can make predictions on new data (James et al., 2013)

Table 2.1 Explanations of methods (Continued)

Prediction	In the context of predictive algorithms, prediction refers to the general process of using historical data to forecast future outcomes. It encompasses both classification and regression tasks. Predictive algorithms learn patterns from historical data to make predictions or forecasts about unknown or future data points. Prediction algorithms are used in various domains, including finance, weather forecasting, stock market analysis, and sales forecasting. They help organizations make informed decisions and plan for future events (Shmueli et al, 2019)
Time Series Analysis	Time series analysis focuses on predicting future values based on historical data collected over time. It considers the temporal order and dependencies in the data to make forecasts. Time series algorithms, such as ARIMA (AutoRegressive Integrated Moving Average), exponential smoothing, and recurrent neural networks (RNNs), are commonly used. Time series analysis is used in financial forecasting, demand forecasting, stock market prediction, and other applications involving temporal data. It helps identify trends, seasonality, and patterns in time-dependent data to make accurate predictions (Hyndman et al, 2018)

Predictive Modeling

Predictive modeling is a statistical technique for predicting future behavior or outcomes that is widely used. Predictive modeling solutions are a type of data-mining technology that analyses past and current data to create a model that can be used to forecast future results (Edwards, 2019). A predictive model is not static which is it will be updated or evaluated on a regular basis to adjust for changes in the primary data. This also means that it is not a prediction model that will be complete at once. Predictive models make predictions based on what has already occurred and what is currently occurring. If the recent data shows that what is currently happening has changed, the expected future outcome also will be affected and must be reviewed (Ali, 2020). The model is built and trained over time to appropriately respond to new data (Prasad et al., 2019).

The predictive modeling process begins with data gathering, followed by the formulation of a statistical model, prediction, and revision of the model when new data becomes available. To increase underwriting accuracy, risk models can be constructed that combine member information in various ways with demographic and lifestyle data from external sources. Predictive models look at past behavior to determine how likely a client is to repeat it in the future. Most prediction models are quick to respond and frequently complete calculations in real-time (Ali, 2020).

2.3 LINEAR REGRESSION ALGORITHM

Linear Regression and Multiple Linear Regression

This is a well-known regression technique as well as one of the most common machine learning modeling strategies (Sarker, 2021). The dependent variable is continuous in this technique, the independent variables might be continuous or discrete, and the regression line is linear. Linear regression uses the best-fit straight line to build a connection between the dependent variable (Y) and one or more independent variables (X) which is also known as a regression line. Equations (1) - (2) show two different formulas of linear regression.

$$y = a + bx + e\text{, and} \tag{1}$$

$$y = a + b1x1 + b2x2 + bnxn + e \tag{2}$$

These equations define where a denotes the line's intercept, b denotes the slope, and e denotes the error term. Based on the predictor variable, these equations can be used to predict the target variable's values. Multiple linear regression is a variation of simple linear regression in which two or more predictor variables 16 are used to model a response variable, y, as a linear function, whereas basic linear regression only includes one independent variable (Sarker, 2021).

2.3.1 EVALUATION OF LINEAR REGRESSION ALGORITHM

Table 2.2 Formula and Explanation of Linear Regression

Formula	Explanation
R Square/Adjusted R Square: $R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$ (Source TechNotes, 2019)	Evaluating scores between 0 and 1. If the score is close to 1, we think the model accurately represents the data. The model's fit to the dependent variables may be assessed using R Square. However, overfitting is not regarded. Due to its complexity, a regression model with many independent variables may perform well on training data but badly on testing data. It is best to evaluate from all angles. Discuss R2's genuine meaning. The total of squared prediction error is divided by the sum of squares that replace the calculated forecast with the mean to produce R2. The projected and actual numbers match better with a higher R Square value (TechNotes, 2019).

Table 2.2 Formula and Explanation of Linear Regression (Continued)

Mean Square Error (MSE): $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ (Source TechNotes, 2019)	Mean Square Error is an absolute measure of model fit. MSE is determined by summing the squares of prediction error. Where prediction error is the difference between genuine values and predicted values, we avoid negative error scores by squaring the error. Its output indicates the deviation from the genuine number. It may have a larger number, which is comparable to being rare. You may wonder why the error score is so high (TechNotes, 2019).
Root Mean Squared Error (RMSE): $RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$	Assessing model fit using Root Mean Square Error (RMSE) is crucial. It summarises prediction error squares to show the difference between predicted and actual values (TechNotes, 2019). RMSE is similar MSE, except it takes the square root of the error, making the score more interpretable in the original data's units. RMSEs with larger differences from real values indicate uncommon or unexpected events. A score around 1 indicates the model correctly fits the data in RMSE.
Mean Absolute Error (MAE): $MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$ (Source TechNotes, 2019)	This metric is nearly identical to Mean Square Error, with the exception that MAE uses absolute error value instead of the square of predicted error to prevent a negative score. However, we do not need to calculate Root of MAE score in this case. We can interpret the score using actual values (TechNotes, 2019).

2.3.2 ADVANTAGES OF LINEAR REGRESSION ALGORITHM

Linear regression is simple and easy to understand since it uses a basic method. Coefficients for each feature are included in the model's output, directly representing the connection and impact of the characteristics on the dependent variable (Draper and Smith,1981). Aside from that, linear regression's processing efficiency makes it a good choice for large datasets with numerous properties. Since the Normal Equation provides a closed-form solution to the method, the model parameters may be quickly determined (Hastie et al., 2019). Lastly, linear regression offers statistical measurements, such as p-values and confidence intervals, that enable testing hypotheses and evaluating the importance of attributes through inference. This may be useful for determining the model's validity and reliability (Montgomery et al, 2012).

2.3.3 IMPLEMENTATION OF LINEAR REGRESSION ALGORITHM IN VARIOUS PROBLEMS

The authors of this research, Siddhant et al. (2021), are now undertaking research on how linear regression might be used to predict housing prices. This study reveals that purchasing a home in Mumbai is challenging. Multiple real estate agents, a list of needs, and property comparisons require considerable time. Without knowledge of the city's price range, newcomers may be misled. Consequently, a model that precisely determines the price of a property is required. Due to this issue, the author intends to develop a model for predicting the price of real estate in Mumbai using a linear regression algorithm. As their objective, they take into account the number of accommodations and the availability of various kinds of facilities. As a consequence of this study, the R2-score of the prediction model was 0.8643. Our model explains 86% of the volatility of the dependent attribute/variable, leaving 14% unexplained. This algorithm is capable of predicting property values in additional Indian cities and rural areas.

Bum et al. (2019) are conducting research on Prediction of Student's Academic Performance Utilizing Linear Regression. This study's problem is that college students drop out or earn degrees below their intellectual capacity. Some prospective college students may be oblivious to the distinctions between degree levels and the amount of work required to obtain them. Academic performance forecasting is necessary for making informed decisions. Consequently, the purpose of this study is to predict the performance of Benue State University mathematics/computer science students using linear regression. Predict who will fail a test and who will pass, along with the quantity of failures or successes. As a result of this objective, we discovered that our suggested strategy predicted the student's subsequent outcomes with an accuracy rate of up to one hundred percent. This strategy has the potential to substantially increase university academic achievement if implemented correctly.

Author Age Prediction from Text Using Linear Regression was published by Nguyen et al. in 2011. Using a linear regression algorithm, the purpose of this study is to investigate the relationship between the linguistic patterns or content of a text and the author's age in order to develop a method that is accurate and reliable. Sociolinguistics investigates the relationship between the way individuals use language and their group membership. There are numerous methods to define a group, including age, gender, socioeconomic status, and political affiliation. A person is a member of multiple communities, so their identity and language are influenced by numerous factors. We discovered correlation and mean absolute error as efficient evaluation metrics as a consequence of this endeavor. The range of mean absolute errors was between 4.1 and 6.8 years, and the correlations reached 0.74. Content and aesthetic characteristics were significant indicators of age in three corpora. Numerous POS patterns indicate advanced age, and a baseline containing only unigrams performs admirably.

The authors of the study Linear Regression Approach to Predict Crop Yield are Murugan et al. (2020). The technique employed is linear regression. The reason why the investigation was conducted is because as a result of global warming and pollution, the quality and production of commodities are severely diminished. The strategy for this initiative seeks to address the issue of cost loss. where the majority of farmers are oblivious of the novel crops that can increase yields and profits if they are grown there. Consequently, the research objective is to develop a linear regression-based model that can predict harvest yield in agricultural settings. The purpose of the study is to determine how various factors, such as weather, soil quality, and crop management practices, influence food growth. Using linear regression analysis, agricultural production variability factors are identified and an effective prediction model is developed. At the conclusion of this study, data cleaning and feature selection separate input and output features. The algorithm then anticipates the optimal temperature and humidity for crop cultivation. Seasonal and climatic range values fluctuate from year to year, necessitating the application of a regression technique.

Seethalakshmi (2018) is the author of Analysis of stock market predictor variables using Linear Regression. Numerous variables impact the stock market's complexity and unpredictability, thereby necessitating this investigation. Analysts, investors, and policymakers are required to comprehend the relationship between predictor factors and stock market performance. The difficulty lies in precisely identifying and measuring the effects of these predictive factors on stock market movements. then, conduct objective investigation Using Linear Regression, evaluate predictor factors and stock market performance. This article examines the best independent predictors of stock market closing values. This study determines which variables have the greatest impact on closing pricing. Consequently, the conclusion of this study is that Comprehensive Model 1 has an R2 value of 0.997%. Open, high, low, volume, and adj close are required to predict closing value accurately. With an R2 value of 0.992%, Model 2 predicts proximity. Adjust close has no effect on the forecast close value. This study demonstrates that open, high, low, and volume can approximate value.

Table 2.3 Implementation of Linear Regression Algorithm in Various Problem

No.	Title	Technique/ Algorithm	Objective	Problem	Result	Reference
1.	Housing Price Prediction Using Linear Regression	Linear Regression	Using a linear regression algorithm, develop a model for predicting the price of real estate in Mumbai. While doing so, the number of accommodations and the availability of various types of facilities are taken into consideration.	In Mumbai purchasing a house is difficult. Multiple real estate agents, a list of requirements, and property comparisons take time. Newcomers may be misled since they don't know the city's price range. Therefore, a model that accurately determines a property's price is needed.	Our prediction model's R2-score was 0.8643. Our model explains 86% of the dependent attribute/variable's volatility, leaving 14% unaccounted for. This algorithm can anticipate property values in more Indian cities and rural regions.	(Siddhant et al., 2021)
2.	Prediction of Student's Academic Performance Using Linear Regression	Linear Regression	Using linear regression to predict Benue State University mathematics/computer science student performance. Predict who will fail and who will pass a test, as well as the number of failures or successes.	Tertiary students drop out or get degrees below their intellectual capacity. Some incoming college students may not even be aware of the distinctions between degree levels or the amount of work required to get them. To make educated decisions, academic performance prediction is required.	We found that our suggested strategy had an accuracy rate of up to 100% in predicting the following outcomes of the student. If implemented correctly, this approach has the potential to significantly raise academic achievement at universities.	(Bum et al., 2019).

Table 2.3 Implementation of Linear Regression Algorithm in Various Problem (Continued)

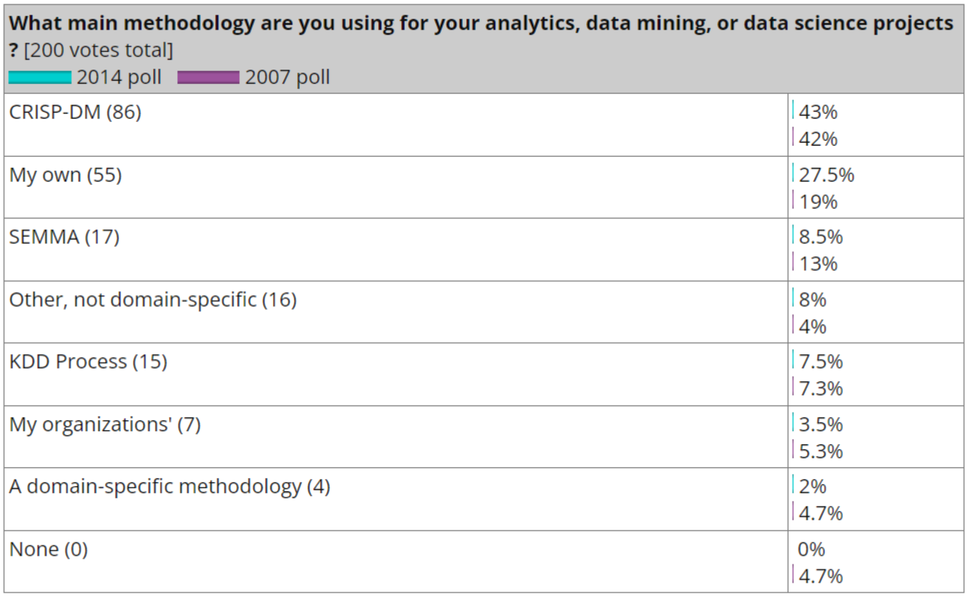
3.	Author Age Prediction from Text using Linear Regression	Linear Regression	The purpose of this study is to investigate the relationship between the linguistic patterns or content of a text and the author's age by using a linear regression algorithm in order to develop a method that is accurate and reliable.	Sociolinguistics tries to figure out how the way people use language is related to their participation in a group. A group can be defined in many ways, such as by age, gender, financial level, and political affiliation. A person is a part of more than one community, so their identity and words are affected by many things.	We found a correlation and mean absolute error efficient evaluation metrics. Mean absolute errors ranged from 4.1 to 6.8 years and correlations reached 0.74. In three corpora, content, and aesthetic traits were significant markers of age. Many POS patterns indicate old age and even a unigram-only baseline performs well.	(Nguyen et al., 2011)
4.	Linear Regression Approach to Predict Crop Yield	Linear Regression	To develop a predictive model that can predict crop yield in agricultural settings using linear regression. The objective of the study is to find out how different things, like weather conditions, soil qualities, and crop management practices, affect food growth. Linear regression analysis is used to find agricultural production variability factors and create an effective prediction model.	Due to global warming and pollution factors, there is limited or much decrease in the quality and production of crops. This project's strategy aims to address the issue of cost loss. where the majority of farmers are unaware of the new commodities that can increase yields and profits if grown in that region.	Data wrangling cleans the data and feature selection splits input and output features. After that, the algorithm predicts the ideal climate and humidity for crop cultivation. It is necessary to use a regression technique due to the fact that seasonal and climatic range values change from year to year.	(Murugan et al., 2020)

Table 2.3 Implementation of Linear Regression Algorithm in Various Problem (Continued)

5.	Analysis of stock market predictor variables using Linear Regression	Linear Regression	To evaluate predictor factors and stock market performance using Linear Regression. This article investigates the best independent factors to predict stock market closing values. This research determines which variables most affect closing prices.	Numerous variables affect the stock market's complexity and unpredictability. Analysts, investors, and policymakers must understand the link between predictor factors and stock market performance. The problem is precisely identifying and measuring these predictive factors' effects on stock market movements.	All-featured Model 1 with R2 value 0.997. Open, high, low, volume, and adj close are needed to correctly anticipate closing value. Model 2 predicts close with R2 = 0.992. Adj close does not alter the close value forecast. This research shows that open, high, low, and volume may estimate near value.	(Seethalakshmi et al., 2018)

2.4 CRISP-DM

The number of data mining (DM) and knowledge discovery (KD) initiatives has grown significantly. As DM and KD development projects became more complicated, project planning delays, low productivity, and failure to meet customer expectations emerged. KD and DM histories are similar. When Information Discovery in Databases (KDD) processing was introduced in the early 1990s, there was a rush to design DM algorithms that could solve all the problems of searching for knowledge in data. DM algorithms were designed and tools were made to implement them. CRISP-DM's publication in 2000 was the most important turning point in DM and KD process models (Marbán et al., 2009). A collaboration of DaimlerChrysler, SPSS, and NCR created Cross-Industry Standard Process for Data Mining (CRISP-DM) (Azevedo and Santos, 2008). It remains the "de facto standard" for data mining and knowledge discovery initiatives, according to several surveys and user polls (Plumed et al., 2021). The most popular approach for designing DM projects is CRISP-DM (Marbán et al., 2009). Based on the KDnugget website, Figure 2.1 shows that CRISP-DM was the most utilized approach for analytics, data mining, and data science (KDnugget, 2014).



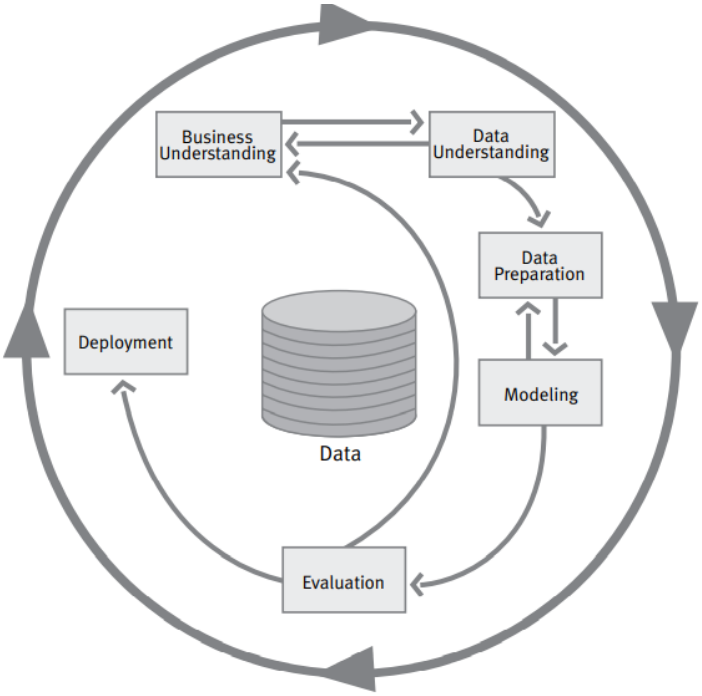
What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]
🟦 2014 poll 🟪 2007 poll

CRISP-DM (86)	43% 42%
My own (55)	27.5% 19%
SEMMA (17)	8.5% 13%
Other, not domain-specific (16)	8% 4%
KDD Process (15)	7.5% 7.3%
My organizations' (7)	3.5% 5.3%
A domain-specific methodology (4)	2% 4.7%
None (0)	0% 4.7%

(Source KDnugget, 2014)

Figure 2.1 The Polls of Most Used Methodology

CRISP-DM specifies the steps that must be completed in a DM project. CRISP-DM also specifies the tasks and deliverables for each phase. CRISP-DM is broken down into six stages. Based on Figure 2.2, the stages discussed are business understanding, data understanding, data preparation, modeling, evaluation, and deployment



(Source Chapman et al., 2000)

Figure 2.2 Phases of the CRISP-DM Methodology

Hierarchical process models describe CRISP-DM. Phases, generic tasks, specialized tasks, and process instances are organized from general to specific in this approach. CRISP-DM hierarchical model in Figure 3.2. The CRISP-DM hierarchical process model's stages split the data mining process into small parts, as seen in the picture. Each phase contains many second-level generic tasks. The second level is general since it covers all data mining situations. General responsibilities should be thorough and stable. Complete means incorporating all data mining processes and applications. Stable means the model should apply to unforeseen changes like new modeling methods. The third level, the specialized task level, defines generic task activities for specific contexts. Phases and tasks as discrete stages performed in a prescribed order constitute an idealized sequence of operations. Many of the tasks may be done in a different sequence, and it may be necessary to repeat procedures. The fourth level, the process instance level, records actual data mining engagement actions, choices, and results. A process instance is organized by upper-level tasks but illustrates what really happened in a specific encounter.

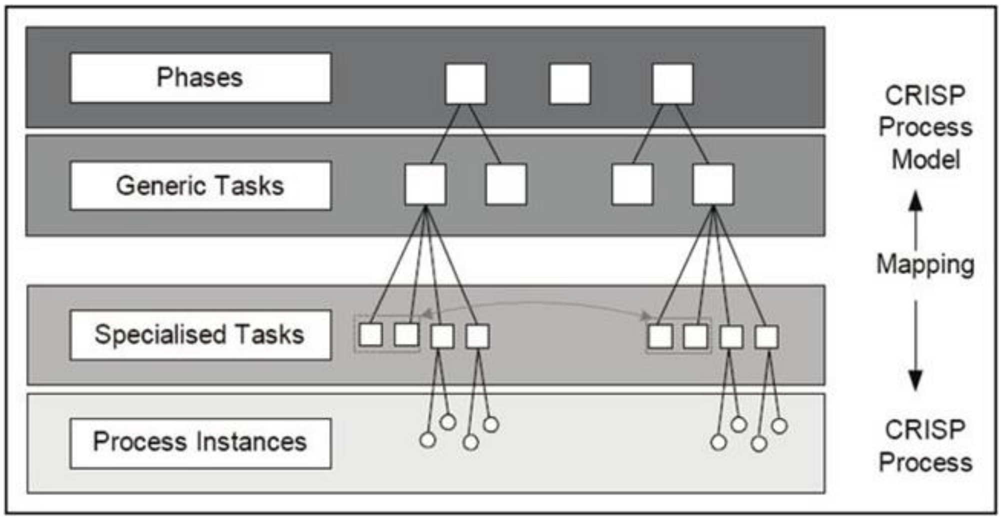


Figure 2.3 CRISP-DM Hierarchical Process Model

2.4.1 BUSINESS UNDERSTANDING

As a component of the CRISP-DM framework, business understanding refers to the process of obtaining a thorough understanding of the business problem, objectives, and organizational factors. It involves working together with different stakeholders, experts on the subject area, and professionals in the relevant domain to set clear corporate objectives and scope. This knowledge helps align analytical efforts with business objectives, guaranteeing that data-driven choices are made based on a full grasp of the business environment. Aligning analytical efforts with business objectives helps ensure that businesses achieve their goals (Wirth et al., 2000).

2.4.2 DATA UNDERSTANDING

Data understanding, which is a component of the CRISP-DM framework, refers to the investigation and familiarization with the available data sources. It entails gathering the data and studying it in order to get insights on the quality of the data, the relevance of the data, and the possibility for analysis that it offers. Organizations are able to make educated judgments on the usefulness of the data for the project's objectives if they have an accurate understanding of the data, its structure, and the limits of the data (Wirth et al., 2000).

2.4.3 DATA PREPARATION

Critical to the CRISP-DM framework, data preparation entails collecting, cleansing, integrating, and transforming data to ensure its suitability for analysis. Data collecting, data cleaning to manage missing or inconsistent information, data integration to integrate different data sources, and data transformation to turn data into a suitable format for analysis are all examples of activities that are included in this phase. In order to provide more efficient analysis and modelling in following phases of the CRISP-DM framework, the objective of the data preparation step is to confirm that the data are true, comprehensive, and presented in the appropriate format (Wirth et al., 2000).

2.4.4 MODELING

Modelling is a crucial phase of the CRISP-DM framework, in which data is analyzed and algorithms are utilized to develop predictive or descriptive models. During this phase, you will pick and use relevant modelling approaches, refine model parameters, and evaluate the performance of the model. The objective here is to create models that faithfully reflect the connections contained within the data and that may be used for making forecasts or gaining new perspectives of models that faithfully reflect the connections contained within the data and that may be put to use for making forecasts or gaining new perspectives is the objective here. (Wirth et al., 2000) The goal of modelling is to maximize the predictive or descriptive capacity of the models, ensuring that they are dependable and helpful to find the purpose of finding a solution to the business issue at hand.

2.4.5 EVALUATION

The evaluation phase is an essential component of the CRISP-DM framework. During this step, the functionality and efficiency of the models that have been generated are evaluated. During this stage, the models are validated by being tested on an external dataset, the accuracy of the models is measured, and the models are compared to a set of predetermined assessment criteria. The purpose of the assessment is to ascertain whether the models can produce accurate results when applied to actual-life situations and whether they align with the company's goals the models are capable of producing accurate results when applied to actual-life situations and whether or not they are in line with the goals of the company. Evaluation assists in finding any flaws or areas for improvement in the models, allowing for modifications or changes to be made before their deployment in the operational environment (Wirth et al., 2000). Evaluation also helps in determining whether a model is fit for its intended purpose.

2.4.6 DEPLOYMENT

Deployment is the final phase in the CRISP-DM architecture for data mining, and it involves putting the models that have been generated into an actual production setting. During this step, the models are integrated into the current systems, and it is made sure that they are correctly deployed and are operating as the designers planned. In addition, the provision of user documentation, training, and support as part of the deployment process is necessary to facilitate the end-user's easy uptake and utilization of the models. The purpose of the deployment is to put the models into use so that the organization may benefit from their predictive or descriptive capabilities and draw actionable insights for decision-making (Wirth et al., 2000). The objective of deployment is to put the models into action.

2.5 SIMILAR WORKS

One of the similar work is the research was carried out by Subha, (2020). This study is being conducted with the objective of predicting the impact that advertising on social media will have on the income of businesses. The purpose of this research is to construct a linear regression model with the intention of using YouTube marketing dollars in order to forecast sales. The method or algorithm that is being used in this study is known as linear regression. The findings of this study indicate that advertising on YouTube is a more reliable indicator of business sales than any other factor.

Research goals by Cong et al. (2021) attempts to analyze how Twitter is used by corporations and whether or not it can be used to predict business outcomes. The study uses statistical modeling and regression analysis to investigate the connection between Twitter usage and metrics of business performance including sales, revenue, and customer engagement. To what extent can businesses utilize Twitter to reach their target audience and advance their goals is the focus of this research. This study's findings suggest that advertising on Twitter might be beneficial to companies' bottom lines. In addition, a new statistical model, the exponential T-X exponentiated exponential, is introduced.

Researchers Lin et al. (2021) compared the effects of newspaper and Facebook ads on sales. Linear Regression is the chosen methodology for this study. the study's goal is to analyze using a technique in statistics known as simple linear regression modeling. Certain statistical tests are used to examine the importance of the advertisement for sale. This research found that ads may significantly affect financial outcomes. Spending money on advertising has been found to boost return on investment by a factor of four compared to doing nothing at all. It is more effective to advertise on the Internet than in paper. Income from advertising on the Internet is three times that from print.

Kohli et al. (2021) publish an article titled Sales Prediction Using Linear and KNN Regression. We use linear regression and the K-nearest neighbors (KNN) method in this study. The goal of this study is to evaluate the relative merits of the linear regression and K-nearest neighbors (KNN) regression models for making sales forecasts. This study concludes that KNN regression is a classic case of an overfitting model. A regression model score may be used to rank each of the following classifiers. The model score obtained through linear regression is somewhat better than that obtained from KNN regression.

Assessing the Effect of Advertising Expenditures on Sales: A Bayesian Structural Time Series Model by Gallego et al. (2019) is another relevant paper. Therefore, The Nerlove-Arrow model has been used in studies. The goal is to build a robust model connecting weekly sales and advertising expenditures using the framework of Bayesian structural time-series models. The model's modular construction and adaptability make it suitable for a wide range of applications. As a consequence, we can say that the model can be built in modules to accommodate changing business needs, has low anticipated error rates, and is easily interpretable. Spike-and-slab prior knowledge is used in the model's variable selection process. It accounts for outliers and recommends ads that are more likely to bring in customers. The model may be used as a tool to help decide how money is spent on advertisements by upper management.

Table 2.4 Similar Works on Research Study

No.	Title	Technique/ Algorithm	Objective	Result	Reference
1.	Social Media Advertisement and its Effect in Sales Prediction - An Analysis	Linear Regression	The purpose of this research is to foresee the effects of social media advertising on business revenue. The goal of this study is to develop a linear regression model for estimating sales using YouTube marketing budgets.	The result of this research shows that YouTube advertising is a better predictor of company sales	(Subha, 2020)
2.	The Role of Twitter Medium in Business with Regression Analysis and Statistical Modelling	Linear Regression	Aims to analyze Twitter's function in corporate communication and if it can forecast company performance. The research uses regression analysis and statistical modeling to examine the link between Twitter activity and company success variables including sales, revenue, and customer engagement. This study seeks to determine whether Twitter is a useful tool for companies to reach their consumers and objectives.	The results of this research show that advertising on Twitter may boost business. Additionally, the exponential T-X exponentiated exponential is presented as a novel statistical model.	(Cong et al., 2021)

Table 2.4 Similar Works on Research Study (Continued)

3.	Impact of Facebook and Newspaper Advertising on Sales: A Comparative Study of Online and Print Media	Linear Regression	For analysis purposes, we consider a statistical tool called simple linear regression modeling. To check the significance of the advertising on sale, definite statistical tests are applied.	According to the results of this study, commercials have a substantial impact on revenue. It has been shown that advertising expenditures result in a fourfold increase in ROI compared to no spending at all. Internet advertising outperforms print advertising. Internet advertising triples revenues compared to print.	(Lin et al., 2021)
4.	Sales Prediction Using Linear and KNN Regression	Linear Regression and KNN	To compare the performance and effectiveness of linear regression and K-nearest neighbors (KNN) regression models in predicting sales.	KNN regression is an obvious example of a model that overfits data. Each of the preceding classifiers may be evaluated according to their regression model score. Compared to KNN regression, linear regression produces a slightly higher model score.	(Kohli et al., 2021).

Table 2.4 Similar Works on Research Study (Continued)

5.	Assessing the effect of advertising expenditures upon sales: a Bayesian structural time series model	The Nerlove-Arrow model	Using the structure of Bayesian structural time-series models, a strong model that links advertising costs and weekly sales can be made. Because the model is flexible and made up of separate parts, it can be used in many different places or situations.	The model has minimal predicted errors, is interpretable, and can be constructed modularly to adapt to new business circumstances. The model selects variables and incorporates spike-and-slab prior knowledge. It handles non-gaussian variations and suggests sales-boosting advertising strategies. The management may utilise the model as a decision support system to allocate ad investments.	(Gallego et al., 2019)

2.6 IMPLICATION OF LITERATURE REVIEW

This literature review is conducted to make research on sales prediction for media platform advertising expenditure using a linear regression algorithm. This literature review has studied sales in media platform advertising expenditure which is sales in media platform advertising expenditure refers to the income generated by businesses from advertising on various digital platforms. It represents the financial returns or sales outcomes generated by the advertising expenditures on these platforms. Predicting sales based in the advertising expenditure provides businesses with valuable insights into future performance, allowing them to allocate resources effectively, establish realistic goals, evaluate performance, and make data-driven decisions.

A system using an efficient algorithm is required to do early sales prediction in a quicker and more accurate way. This chapter has shown that one of the most efficient approaches to get an accurate result is using a linear regression algorithm for predicting the sales in media platform advertising expenditure. Therefore, it has the potential to increase sales since the company can plan its advertising budget more precisely.

The linear regression approach that will be implemented to develop this system is widely used in the field of sales prediction due to its high accuracy and ease of use. Thus, implementing the linear regression algorithm to develop this proposed research will be able to provide a reliable system for professionals to predict sales in advertising expenditure accurately and faster, which also can help to save more time and cost. It also can improve the efficiency of the existing method of sales prediction by creating this proposed system.

2.7 CONCLUSION

In conclusion, in order to gather relevant information for the proposed research, relevant articles, journals, websites, and conference papers were reviewed. This chapter has discussed sales prediction in media platform advertising expenditure systems. The first part has covered the description of the predictive modeling area as well as the explanation of applications, phases, and techniques of the predictive modeling system. Then, the second part elaborated on sales prediction in media platform advertising expenditure, followed by the implementation of prediction systems in various areas. The third part of this literature review has described the linear regression algorithm-based o how it works, the step to implement the algorithm, and the advantages of the linear regression algorithm. Then discusses the similar works of the proposed project and the implication of the literature review in the last part. The research methodology for this proposed project will be discussed in the next chapter.

CHAPTER 3: METHODOLOGY

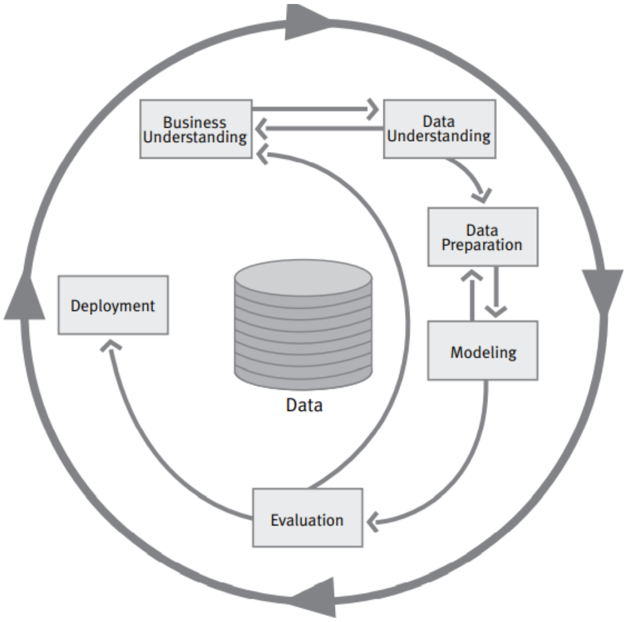
According to Pamplana (2022), "research methodology is a systematic framework for addressing study problems by using the best and most practical ways to conduct the research in alignment with the purpose and goals of the research." Methods of data collection and analysis used in this study are presented and explained in detail in this section. Research design is the process through which a study is planned in enough detail and precision to provide results that can be trusted and which also achieve the study's objectives. The researcher plans to use the Crisp-DM methodology to ensure the study is conducted well. To bring different points of view together and provide a complete picture, this research will use a mixed-method approach that will blend qualitative and quantitative techniques.

3.1 CRISP-DM

The complexity and scope of the study of extracting value from data has increased under the umbrella of data science. Today, data science is favored and utilized more frequently than data mining (Martínez, 2019). According to Martínez (2019), if data mining is analogous to gold mining from a metaphorical discourse perspective, then data science is analogous to prospecting or seeking for profitable mining sites. Data mining is used for consistency purposes in this paper. Moreover, the term "data mining" is utilized in the Cross-Industry-Standard-Process paradigm for Data Mining (CRISP-DM) (Wirth et al., 2000), which has been thoroughly implemented over the past two decades (Martínz, 2019). Data mining is goal-driven and more concerned with processes, whereas data science is data-driven and more concerned with the exploration of values, including goal-driven values (Martínez, 2019). Thus, the CRISP-DM process paradigm is derived from goal-oriented perspectives, yet it can still be applied to data science efforts.

The CRISP-DM is a generic data mining process model that offers an overview of data mining project life cycles (Wirth et al., 2000). According to user questionnaires and numerous surveys, CRISP-DM is regarded as the de facto standard for knowledge discovery and data mining projects (Martínez, 2019). CRISP-DM is popular in both industry and academia (Mariscal, 2010), and it is regarded as the de facto standard for knowledge discovery and data mining projects (Mariscal, 2010). Using the CRISP-DM reduces project costs and timelines, reducing the CRISP-DM reduces project costs and timelines, and reduces the knowledge requirements for data mining. In addition, using CRISP-DM expedites training, knowledge transmission, documentation, and the capture of best practices (Chapman et al. 2000). More importantly, especially for researchers and practitioners, data mining can be applied to innovation (Kong, 2017) endeavors. CRISP-DM is therefore beneficial for innovative endeavors.

CRISP-DM specifies the steps that must be completed in a DM project. CRISP-DM also specifies the tasks and deliverables for each phase. CRISP-DM is broken down into six stages. Based on Figure 3.1, the stages discussed are business understanding, data understanding, data preparation, modeling, evaluation, and deployment



(Source Chapman et al., 2000)

Figure 3.1 Phaes of the CRISP-DM Methodology

Table 3.1 CRISP-DM Project Framework

Research Methodology	Objective	Task	Activities	Deliverables
Preliminary Phase: Business understanding	**<u>Objective 1:</u>** To study linear regression algorithm in sales prediction for media platforms' advertising expenditure.	Literature review	1. Reading articles, journals, books, and conference papers. 2. Watching YouTube videos that related to the title project and prosed algorithm 3. Identify the problem statement 4. Identify the project's title 5. Decide the suitable algorithm for the project	1. Understanding sales in the media platform's advertising expenditure 2. Find potential techniques/ algorithms for Sales Prediction for Media Platform's Advertising Expenditure. 1. Algorithm to be used is linear regression 2. Research background on Sales Prediction for Media Platform's Advertising Expenditure using Linear Regression. 3. Problem statement 4. Objectives 5. Significance
Data understanding		Data collection	Search and download the data about sales in the media platform's advertising expenditure from Kaggle	1. Sales data to be processed 2. Clean data
Data preparation		Data Preprocessing	1. Data cleaning 2. Data transformation 3. Data splitting	
Design And Implementation Phase: Modeling	**<u>Objective 2:</u>** To develop a prototype of sales prediction for media platforms' advertising expenditures using the linear regression algorithm.	Research	1. Design Flowchart, system architecture, and Pseudocode. 2. Implement the linear regression algorithm prototype. 3. Build a prototype of Sales Prediction for Media Platform's Advertising Expenditure.	Prototype of sale prediction for media platforms' expenditure using linear regression algorithm.
		Development		
Evaluation Phase: Evaluation	**<u>Objective 3:</u>** To evaluate accuracy of linear regression algorithm in predicting the sales for the media platform's advertising expenditure.	Algorithm evaluation	Evaluate the accuracy of the linear regression algorithm in forecasting the expected sales in the media platform's advertising expenditure.	Accuracy of Linear Regression Algorithm in Sales Prediction for Media Platform's Advertising Expenditure.
Deployment		Report writing	Write the full report	1. Final Report 2. Final Presentation

3.2 BUSINESS UNDERSTANDING

When initiating a new data mining project, the researcher must understand the retail and commercial aspects of the endeavor. In this early phase, the primary focus should be placed on identifying the requirements and objectives of the project from a business perspective. With this understanding in hand, the next step is to define the data mining issue and develop a general strategy for achieving the objectives. Thus, both the subsequent actions and the outcome must always be in agreement with the first step (Gupta, 2014). In most situations, management is required to choose the group that will be in charge of the data mining project as well as establish an initial expenditure plan for the endeavor. In contrast, there will not be any financial outlays necessary to collect the necessary data for the investigation as part of this investigation. The researcher will make the decision as to which components of the organization should be investigated, develop a business model if one is necessary, outline the study's aims and expected outcomes, specify the project's objectives, and discover data-collection tools.

3.2.1 LITERATURE REVIEW

An in-depth analysis of the existing literature provides context for the proposed endeavor and reveals overlooked concepts, methods, and gaps in the research. Literature review entails searching for and evaluating appropriate texts like books and academic journals, then synthesizing, and analyzing the findings. First, some background reading and research into the desired subject and algorithm to be applied have been done to accomplish the study of the linear regression approach in predicting the predicted sales in advertising spend on media platforms.

Numerous articles, books, journals, and conference papers relevant to the proposed title were used to conduct the research and study. These citations are sourced from several online resources, including ResearchGate, Science Direct, Google Scholar, and the IEEE. The UiTM Library website has been very helpful since it has a collection of research papers pertaining to the subject from all around the globe. In addition, because it is crucial to acquire a comprehensive comprehension of the algorithm to be applied, more research has been carried out by perusing several movies available on YouTube that offered the relevant details. Because of this exploratory research, we now have a better understanding of the context, the issue, our goals, the suggested algorithm, our assessment method, and the benefits of the new system, all of which will inform the development of the system proposal.

3.3 DATA UNDERSTANDING

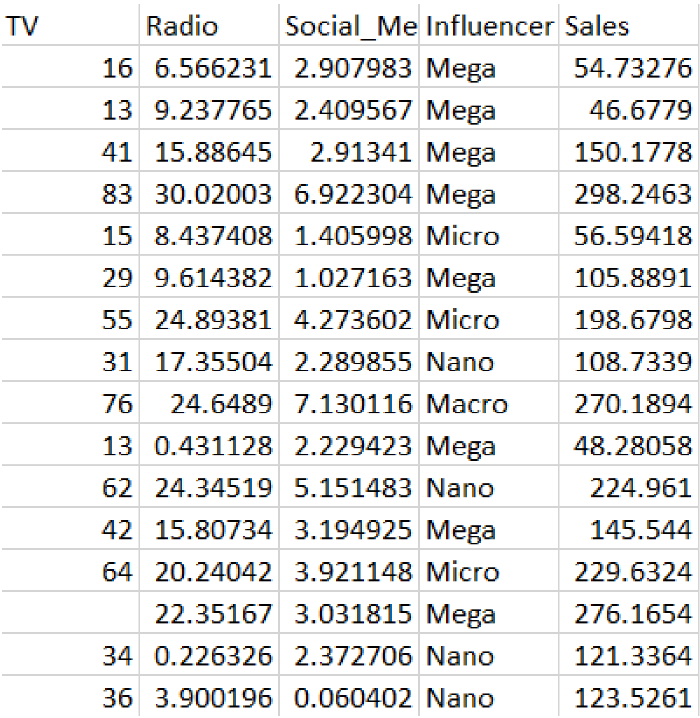
In the second phase of the model, the necessary data is collected through web scraping and by obtaining it from trustworthy sources such as GitHub and Kaggle. The appropriate data are chosen, which will serve as the basis for determining how accurate the hypotheses developed during the preliminary phase were. This involves keeping a record of the sources and challenges encountered when gathering the data to make it accessible for evaluation in the event that the process is repeated. This phase also includes reflecting on the business objectives to have a better understanding of the data and checking the data quality of the data that was acquired from a variety of sources. As a result, you often could switch back and forth between these two early stages in order to better grasp the data. (Gupta, 2014).

3.3.1 DATA COLLECTION

Data collection refers to the process of obtaining data for the purpose of using it in the process of developing the project prototype. All the data that obtain comes from Kaggle, and it consists of a total of 4573 data. Television, radio, social media, and influencer the attributes that are considered inputs in this dataset, whereas sales are the asset that is considered an output.

Table 3.2 Description of Attributes

Attribute	Description	Type
Television	Amount of advertising expenditure on television	Numeric
Radio	Amount of advertising expenditure on radio	Numeric
Social Media	Amount of advertising expenditure on newspaper	Numeric
Influencer	Collaboration with influencer (Mega, Macro, Nano, Micro)	Categorical
Sales	Total of sales	Numeric



TV	Radio	Social_Me	Influencer	Sales
16	6.566231	2.907983	Mega	54.73276
13	9.237765	2.409567	Mega	46.6779
41	15.88645	2.91341	Mega	150.1778
83	30.02003	6.922304	Mega	298.2463
15	8.437408	1.405998	Micro	56.59418
29	9.614382	1.027163	Mega	105.8891
55	24.89381	4.273602	Micro	198.6798
31	17.35504	2.289855	Nano	108.7339
76	24.6489	7.130116	Macro	270.1894
13	0.431128	2.229423	Mega	48.28058
62	24.34519	5.151483	Nano	224.961
42	15.80734	3.194925	Mega	145.544
64	20.24042	3.921148	Micro	229.6324
	22.35167	3.031815	Mega	276.1654
34	0.226326	2.372706	Nano	121.3364
36	3.900196	0.060402	Nano	123.5261

Figure 3.2 Sales in Media Platform Advertising Expenditure

3.4 DATA PREPARATION

The CRISP-DM model's third phase is data preparation, which can be divided into selecting, pre-processing, and transformation. The initial stage is to choose the data to be analyzed by checking its consistency, consistency, and plausibility. Next, the data is assessed in the pre-processing phase, and based on its significance, it is either kept or discarded from the main data set. A similar term is "data cleansing." The next step is transformation when the information is standardized and encoded. Experts estimate that data preparation takes over 80% of the total time needed for a data mining project (Sharda et al., 2015), making it the most laborious and time-consuming phase of the CRISP-DM framework.

3.4.1 DATA PRE-PROCESSING

The preparation of data is a crucial phase that must be performed before any machine learning models are trained. The objective of the data preprocessing phase is to optimize the training and testing procedure through transforming unprocessed information into formats that are accessible and practical.

3.4.1.1 DATA CLEANING

The data cleaning process in this study specifically focuses on handling missing values, aiming to ensure data accuracy and mitigate potential issues during the subsequent data mining stages. To address the absence of values, the approach involves employing imputation techniques, such as filling in missing values with the attribute mean, mode, or median. This method is particularly crucial for maintaining the integrity of numerical data, which is foundational to the proposed project. By imputing missing values with statistical measures like the mean or median, the dataset is prepared for analysis, reducing the risk of inaccuracies and incorrect results that could arise from incomplete information. This meticulous handling of missing values underscores the commitment to robust data quality and enhances the reliability of the dataset for the research objectives.

3.4.1.2 DATA TRANSFORMATION

Within the data preparation phase, a pivotal step is the "data transformation" process, where the raw data undergoes modifications that enable thorough analysis. While including various techniques like normalization, standardization, and feature engineering, a significant emphasis is placed on encoding categorical variables, especially the "Influencer" column. When considering the categorical characteristics of the attribute "Influencer" (Mega, Macro, Nano, Micro), it is critical to highlight the use of one-hot encoding. This technique makes sure that the numerical format of the categorical data is suitable, which improves the dataset's machine learning model compatibility. Through the application of one-hot encoding to the "Influencer" column in particular, the data transformation procedure enhances the accuracy and consistency of numerical values while also improving machine learning algorithm performance, creating an environment that supports insightful and accurate analysis.

3.4.1.3 DATA SPLITTING

Data splitting is a technique that divides a dataset into two sections, the training set and the testing set. The algorithm model will construct a trained prediction model by learning from the data using a training set. The training/testing ratios considerably influenced the prediction capabilities of ML models, with the 70/30 ratio performing best. The findings reported here indicated an efficient way to pick the right dataset ratios and the best ML model to reliably estimate soil shear strength, which would aid building project planning and engineering (Nguyen et al., 2021). In this proposed project, the dataset (4573 samples) will be divided 7:3 ratio, meaning that 70% (3201 samples) of the data will be used for training and 30% (1372 samples) will be used for testing. While the common 7:3 training-to-testing data ratio is prevalent, alternative ratios like 9:1 or 8:2 can also be suitable depending on research project requirements. Opting for a 7:3 ratio offers benefits such as providing ample training data (70%) for the model to capture intricate predictor-outcome relationships, reserving a reasonable testing data (30%) for robust performance evaluation, and striking a balance between bias and variance. Higher ratios might lead to a smaller testing set, affecting performance assessment, particularly when data is limited. This balanced approach ensures effective model learning without overfitting or underfitting.

Table 3.3 Data Splitting of Salas Advertising Dataset

Dataset	Ratio	Number of Samples
Training Set	70	3201
Testing Set	30	1372

3.5 MODELING

Within this phase, multiple modeling strategies are selected and applied to a previously created data collection in order to meet a specific business requirement. Evaluation and comparison of the numerous models developed is also a part of the model construction procedure. Due to the lack of a universally accepted optimal technique or algorithm for data mining, the "best" method for a given objective must be determined using 49 distinct model types and a well-defined testing and evaluation plan. Following this stage, data visualization becomes the primary objective of the modeling phase (Gupta, 2014).

In addition, at this stage, the data is transformed into charts that may be used to evaluate different data movement and storage options. In this stage, the user interface design is also finalized. All the parts of the system are settled on during design.

3.5.1 SYSTEM ARCHITECTURE

The system architecture for sales prediction using linear regression involves a comprehensive pipeline, beginning with the data collection phase where information on advertising expenditures across various platforms and corresponding sales is gathered. This data is thoroughly preprocessed, which includes cleaning and a key 80/20 split of training and testing sets. By determining the most relevant features for sales prediction, feature selection is taken into consideration, providing flexibility in improving the model's performance. The use of a linear regression model provides the basis of the architecture. To discover the correlations between advertising expenses and sales, the training dataset is loaded, the model is instantiated, and training begins. The influence of each prediction on the target variable is then shown by extracting the model's parameters, which include coefficients and an intercept.

The system design additionally involves a phase for evaluating the model, in which the testing set is used to thoroughly evaluate the trained linear regression model to verify its prediction accuracy and ability to generalize to new data. A user interface component might be added to the architecture to enable users to communicate with the model and get predictions and insights in real time. In the context of media platform advertising, this user interface improves stakeholders' ability to enter new information on advertising expenditures and obtain predictions immediately. This helps with strategic planning and well-informed decision-making.

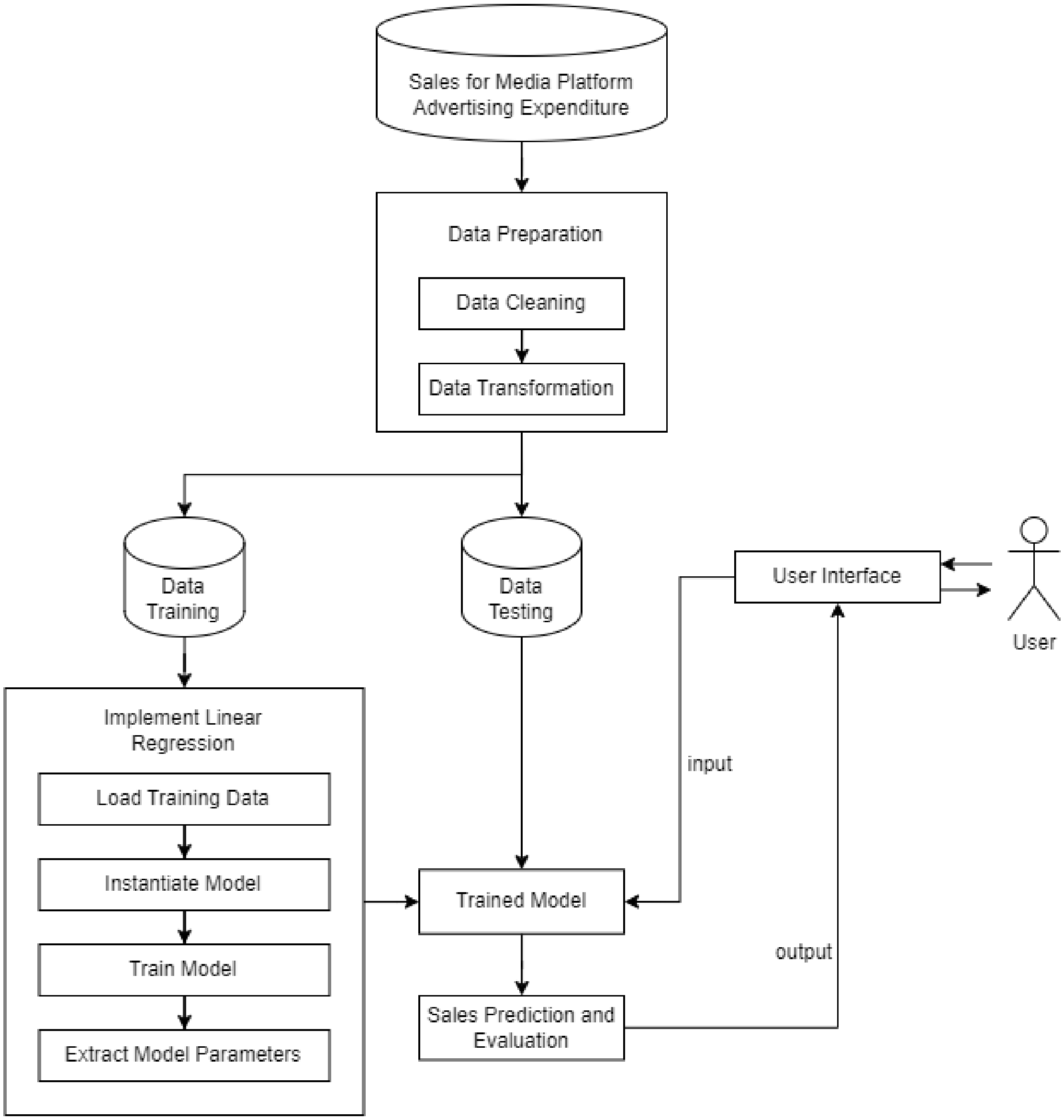


Figure 3.3 System Architecture of the Sales Prediction for Media Platforms' Advertising Expenditure using Linear Regression

3.5.2 FLOWCHART

A flowchart is a method of presenting information in a structured manner, enabling simpler understanding for readers and minimizing subjectivity. A flowchart is a graphical representation that depicts the functions and processes of a complex system. Qualitative research flowcharts function as navigational implements, guiding researchers from the first stages of their investigation to the ultimate conclusion of their results. a system flowchart also depict the relationship between input, output, and process (Lithmee, 2018). Figure 3.4 below displays the flowchart of the proposed prediction model.

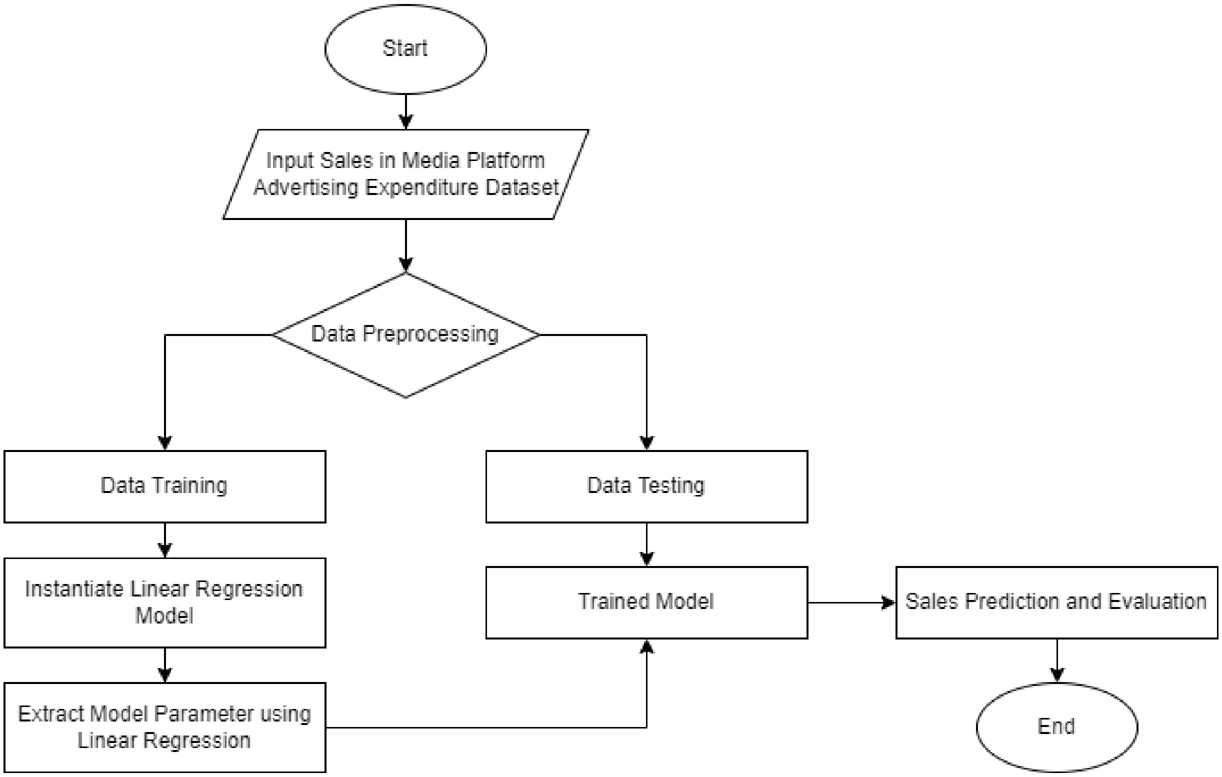

Start
Input Sales in Media Platform Advertising Expenditure Dataset
Data Preprocessing
Data Training
Data Testing
Instantiate Linear Regression Model
Trained Model
Sales Prediction and Evaluation
Extract Model Parameter using Linear Regression
End

Figure 3.4 Flowchart of the Sales Prediction for Media Platforms' Advertising Expenditure using Linear Regression

3.5.3 USER INTERFACE DESIGN

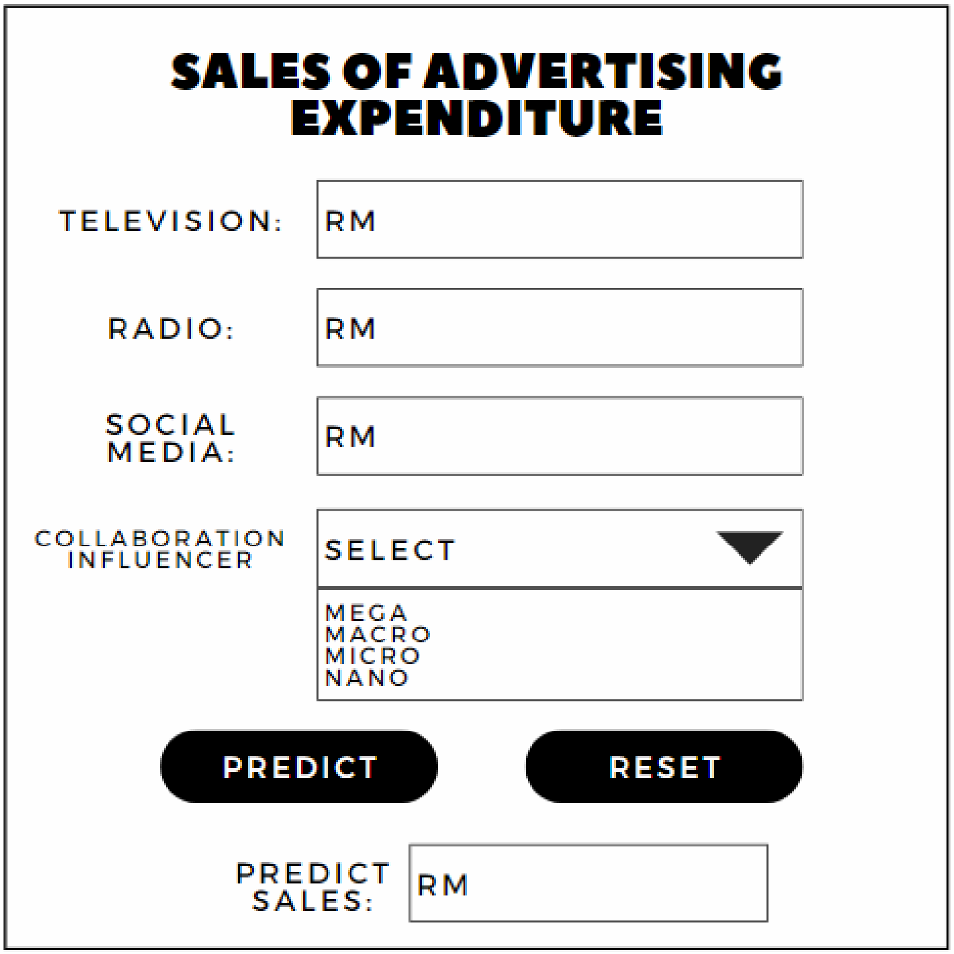
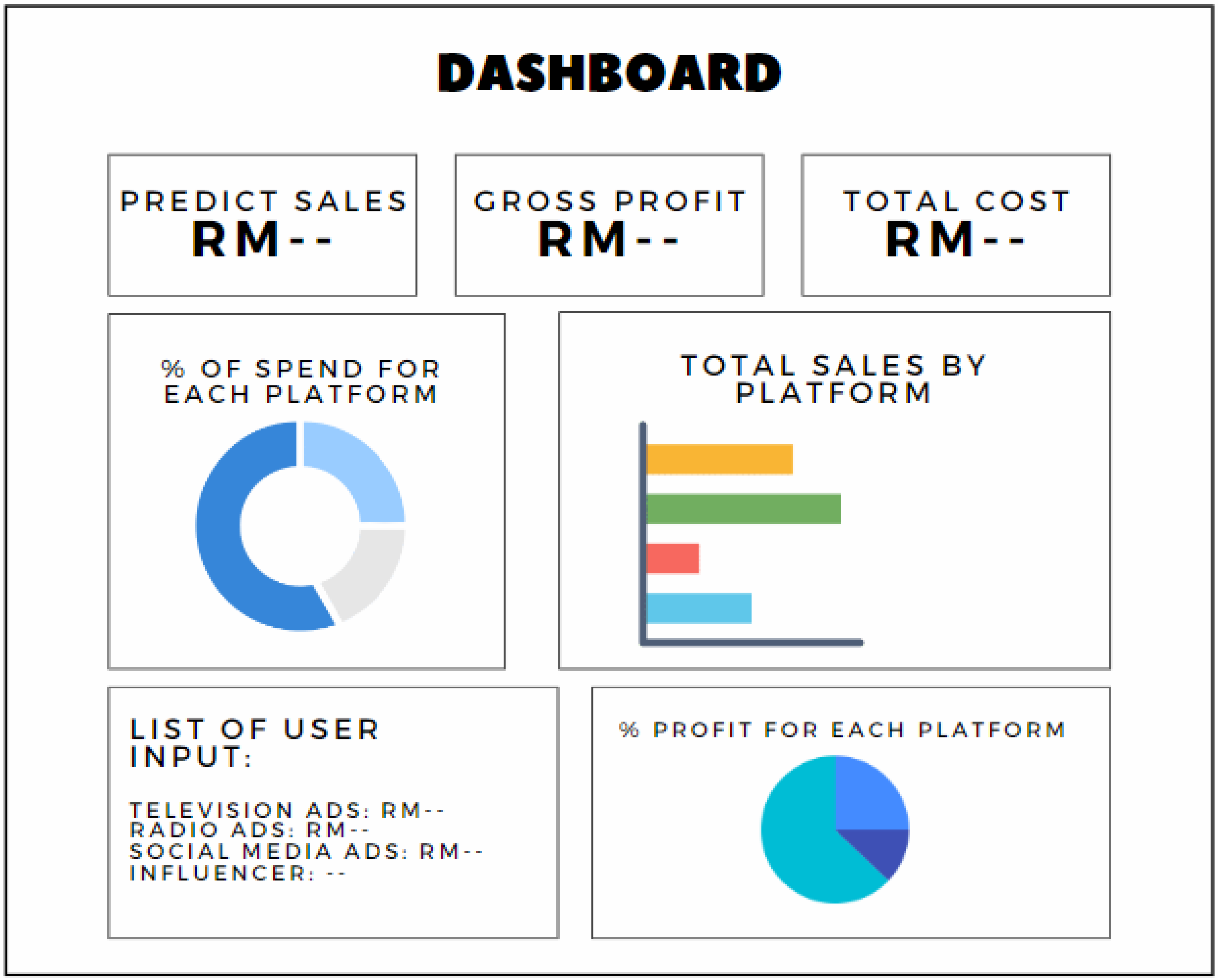


Figure 3.5 User Interface of Sales Prediction for Media Platforms' Advertising Expenditure using Linear Regression (Form Prediction)



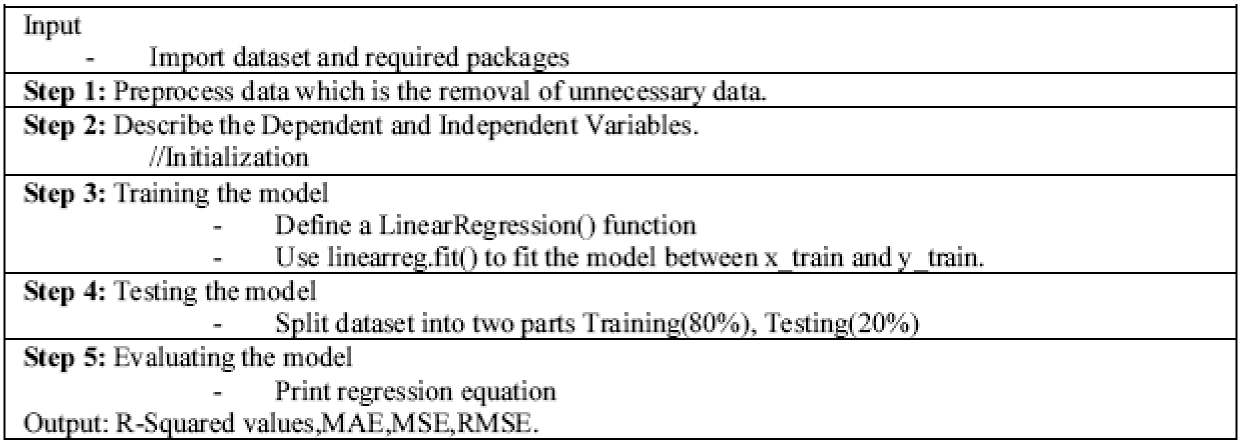
LIST OF USER INPUT:

TELEVISION ADS: RM--
RADIO ADS: RM--
SOCIAL MEDIA ADS: RM--
INFLUENCER: --

Figure 3.6 User Interface of Sales Prediction for Media Platforms' Advertising Expenditure using Linear Regression (Output Dashboard Result)

3.5.4 PSEUDOCODE OF SELECTED ALGORITHM

Pseudocode is a method of presenting computer algorithms that integrates plain language with programming, simplifying the algorithm development process. It is simply a step in the direction of the actual code development. Pseudocode demonstrates how an algorithm should be organized and sequenced without following to the exact coding syntax.



```
Input
    -   Import dataset and required packages
Step 1: Preprocess data which is the removal of unnecessary data.
Step 2: Describe the Dependent and Independent Variables.
        //Initialization
Step 3: Training the model
            -   Define a LinearRegression() function
            -   Use linearreg.fit() to fit the model between x_train and y_train.
Step 4: Testing the model
            -   Split dataset into two parts Training(80%), Testing(20%)
Step 5: Evaluating the model
            -   Print regression equation
Output: R-Squared values,MAE,MSE,RMSE.
```

(Source Malathib, P. et al., 2022)

Figure 3.7 Pseudocode of Linear Regression Algorithm

3.5.5 PROTOTYPE IMPLEMENTATION

The implementation phase of prototype development contains testing and debugging, both of which are crucial tasks for arising the conceptualized system. Python is a programming language that is both adaptable and extensively used. It is utilized for coding because of its readability, simplicity of use, and large library, which allow for flexibility in creating different functionality. By providing features including code completion, syntax highlighting, and an integrated debugger, VS Code assists in the development of an efficient coding experience. These instruments improve the efficacy and productivity of the development procedure by providing the prompt identification and solution of problems. The strategic choice of hardware and software components is crucial in the implementation of the prototype, as it performs to establish a reliable framework for the undertaking and guarantee seamless progress. The specifics of each piece of hardware and software picked add to the general strength and dependability of the system being suggested.

Table 3.4 Hardware Requirements

Item	**Description**
Processor	Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz
Memory	4 GB
System Type	64-bit operating system, x64-based processor

Table 3.5 Software Requirements

Item	**Description**
Operating System	Windows 10 Home Single Language
Programming Tools	VS Code (python)

3.6 EVALUATION

The CRISP-DM methodology then proceeds to the evaluation phase, which involves looking at the outcomes, taking a review of the procedure, and planning for the future. After modeling, the outcomes are analyzed to see how well the model or representation serves the company's goals. Evaluation and assessment of outcomes, comparison of results, and interpretation of patterns with respect to validity, understandability, and interest are all part of this stage. Data mining initiatives benefit from having outcomes prioritized in terms of business success criteria, with consequences checked and lessons learned stated. To put it another way, (Chapman, 2000).

During the modelling phase, assessment stages addressed issues such as model correctness and generality. This activity evaluates the model's ability to achieve business goals and attempts to establish whether there is a business reason why this model is poor. A further in-depth examination of the information mining engagement is required to see if anything critical was overlooked. Finally, depending on the assessment results and the process evaluation, the researcher selects the next actions. Decisions must be made on whether to finish this project and proceed with the deployment phase, whether to do more iterations, or whether to begin new data mining projects.

3.6.1 LINEAR REGRESSION

During the evaluation phase of our research relating to the implementation of linear regression for predicting sales in media platform advertising expenditure, we implement a wide variety of performance metrics to evaluate the efficacy of our predictive model. The accuracy as well as reliability of our linear regression model can be thoroughly evaluated using a variety of metrics, which consists of r2 (coefficient of determination), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Table 3.6 Evaluation Metrics R Square and MSE

Evaluation Metrics	Formula
R Square/Adjusted R Square:	$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$
Mean Square Error (MSE):	$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

The r2 value is an essential metric for knowing what percentage of the variance in the sales data can be captured by our advertising expenditure model. A better fit is indicated by a higher value of r2, which represents the model's capability to capture the basic connections in the data. On the other hand, MSE calculates the average of the squared deviations in sales values between what was predicted and what occurred. A lowered mean squared error (MSE) signifies improved predictive precision, highlighting the model's ability to minimize prediction errors.

Table 3.7 Evaluation Metrics RMSE and MAE

Evaluation Metrics	Formula
Root Mean Squared Error (RMSE):	$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$
Mean Absolute Error (MAE):	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$

The MAE metric offers valuable information regarding the mean magnitude of differences that happen between predicted and actual sales figures. Because it is so simple, it is easy to understand how well the model is working and what the usual estimate error is. Additionally, RMSE, which is the square root of MSE, provides an improved assessment of prediction accuracy by giving greater weight to larger errors.

Our objective is to thoroughly evaluate the efficacy of our linear regression model in predicting sales by applying the advertising expenditures of media platforms as the basis for these metrics. By implementing the combined metrics of RMSE, r2, MSE, and MAE, we can acquire an in-depth understanding of the model's qualities and flaws. This understanding performs as a robust basis for improving and maximizing our prediction capacities with relation to advertising expenditure on media platforms.

3.7 DEPLOYMENT

The final phase is deployment, in which the data is converted into concrete steps and utilized by decision-makers such as the team leader or the business CEO. The analysis should be understandable by the relevant departments so that they may make educated decisions. If the organization needs comparable analyses in the future, the CRISP-DM procedure must be automated. (Gupta, 2014).

The process of deploying the created machine learning model or data mining solution into a real-world operational environment is referred to as deployment in the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach. Documentation is critical to ensuring a successful deployment. At the end of the activity, the final report will be created, followed by the final presentation, which will conclude the whole research study. The presentation often contains a major portion of the material from the final report.

3.7.1 GANTT CHART

Table 3.8 Gannt Chart of the Project

ACTIVITIES	MAR 2023		APR 2023			MAY 2023				JUNE 2023			JULY 2023		OKT 2023			NOV 2023			DIS 2023				JAN 2024			FEB 2024
	Duration (Weeks)																											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	1	2	3	4	5	6	7	8	9	10	11	12	13	14
PHASE 1: BUSINESS UNDERSTANDING	■	■	■	■	■	■																						
Literature review	▨	▨	▨																									
Identify the problem statement				▨	▨	▨																						
PHASE 2: Data UNDERSTANDING							■	■	■	■	■	■	■	■	■	■	■											
Data collection							▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
PHASE 3: DATA PREPARATION											■	■	■	■	■	■	■	■	■	■	■							
Determine Table, Case and Attribute of data											▨	▨	▨	▨	▨	▨	▨											
Construct data into Final data												▨	▨	▨	▨	▨	▨	▨	▨	▨	▨							
Transformation and Cleaning of data												▨	▨	▨	▨	▨	▨	▨	▨	▨	▨							
PHASE 4: MODELING														■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Produce flowchart and system architecture														▨	▨	▨	▨	▨	▨	▨								
Design the user interface of the system															▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Implementation and development																	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
PHASE 5: EVALUATION																							■	■	■	■	■	■
Testing the system																							▨	▨	▨	▨	▨	▨
Evaluate Model																							▨	▨	▨	▨	▨	▨
PHASE 6: DEPLOYMENT	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Final report	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Final presentation	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

3.8 CONCLUSION

This chapter describes in detail how a researcher implements linear regression in regression analytics for data mining in order to obtain insight using the CRISP-DM methodology. Business understanding, data understanding, data preparation, modelling, evaluation, and deployment are the 6 phases of the CRISP-DM approach that must be followed in order to realize the study goal. Insight on how to carry out each step effectively is provided by the technique.

Expenditures on advertising across media platforms were predicted using the CRISP-DM approach. The study was built around a foundation of established business goals and data needs. Information on budgets for advertising and product sales was collected and cleaned and transformed. Multiple prediction techniques, such as those based on time series analysis and machine learning, were used to construct robust prediction models. To determine which model was more accurate, we used measures like mean absolute error and root mean squared error. The selected model was implemented to provide continuing sales forecasts, and its performance was checked and maintained on a regular basis. Businesses that used CRISP-DM saw an increase in income as the outcome of more knowledgeable advertising alternatives.

CHAPTER 4: RESULT AND FINDING

The findings and results obtained through the methodology described in the previous chapter are detailed in this chapter. The results and analyses of the sales predicting in advertising expenditures through the implementation of a developed linear regression algorithm will be discussed into in enormous detail. This chapter will encompass the following six subjects: conceptual framework, Linear Regression algorithm programmed codes, prototype interface, evaluation results, discussion, and conclusion.

4.1 CONCEPTUAL FRAMEWORK

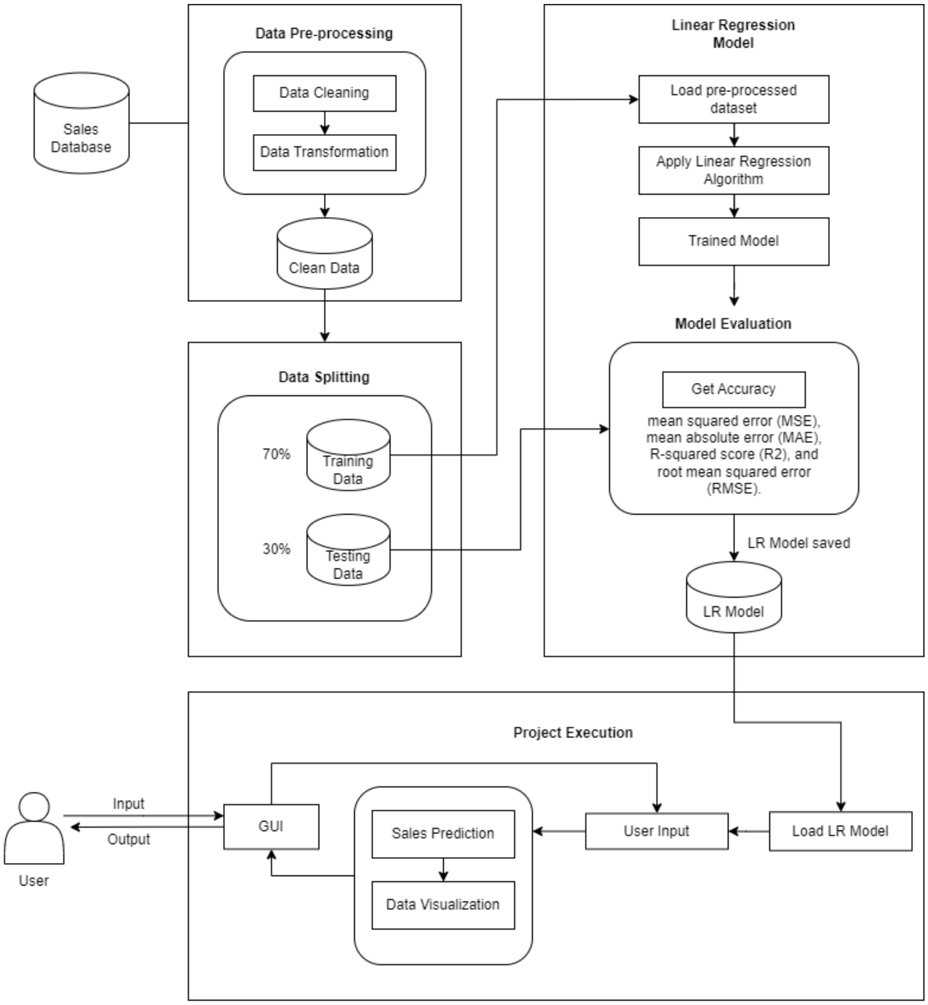
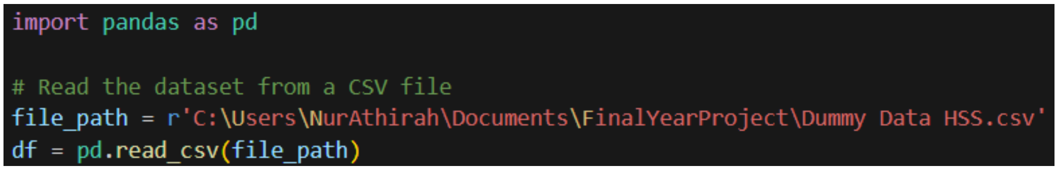


Figure 4.1 Conceptual Framework of Sales Prediction in Media Platforms Advertising Expenditure using Linear Regression

4.2 PROGRAM CODE FOR ALGORITHM

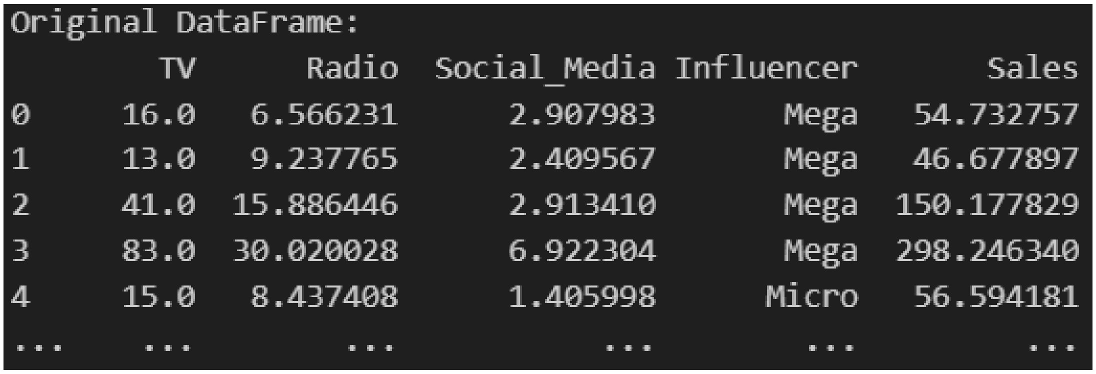
4.2.1 DATA COLLECTION

The dataset applied by this system is the Advertising Expenditure dataset from Sales in Media Platforms. To make it easier the import of the dataset into the system, a CSV file will be applied for saving the dataset. The dataset consists of a total of 4573 data. Input variables in this dataset consist of television, radio, social media, and influencers. Output variables are assets such as sales. Before loading the dataset into the system, the Python code supplied reads the dataset from a CSV file using the Pandas library. The file path to the CSV file is stored in the variable file_path, and the pd.read_csv(file_path) function is used to create a Pandas DataFrame (df) containing the tabular data from the CSV file. For a raw string literal, the r prefix is appended to the file path string. The resulting DataFrame (df) can be employed for various data manipulation and analysis tasks using Pandas functionalities.



```
import pandas as pd

# Read the dataset from a CSV file
file_path = r'C:\Users\NurAthirah\Documents\FinalYearProject\Dummy Data HSS.csv'
df = pd.read_csv(file_path)
```

Figure 4.2 Program Codes to Load the Dataset into System

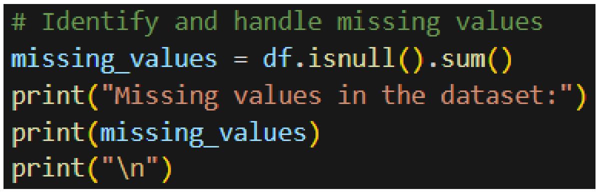


```
Original DataFrame:
        TV      Radio  Social_Media Influencer       Sales
0     16.0   6.566231      2.907983       Mega   54.732757
1     13.0   9.237765      2.409567       Mega   46.677897
2     41.0  15.886446      2.913410       Mega  150.177829
3     83.0  30.020028      6.922304       Mega  298.246340
4     15.0   8.437408      1.405998      Micro   56.594181
...    ...        ...           ...        ...         ...
```

Figure 4.3 Data Sample Dataset Loaded into System

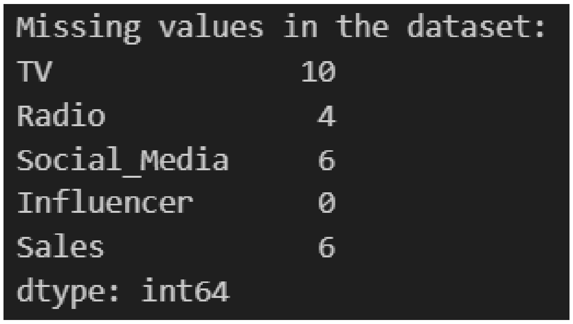
4.2.2 DATA PREPARATION

The Python function handles Pandas DataFrame missing values. Isnull() is used to find missing values and generate a missing_values Dataset that counts each column's missing values. Then, the code uses dropna() with inplace=True to delete rows with missing data from the DataFrame to apply the changes immediately. The total of the data samples in the dataset used are reduced from 4572 data to 4546 data.



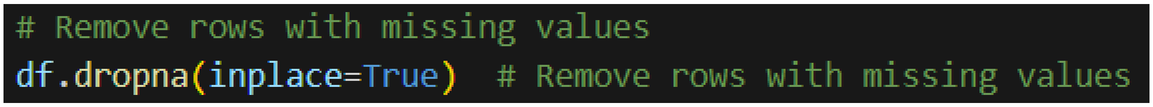
```
# Identify and handle missing values
missing_values = df.isnull().sum()
print("Missing values in the dataset:")
print(missing_values)
print("\n")
```

Figure 4.4 Program Code to Identify Missing Value



```
Missing values in the dataset:
TV              10
Radio            4
Social_Media     6
Influencer       0
Sales            6
dtype: int64
```

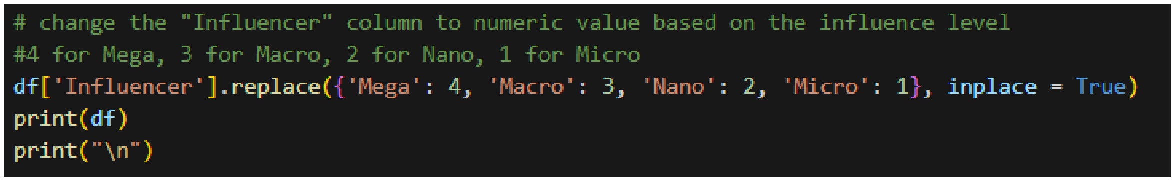
Figure 4.5 Output from Program Code to Identify Missing Value



```
# Remove rows with missing values
df.dropna(inplace=True)  # Remove rows with missing values
```

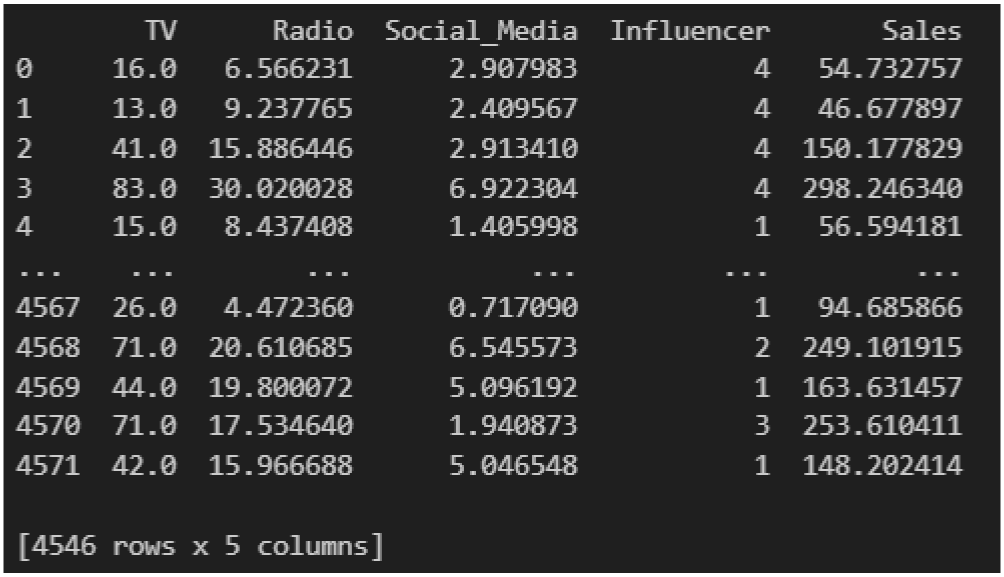
Figure 4.6 Program Code to Remove Rows with Missing Values

This code snippet is transforming the categorical values in the "Influencer" column of the DataFrame (df) into corresponding numeric values, representing different levels of influence. The levels 'Mega', 'Macro', 'Nano', and 'Micro' are replaced with numeric values 4, 3, 2, and 1, respectively. This numeric encoding likely reflects the hierarchical influence scale, where 'Mega' holds the highest influence level, followed by 'Macro', 'Nano', and 'Micro'. The inplace=True parameter ensures that the changes are applied directly to the original DataFrame. Finally, the modified DataFrame is printed to display the updated "Influencer" column with numeric values, providing a clearer numerical representation of influence levels for further analysis or modeling purposes.



```python
# change the "Influencer" column to numeric value based on the influence level
#4 for Mega, 3 for Macro, 2 for Nano, 1 for Micro
df['Influencer'].replace({'Mega': 4, 'Macro': 3, 'Nano': 2, 'Micro': 1}, inplace = True)
print(df)
print("\n")
```

Figure 4.7 Program Code of mapping Influencer column

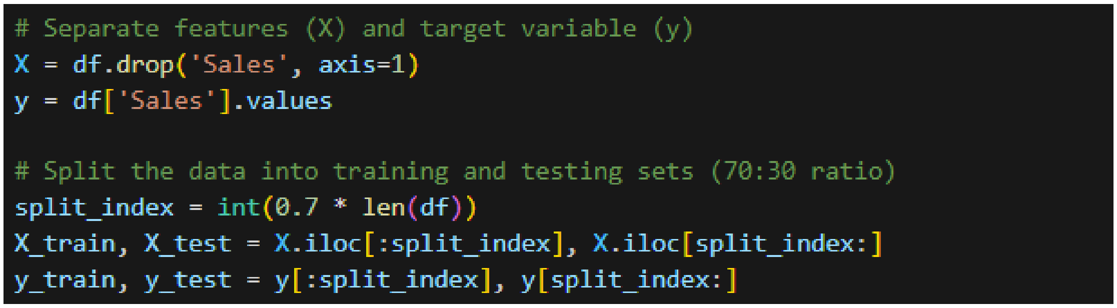


```
        TV      Radio  Social_Media  Influencer       Sales
0     16.0   6.566231      2.907983           4   54.732757
1     13.0   9.237765      2.409567           4   46.677897
2     41.0  15.886446      2.913410           4  150.177829
3     83.0  30.020028      6.922304           4  298.246340
4     15.0   8.437408      1.405998           1   56.594181
...    ...        ...           ...         ...         ...
4567  26.0   4.472360      0.717090           1   94.685866
4568  71.0  20.610685      6.545573           2  249.101915
4569  44.0  19.800072      5.096192           1  163.631457
4570  71.0  17.534640      1.940873           3  253.610411
4571  42.0  15.966688      5.046548           1  148.202414

[4546 rows x 5 columns]
```

Figure 4.8 Data Sample after Data Preparation

4.2.3 DATA SPLITTING

In this segment of the machine learning workflow, the code separates the dataset into features (X) and the target variable (y). In order to determine the features, represented as X, the 'Sales' column is removed from the original DataFrame(df). Meanwhile, the target variable, identified as y, is obtained through the process of separating the 'Sales' column. The dataset is then divided into training and testing sets utilizing the train_test_split function provided by the sklearn.model_selection module. The resulting sets, represented y_train, y_test, X_train, and y_test, consist of the target variable and features for both the training and testing phases in a ratio of 70:20. This process of partitioning ensures a dependable assessment of the performance of machine learning models on unseen data by enabling subsequent training and evaluation.

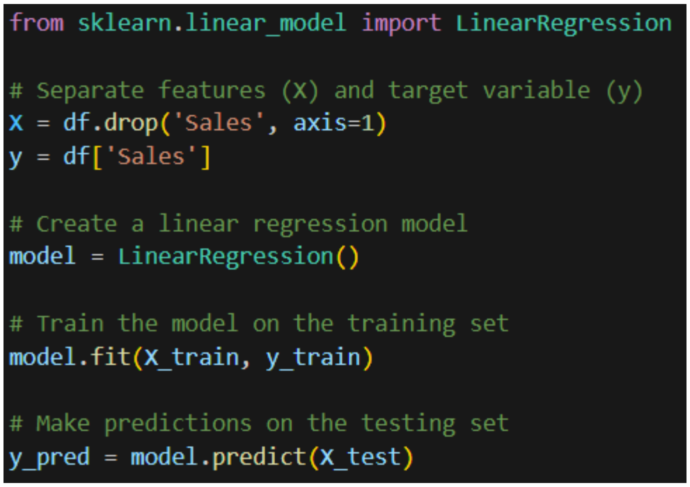


```
# Separate features (X) and target variable (y)
X = df.drop('Sales', axis=1)
y = df['Sales'].values

# Split the data into training and testing sets (70:30 ratio)
split_index = int(0.7 * len(df))
X_train, X_test = X.iloc[:split_index], X.iloc[split_index:]
y_train, y_test = y[:split_index], y[split_index:]
```

Figure 4.9 Program Code of Splitting the Data into Training and Testing Data

4.2.4 LINEAR REGRESSION ALGORITHM

The proposed approach involves the application of a linear regression (LR) algorithm to predict sales on a media platform, applying data related to advertising expenditure. The algorithm implemented in this system is constructed from scratch without relying on the scikit-learn library. The building of the LR algorithm relies on understanding the mathematical model, in which the dependent variable (sales) is expressed as a linear function of the independent variable (advertising spend). The manual implementation covers key steps such as initializing model parameters, employing an iterative optimization method like gradient descent, and updating the slope and intercept to minimize the difference between predicted and actual sales in the training data. This approach provides an in-depth understanding of the basic mechanisms of the LR algorithm. However, when using the linear regression approach with the scikit-learn module, the first process is to separate the features (independent variables) from the target variable (sales). Scikit-learn simplifies the implementation process by providing a very efficient Linear Regression model. The model is trained by using the fit method, which optimizes parameters such as slope and intercept to learn insight into patterns and correlations within the training data. The trained model can then use the predict technique for predicting new data, providing a simple and efficient way to use linear regression to predict sales in context of media platform advertising expenditures.



```python
from sklearn.linear_model import LinearRegression

# Separate features (X) and target variable (y)
X = df.drop('Sales', axis=1)
y = df['Sales']

# Create a linear regression model
model = LinearRegression()

# Train the model on the training set
model.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = model.predict(X_test)
```

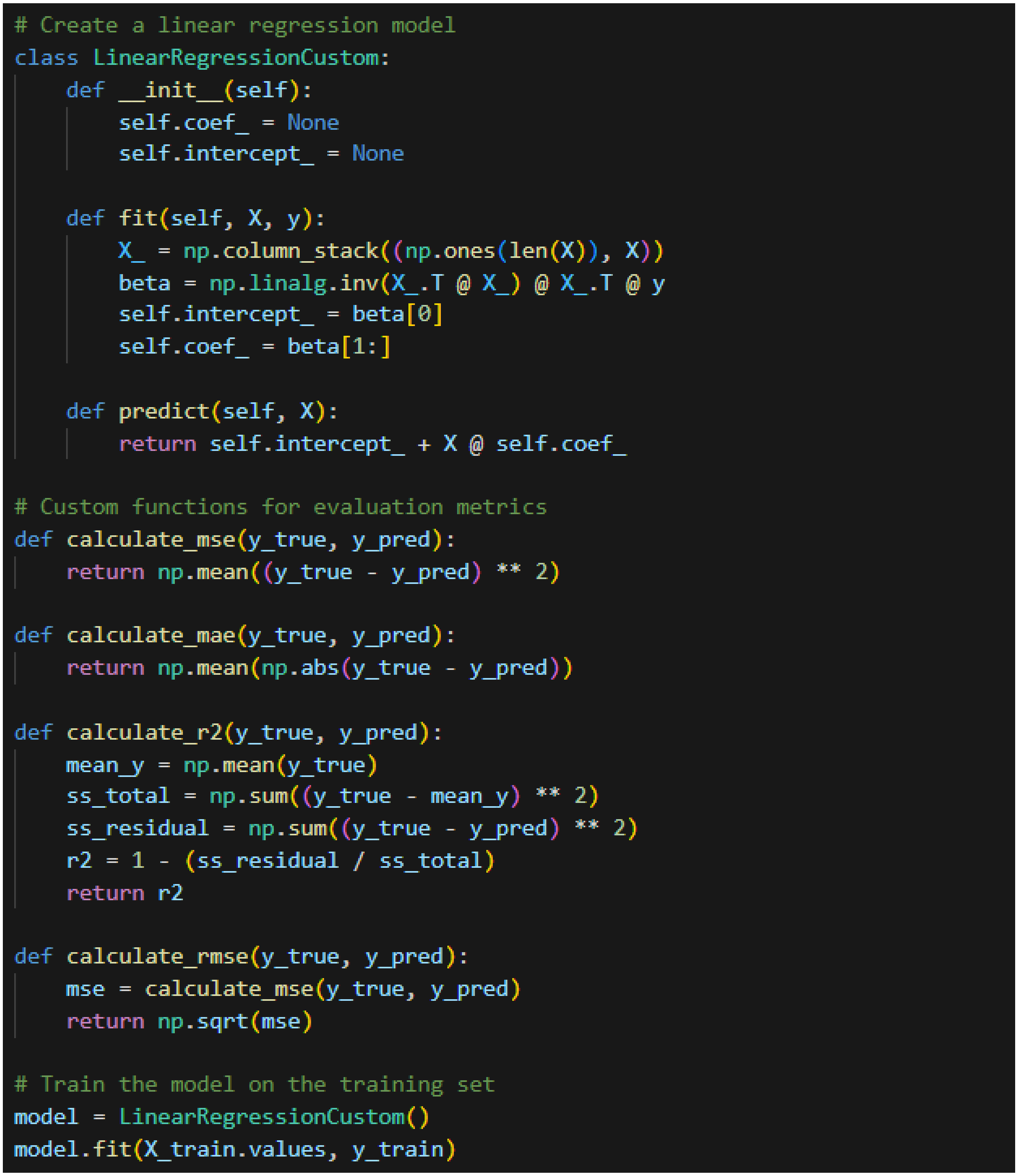
Figure 4.10 Program Code of the Linear Regression Algorithm using Scikit Learn Library

The provided code initiates the implementation of a custom linear regression (LR) algorithm from scratch, emphasizing a manual programming approach for a comprehensive understanding of the underlying principles. The dataset, loaded from a CSV file, is split into training and testing sets (70:30 ratio). The LR model is defined in a class with an initialization function (__init__) setting initial weights and intercept to None. The fit function employs the method of ordinary least squares for training the model, adjusting weights and intercept to minimize the mean squared error. The custom LR model is then evaluated on both the training and testing sets using calculated metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R2), and Root Mean Squared Error (RMSE).

Following the manual training and evaluation of the LR model, the code displays the performance metrics for both the training and testing sets. The calculated metrics provide insights into how well the LR model generalizes to new data (testing set) and fits the training data. This manual approach underscores a deep understanding of the LR algorithm, covering initialization, training, prediction, and evaluation processes without relying on external libraries. Additionally, it showcases the importance of evaluating the model's predictive performance using diverse metrics, offering a comprehensive assessment of its effectiveness.

Table 4.1 Explanation Each Function in LR Class

Function	Explanation
__init_()_	The LR model is defined in a class with an initialization function (__init__) setting initial weights and intercept to None. The fit function employs the method of ordinary least squares for training the model, adjusting weights and intercept to minimize the mean squared error.
fit()	The fit function trains Linear Regression utilizing features (X) and target variable (y). Initialize the weights and bias, then update them using gradient descent to minimize MSE. To monitor model performance, training progress is displayed every 100 epochs along with MSE, R2, MAE, and RMSE.
predict()	The predict function generates predictions by utilizing the input features (X) and the trained model. The final results are made by taking the dot product of the feature matrix and the weights and adding the bias term.
mean_squared_error()	Calculates the Mean Squared Error (MSE) between the actual and predicted values.
r2_score()	calculates the R-squared (R2) score, which measures how predictable the dependent variable's variance is from the independent variables. A higher R2 value suggests a better model-data fit.
mean_absolute_error()	Calculates the Mean Absolute Error (MAE) between the actual and predicted values.
root_mean_squared_error()	Calculates the Root Mean Squared Error (RMSE)

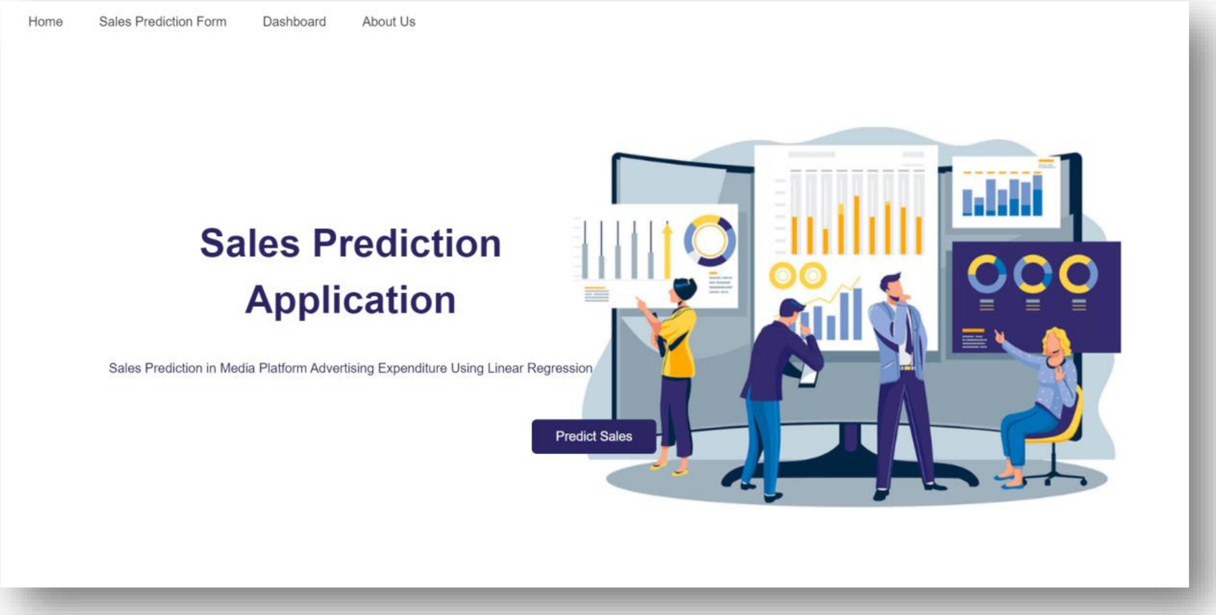


```python
# Create a linear regression model
class LinearRegressionCustom:
    def __init__(self):
        self.coef_ = None
        self.intercept_ = None

    def fit(self, X, y):
        X_ = np.column_stack((np.ones(len(X)), X))
        beta = np.linalg.inv(X_.T @ X_) @ X_.T @ y
        self.intercept_ = beta[0]
        self.coef_ = beta[1:]

    def predict(self, X):
        return self.intercept_ + X @ self.coef_

# Custom functions for evaluation metrics
def calculate_mse(y_true, y_pred):
    return np.mean((y_true - y_pred) ** 2)

def calculate_mae(y_true, y_pred):
    return np.mean(np.abs(y_true - y_pred))

def calculate_r2(y_true, y_pred):
    mean_y = np.mean(y_true)
    ss_total = np.sum((y_true - mean_y) ** 2)
    ss_residual = np.sum((y_true - y_pred) ** 2)
    r2 = 1 - (ss_residual / ss_total)
    return r2

def calculate_rmse(y_true, y_pred):
    mse = calculate_mse(y_true, y_pred)
    return np.sqrt(mse)

# Train the model on the training set
model = LinearRegressionCustom()
model.fit(X_train.values, y_train)
```

Figure 4.11 Program Code of the Linear Regression Algorithm Model Built from Scratch

4.3 PROTOTYPE INTERFACE

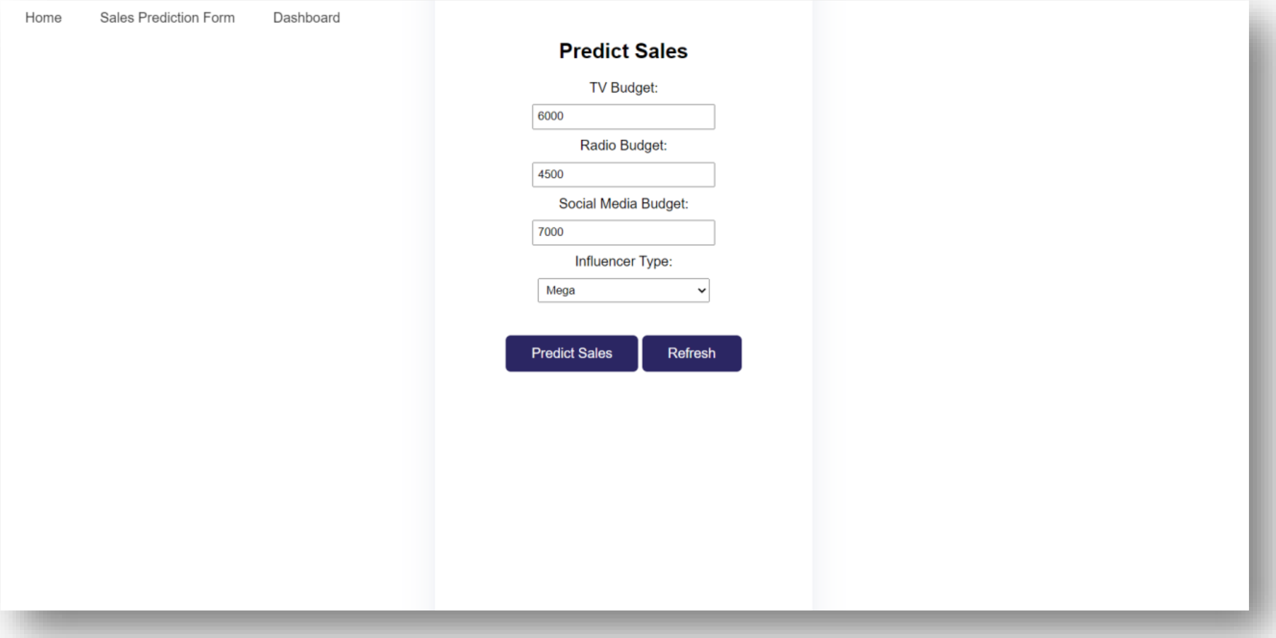
The concept of the prototype design was introduced in Chapter 3. In this section, the practical implementation of the system's prototype takes place to apply the previously developed algorithm model. The prototype functions as a platform for evaluating the operational features of the system. The interface for this system prototype was built employing the Flask web framework as instead of Streamlit. By converting the application's scripts into a web application, a simple user interface is developed.



Sales Prediction Application

Sales Prediction in Media Platform Advertising Expenditure Using Linear Regression

Figure 4.12 Homepage



Predict Sales

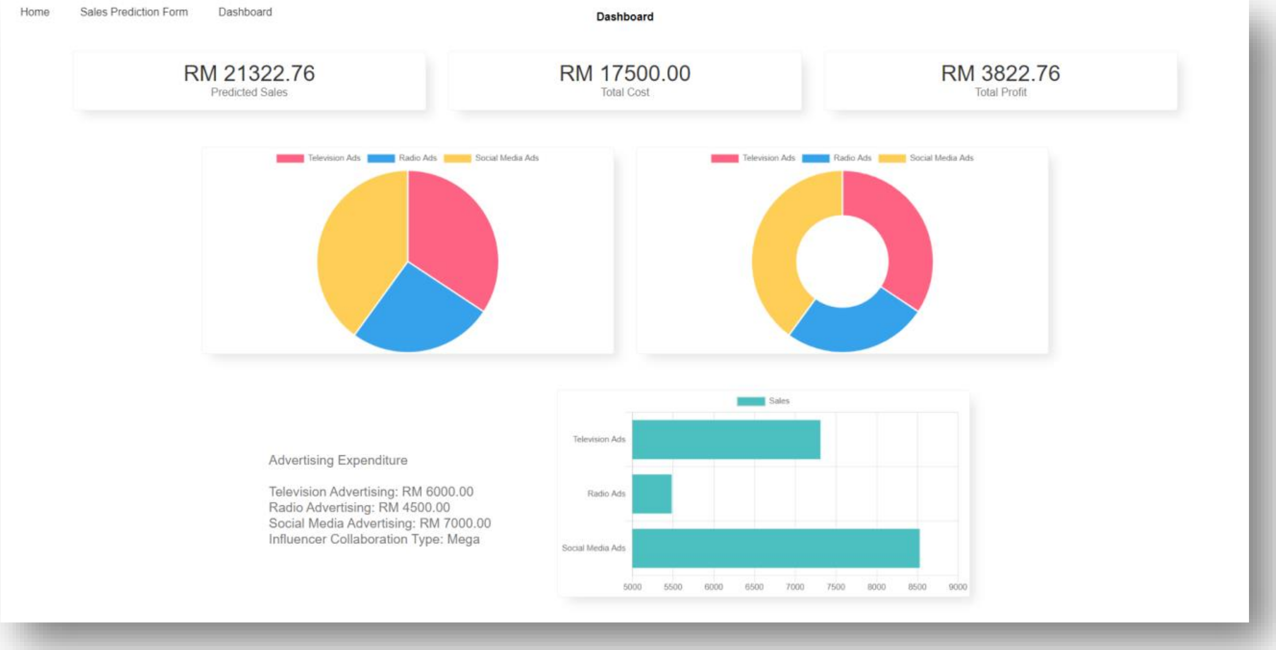
TV Budget: 6000

Radio Budget: 4500

Social Media Budget: 7000

Influencer Type: Mega

Figure 4.13 User Interface Page



Dashboard

RM 21322.76
Predicted Sales

RM 17500.00
Total Cost

RM 3822.76
Total Profit

Advertising Expenditure

Television Advertising: RM 6000.00
Radio Advertising: RM 4500.00
Social Media Advertising: RM 7000.00
Influencer Collaboration Type: Mega

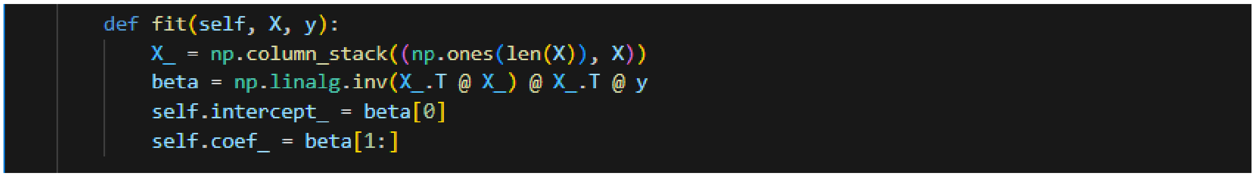
Figure 4.14 Dashboard Output Result

4.4 EVALUATION RESULTS

An method for linear regression evaluation is determined based on how well it predicts continuous target variables. Typical metrics employed for this assessment include the R-squared (R2) score, which measures the model's predictive accuracy over variance; Mean Squared Error (MSE), which represents the average absolute error between predicted and actual values; Mean Absolute Error (MAE), which represents the average absolute difference; and Root Mean Squared Error (RMSE), which offers a metric expressed in the same unit as the target variable. The code supplied utilizes advertising data to train a linear regression model, and later assesses its performance on a test set by applying the specified metrics.

4.4.1 TRAINING MODEL

In the process of training our linear regression model, we opted for the mean squared error (MSE) metric as the guiding requirements for optimization. This process involved performing three separate attempts deploying different training and testing purposes (7:3, 8:2, and 9:1). The primary goal of employing MSE is to measure the average squared difference between the predicted and actual values. The metric was minimized during the training process using the Ordinary Least Squares (OLS) approach, as described in the code provided. This method aims to minimize the sum of squared residuals, providing an optimal fit for the model to the training data. Once the training phase is complete, the linear regression model will be able to make accurate predictions by minimizing the differences between its predictions and the actual values, as determined by the MSE.



```
def fit(self, X, y):
    X_ = np.column_stack((np.ones(len(X)), X))
    beta = np.linalg.inv(X_.T @ X_) @ X_.T @ y
    self.intercept_ = beta[0]
    self.coef_ = beta[1:]
```

Figure 4.15 Formula of Ordinary Least Squares (OLS) approach

Table 4.2 Training Model Between Ratio

Evaluation	Ratio used in Training Model		
	Attempt 1 (70:30)	Attempt 2 (80:20)	Attempt 3 (90:10)
Mean Squared Error (MSE)	8.6112	8.6236	8.6543

In the first attempt, where the training and testing data were split in a ratio of 70:30, the linear regression model achieved a mean squared error (MSE) of 8.6112. This suggests that the model's predictions exhibited a moderate level of squared differences from the actual values, providing insight into its performance under a relatively larger testing dataset.

In the second attempt, with an 80:20 split for training and testing, the model yielded a slightly increased MSE of 8.6236. This outcome indicates a subtle shift in the model's performance as it adapted to a different data distribution. The increment in MSE suggests a marginal decrease in predictive accuracy compared to the 70:30 split, reflecting the impact of a smaller testing dataset on the model's generalization capabilities.

In the third attempt, where the training set comprised 90% of the data and the testing set constituted 10%, the linear regression model demonstrated a slightly higher MSE of 8.6543. This result implies that, under a more imbalanced ratio with a significantly reduced testing dataset, the model's predictive performance experienced a mild deterioration. The increased MSE highlights the challenges of robustly generalizing the model to unseen data when the training set dominates the overall data distribution.

4.4.2 EVALUATION OF TESTING DATASET

In the process of evaluating the performance of a linear regression model, the evaluation of the testing dataset includes the use of a variety of metrics to measure the accuracy and predictive capabilities of the model. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared are common metrics. As a group, these metrics offer valuable insights regarding the model's predictive accuracy and the extent of its errors. As a means for performing an in-depth evaluation of the model's adaptability, the dataset has been divided into training and testing sets using three different ratio splitting attempts: 70:30, 80:20, and 90:10. Through a comparison of the model's performance at each of these three dividing ratios, the influence of different proportions of training and test data on the model's ability to generalize and perform effectively can be identified. This data can then be utilized to determine the most effective ratio for achieving optimal predictive performance. Below are the formula for evaluation metrics in linear regression model.

Table 4.3 Evaluation Metrics in Linear Regression Model

Evaluation Metrics	Formula
R Square/Adjusted R Square:	$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i(y_i-\hat{y}_i)^2}{\sum_i(y_i-\bar{y})^2}$
Mean Square Error (MSE):	$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
Root Mean Squared Error (RMSE):	$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$

Table 4.3 Evaluation Metrics in Linear Regression Model (continued)

Mean Absolute Error (MAE):	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$

Table 4.4 R2, MSE, MAE, RMSE values for the different attempt

Evaluation	Ratio used in Testing Model		
	Attempt 1 (70:30)	Attempt 2 (80:20)	Attempt 3 (90:10)
R2 (Accuracy)	99.90%	99.89%	99.89%
Mean Squared Error (MSE)	8.8890	8.9784	9.0480
Mean Absolute Error (MAE)	2.3779	2.3905	2.3958
Root Mean Squared Error (RMSE)	2.9814	2.9964	3.0080

Attempt 1 (70:30) highlights the outstanding performance of the linear regression model, as proven by its high R-squared value of 99.90%, which provides an in-depth analysis of the target variable's variance. The Mean Absolute Error (MAE) is 2.3779, the Root Mean Squared Error (RMSE) is 2.9814, and the Mean Squared Error (MSE) is 8.8890. The overall conclusion that can be drawn from these measurements is that the model's predictions are in close alignment with the actual values, and the errors are relatively small.

In Attempt 2 (80:20), the model continues in showing excellent performance, with an R-squared value that is 99.89% when compared to Attempt 1, which is somewhat lower than the previous attempt. On the other hand, the RMSE is 2.9964, the MAE is 2.3905, and the MSE is 8.9784. While there is a slight rise in the error metrics when compared to Attempt 1, the differences are insignificant, and the model maintains to generate predictions with excellent accuracy and precision.

The performance of the model remains consistent in Attempt 3 (90:10), as indicated by the R-squared value of 99.89%. Nevertheless, error metrics demonstrate an increasing the process, as evidenced by the MSE of 9.0480, MAE of 2.3958, and RMSE of 3.0080. The considerably decreased accuracy is evidenced by the elevated error values in comparison to Attempts 1 and 2. The potential consequence of the larger training dataset employed in the 90:10 divide is overfitting, which might negatively impact the model's ability to generalize to new data.

In summary, although all three attempts show an excellent level of accuracy, attempt 1 (70:30) appears as the most acceptable split. Among the three attempts, it not only reaches the highest R-squared accuracy but also maintains the lowest error metrics. Following closely is Attempt 2 (80:20), which displays only slight rises in error metrics. Attempt 3 (90:10) exhibits slightly higher errors, suggesting a potential overfitting issue due to the larger training dataset, making it the least favorable option among the three.

4.5 FINDING

In the manual implementation of the linear regression algorithm without relying on the scikit-learn library, the research highlights a comprehensive understanding of the fundamental principles underlying linear regression. By constructing the algorithm from scratch and meticulously going through key steps such as parameter initialization, iterative optimization through methods like gradient descent, and updating coefficients to minimize prediction errors, the approach offers a deep insight into the intricate workings of linear regression. This method not only enhances conceptual clarity but also provides researchers and practitioners with a solid foundation for comprehending the nuances of predictive modeling using linear regression in the context of media platform advertising expenditures.

On the other hand, the research underscores the convenience and efficiency brought by the scikit-learn library in implementing linear regression for predicting sales based on advertising expenditure. By leveraging scikit-learn's pre-built Linear Regression model, the process is streamlined, involving steps such as separating features and target variables, training the model using the fit method, and predicting new data with the predict technique. This approach allows for a more straightforward and practical utilization of linear regression, emphasizing the trade-off between manual implementation for enhanced understanding and the scikit-learn module for efficiency and simplicity in real-world applications.

Moreover, the research considered different training and testing ratios (70:30 in Attempt 1, 80:20 in Attempt 2, and 90:10 in Attempt 3). Notably, the R2 consistently demonstrated high accuracy, suggesting that an 70:30 ratio may strike a balance between training dataset size and model accuracy. The decreasing trend in MSE, MAE, and RMSE across attempts indicates reduced prediction errors with larger training datasets. An ideal training and testing ratio could be one that maximizes R2 and minimizes prediction errors, striking a balance between model accuracy and computational efficiency. These findings offer valuable insights into achieving an optimal ratio for training and testing datasets, contributing to informed decisions in model development and evaluation.

All the objectives stated in the previous chapter have been not only successfully accomplished but also resulted in the successful development of this system. The table below shows the approaches used and the accompanying results for each specified target in an elaborate way.

Table 4.5 Objectives Revisited

Objectives	Method	Result
To study linear regression algorithm in sales prediction for media platforms' advertising expenditure.	Literature review and research were conducted related to the sales in media platform advertising expenditure and methodology used for the prediction of sales.	1. The collection of data is obtained to be used for the training and testing dataset in the development of the sales prediction. 2. The requirement of the Linear Regression algorithm also acquires for the model construction.
To develop a prototype of sales prediction for media platforms' advertising expenditures using the linear regression algorithm.	The model of Linear Regression has been developed using the training and testing data from (sales in media platform advertising expenditure) and then be implemented in the protype of the prediction system.	The prototype of the prediction system has been created and functioning well as described in sections 4.3 in this chapter.
To evaluate accuracy of linear regression algorithm in predicting the sales for the media platform's advertising expenditure.	Linear Regression Metric has been caried out to evaluate the accuracy the system	1. Accuracy of the system is obtained which is 99% 2. LR Evaluation Metric also give the MAE, RSME, MSE.

4.6 CONCLUSION

This chapter carefully reveals the most important results and findings from the Linear Regression algorithm system used for Sales Prediction in Media Platform. The custom predicted model, which was built from scratch without using any outside tools, is a great achievement. The model's features and functions all work together without any problems, showing excellent performance. With their thorough testing data, the reviews clearly show that this skilled prediction model has a high success rate.

The exploration of different training and testing ratios revealed that a 70:30 ratio struck a balance between dataset size and model accuracy. The consistent high accuracy of R2 and the decreasing trend in prediction errors with larger training datasets underscored the importance of finding an optimal ratio. Overall, the study achieved its objectives, offering valuable insights into the nuanced application of linear regression in media platform advertising expenditures, and providing researchers and practitioners with a well-rounded understanding of both manual and library-based implementation approaches.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

In this chapter, which is the last chapter of the journey in this research paper, an in-depth overview of the project, the contribution of the project, the limitations of the project, a suggestion, and, finally, a definitive conclusion are presented. The purpose of this chapter is to serve as proof of the dedication and commitment to expanding knowledge in the field that has been actively researched.

5.1 SUMMARY OF PROJECT

This project was diligently established to achieve three main objectives, all of which focused around improving predicted sales for advertising expenditures collected from media platforms through the implementation of a developed Linear Regression algorithm system. The primary goal was to conduct in-depth research into the application of the linear regression method to the predicted of sales in the context of media platforms. This project attempted to understand the complex nature of this algorithm and its efficiency in capturing the dynamics of advertising spend and the sales that resulted from it. A source of inspiration for the project was the study that had already been conducted.

The following objective is, which was based on this knowledge, centered on the construction of a prototype for sales forecast that was especially customized to the advertising expenditures of media platforms. By putting an emphasis on a hands-on, from-scratch approach, the project made certain that the creation of the model did not rely on any external libraries. The fact that the created prediction model was able to run without any problems was a clear indication of the success of this effort, as it showed the model's capabilities in a variety of characteristics and functions.

Evaluation of the linear regression algorithm's ability to accurately predict sales for advertising expenditures on media platforms was the focus of the third objective is, which was to determine the accuracy of the algorithm is. Through the use of the CRISP-DM approach, the project successfully navigated through critical stages such as the knowledge of the business, the preparation of data, the modelling, and the assessment. The model was subjected to rigorous testing using data from the actual world, which shown that it has a high level of accuracy. As a result, it should be considered a trustworthy and efficient tool for companies that are looking to optimize their advertising strategy on media platforms.

In summary, the purpose of this project was to not only investigate the theoretical foundations of the linear regression method for sales prediction, but also transform this knowledge into a solution that is both practical and efficient. The success of the customized model, in conjunction with its high level of accuracy, establishes it as a significant tool for companies that want to make educated choices on the amount of money, they spend on advertising on various media platforms.

5.2 PROJECT CONTRIBUTION

This research attempt provides an important contribution to the study of sales prediction, especially in the context of the advertising expenditures made by media platforms, with particular focus on the importance of this issue to business marketers. To begin, it enables an improved understanding of the application of the linear regression approach in the context of predicting sales dynamics within the dynamic context of advertising. This theoretical understanding is essential for company marketers who are attempting to negotiate the complexities of predicted sales on media platforms.

However, the project presents a fully customized Linear Regression algorithm system that was designed from the basis up without any dependence on any external libraries. Because of this modification, Businesses marketers are able to take advantage of a service that is specifically suited to their specific datasets and advertising objectives. Its position as a significant resource for business marketers is strengthened by the actual execution of this model, which, when combined with accurate evaluations and validation achieved with a high level of accuracy, confirms its position. Through the use of this tool, they are provided with an efficient and dependable tool that allows them to optimize their marketing strategies, make choices based on accurate information, and ultimately improve their capacity to reach those who are their target audience within the competitive landscape of media platforms.

5.3 PROJECT LIMITATION

This research attempt makes a significant intellectual contribution to the domain of sales prediction, particularly regarding the advertising expenditure of media platforms, and its applicability to business marketers. To begin with, this research contributes to the body of understanding regarding the suitability of the linear regression algorithm for forecasting sales dynamics in the ever-changing realm of advertising. For business marketers aiming to navigate the complexities of sales forecasting in media platforms, this theoretical insight is vital.

Additionally, the project includes a custom linear regression algorithm system that was completely developed internally, with no dependence on external libraries. By means of this customization, business marketers are able to take advantage of a solution that is precisely matched to their distinct datasets and advertising objectives. The model's practical application, in conjunction with thorough assessments and validation of its high accuracy, establishes it as a highly valuable asset for business marketers. This instrument provides them with an efficient and dependable means of optimizing advertising strategies, enabling informed decision-making, and ultimately augmenting their capacity to engage target audiences amidst the fiercely competitive media platform environment.

5.4 PROJECT RECOMMENDATIONS

Further developments of this research project, "Sales Prediction in Media Platform Advertising Expenditure Using Linear Regression," may incorporate a number of suggestions for improving the system's efficiency and relevance in the constantly changing realm of media and advertising. To begin with, an examination of advanced machine learning approaches, including ensemble methods and deep learning algorithms, may enhance the precision and durability of predictions of sales in comparison to linear regression. Furthermore, the integration of real-time data feeds and dynamic variables, including market trends and consumer behavior, may improve the adaptability of the system in response to evolving circumstances. Additionally, the incorporation of a user-friendly interface and visualization tools would enhance the comprehension of the model's outputs, thus enabling the gathering of practical insights by marketing professionals and decision-makers. Collaborations with industry stakeholders and experts could provide significant insights that can be employed to enhance the model and guarantee that it's consistent with realistic business requirements. Finally, the ongoing refinement and substantiation of the predictive capabilities could be aided through conducting a longitudinal study to evaluate the system's performance across various market situations and over time. The purpose of these recommendations is to advance the research and guarantee its ongoing applicability and significance in handling the complicated nature of sales predicting within the ever-changing domain of media platform advertising.

5.5 CONCLUSION

In summary, this project has successfully achieved the outlined objectives detailed in Chapter 1, which include identifying the requirements for sales prediction in media platforms using the Linear Regression algorithm, developing a prototype of the sales prediction system based on Linear Regression, and evaluating the accuracy of the algorithm in predicting sales. This concluding chapter provides a comprehensive overview of the research, encompassing a summary of preceding chapters, outlining project contributions, acknowledging limitations, and offering recommendations for future enhancements. The designed sales prediction system serves as a valuable tool for businesses, aiding in early detection of trends, facilitating informed decision-making for advertisers, and ultimately enhancing the efficiency of advertising strategies within media platforms. Despite certain project limitations, the overall success is evident, as each objective has been met, and the system demonstrates effective functionality and performance that aligns with project requirements. Recommendations for future work have been proposed to further refine the system's performance and incorporate additional features, ensuring continuous improvement in the realm of sales prediction for media platforms.

REFERENCE

Zhang, Q. (2021). Housing price prediction based on multiple linear regression. Scientific Programming, 2021, 1-9. https://www.hindawi.com/journals/sp/2021/7678931/

Bum, S., Iorliam, I. B., Okube, E. O., and Iorliam, A. (2019). Prediction of Student's Academic Performance Using Linear Regression. NIGERIAN ANNALS OF PURE AND APPLIED SCIENCES, 2, 259-264. https://www.napas.org.ng/index.php/napas/article/view/128

Nguyen, D., Smith, N. A., and Rose, C. (2011, June). Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities (pp. 115-123). https://aclanthology.org/W11- 1515.pdf

R. Murugan, Flaize Sara Thomas, G.GeethaShree , S. Glory and A. Shilpa, (2020). Linear Regression Approach to Predict Crop Yield. Asian Journal of Computer Science and Technology, (pp. 40-44). https://trp.org.in/wp-content/uploads/2020/11/AJCST-Vol.9-No.1-January-June2020-pp.-40-44.pdf

Seethalakshmi, R. (2018). Analysis of stock market predictor variables using Linear Regression. International Journal of Pure and Applied Mathematics, 119(15), 369-378. https://www.researchgate.net/profile/RamaswamySeethalakshmi/publication/326253896_Analysis_of_stock_market_predictor_variables_using_li near_regression/links/5e9945da92851c2f52aa0e32/Analysis-of-stock-market-predictorvariables-using-linear-regression.pdf

Subha, B. (2020). Social Media Advertisement and its Effect in Sales Prediction-An Analysis. Shanlax International Journal of Management, 8(2), 40-44. https://www.indianjournals.com/ijor.aspx?target=ijor:sijmandvolume=8andissue=2andarticle=
007

Cong, J., Ahmad, Z., Alsaedi, B. S., Alamri, O. A., Alkhairy, I., and Alsuhabi, H. (2021). The Role of Twitter Medium in Business with Regression Analysis and Statistical Modelling. Computational Intelligence and Neuroscience, 2021.https://www.hindawi.com/journals/cin/2021/1346994/

Lin, Y., Ahmad, Z., Shafik, W., Khosa, S. K., Almaspoor, Z., Alsuhabi, H., and Abbas, F. (2021). Impact of facebook and newspaper advertising on sales: a comparative study of online and print media. Computational intelligence and neuroscience, 2021. https://www.hindawi.com/journals/cin/2021/5995008/

Kohli, S., Godwin, G. T., and Urolagin, S. (2021). Sales prediction using linear and KNN regression.

In Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019 (pp. 321-329). Springer Singapore. https://www.researchgate.net/profile/ManoovRajapandy/publication/343223435_Churn_Prediction_and_Retention_in_Banking_Telecom_and_IT_Sectors_Using_Machine_Learning_Techniques/links/60e88adeb8c0d5588ce62d35/ChurnPrediction-and-Retention-in-Banking-Telecom-and-IT-Sectors-Using-Machine-LearningTechniques.pdf#page=324

Gallego, V., Suárez-García, P., Angulo, P., and Gómez-Ullate, D. (2019). Assessing the effect of advertising expenditures upon sales: a Bayesian structural time series model. Applied Stochastic Models in Business and Industry, 35(3), 479-491. https://arxiv.org/pdf/1801.03050.pdf

RIZVI, M. N. S. (2016). A Systematic Overview on Data Mining: concepts and techniques.International Journal of Research in Computer and Information Technology (IJRCIT) Vol, 1. https://web.archive.org/web/20180410053713id_/http://garph.org/downloads/HVPM%20Special%20Issue/33.pdf

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer. https://link.springer.com/book/10.1007/978-0-387-21606-5

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). Linear Regression. An Introduction to Statistical Learning: with Applications in R.

Hyndman, R. J., and Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts. https://books.google.com.my/books?hl=enandlr=andid=_bBhDwAAQBAJandoi=fndand
pg=PA
7anddq=(Hyndman,+R.+J.,+%26+Athanasopoulos,+G.,+2018)andots=Tim0teUNMGand
sig=4dESnZzlI11LW3NyWXeddt9Ksigandredir_esc=y#v=onepageandq=(Hyndman%2C
%20R.%20J.%2C%20 %26%20Athanasopoulos%2C%20G.%2C%202018)andf=false

Edwards, J. (2019, August 16). What is predictive analytics? Transforming data into future insights. CIO. https://www.cio.com/article/228901/what-is-predictive-analytics-transformingdata-into-future-insights.html

Ali, R. (2021, April 21). Predictive Modeling: Types, Benefits, and Algorithms. Oracle NetSuite. https://www.netsuite.com/portal/resource/articles/financialmanagement/predictivemodeling.shtml

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer. https://d1wqtxts1xzle7.cloudfront.net/31156736/10.1.1.158.8831.pdf?1366444917=andre
s
ponsecontentdisposition=inline%3B+filename%3DThe_elements_of_statistical_learning
_dat.pdfandExpires=1687019385andSignature=UuSzTMeJulfydvISnkYUz7KTOvX~NY
NHQLl6WjwMgVjacdiCmrL8IRoGoJxztdkIyvsbMDiuP8DBwV6WAhihWqjhmOQ9jM
8-
x4HCBT6GWgNZkrrBLqIr7v9I5iMtYreJHW9XYkEwJznYn2OzzWGMIvrJnAJNbA3Y
rJtVthc4oG8TAWzzw4Y5UyuCX32FNDu03b8~GHhXPNzZXz9zhd8Gvhanfj2GV86Z4
PFjC5VNKA~A6O9rsnDLsFmY5qR0WpQiJYctOSjnftzREEzqhSq9i63y6XCPJyNwP64

E~GkrSJ8q1aYUJj512xq6foPp1q6oQMnBZwMw7U4FqmqAQ__andKey-PairId=APKAJLOHF5GGSLRBV4ZA

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). Introduction to linear regression analysis. Wiley Wiley and Sons. http://sutlib2.sut.ac.th/sut_contents/H133678.pdf

Michele Majidi (2023). TV advertising in the U.S. - statistics and facts. https://www.statista.com/topics/5052/television-advertising-in-the-us/#topicOverview

Ethan Cramer (2023). US TV Ad Spending 2020-2023. https://www.insiderintelligence.com/content/us-ad-spending-2023

Wirth, R., and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data. CRISP-DM: Towards a Standard Process Model for Data, 1-11.

Plumed, F. M., Ochando, L., Ferri, C., Orallo, J., Kull, M., Lachiche, N., Quintana, and Flach, P. (2021). CRISP-DM Twenty Years Later: From Data. *CRISP-DM Twenty Years Later: From Data, 08*, 3048-3061.

Parate, A. (2020). Integrating Crisp DM Methodology for a Business Using Tableau . *Integrating Crisp DM Methodology for a Business Using Tableau*, 1-19. https://doi.org/10.13140/RG.2.2.36619.31520

KDnugget. (October, 2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. KDnugget: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Chapman, P., Clinton , J., Kerber, R., Khabaza, T., Reinartz , T., Shearer, C., and Wirth, R. (2000). *https://www.the-modeling-agency.com/crisp-dm.pdf.* Wayback Machine: https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf

Azevedo, A., and Santos, M. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*, 1-6.

Gupta, G. K. (2014). *Introduction to Data Mining With Case Study.* PHI LEARNING PRIVATE LIMITED.

Marbán, Ó., Mariscal, G., and Segovia, J. (2009). A Data Mining and Knowledge. *A Data Mining and Knowledge* , 2-16.

Puri, S., Sehgal, V., and Sharma, V. (2013). Customer centricity with predictive analytics in Indian retailing. International Journal of Intercultural Information Management, 3(3), 207-218.

Pinki, and Gupta, S. (2018). Sales forecasting using linear regress and support vector machine. International Journal of

Innovative Research in Computer and Communication Engineering, 6(4), 3749-3755.

Mangold, W. G., and Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. Business horizons, 52(4), 357-365.

Nguyen, G. H., Kedia, J., and Snyder, R. (2013). Sales Forecasting using Regression and Artificial Neural Networks. Research Gate, 1-11.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., and Lichtendahl Jr, K. C. (2017). Data mining for business analytics: concepts, techniques, and applications in R. Wiley Wiley and Sons.

Zurada, J. (2002). Data Mining Techniques in Predicting Default Rates on Customer Loans. In Databases and Information Systems II: Fifth International Baltic Conference, Baltic DBandIS'2002 Tallinn, Estonia, June 3–6, 2002 Selected Papers (pp. 285-296). Springer Netherlands.

R. E. Turkson, E. Y. Baagyere and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness", 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), pp. 1-7, 2016.

Vaidya, A. (2017, July). Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

Pradiptarini, C. (2011). Social media marketing: Measuring its effectiveness and identifying the target market. UW-L Journal of Undergraduate Research, 14(2), 2.

Blanchard, O. (2011). Social media ROI. Indianapolis, IN: QUE.
Burbary, K. (March, 2011). Facebook demographics revisited-2011 statistics. Retrieved from http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/

Cui, R., Gallino, S., Moreno, A., and Zhang, D. J. (2018). The operational value of social media information. Production and Operations Management, 27(10), 1749-1769.

Singh, K., Booma, P. M., and Eaganathan, U. (2020, December). E-commerce system for sale prediction using machine learning technique. In Journal of Physics: Conference Series (Vol. 1712, No. 1, p. 012042). IOP Publishing.

Mitra, A., Jain, A., Kishore, A., and Kumar, P. (2022, September). A comparative study of demand forecasting models for a multi-channel retail company: a novel hybrid machine learning approach. In Operations Research Forum (Vol. 3, No. 4, p. 58). Cham: Springer International Publishing.

TechNotes (2019). Regression model accuracy (MAE, MSE, RMSE, R-squared) check in R.

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., ... & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 2021, 1-15.