AN ANALYTIC DASHBOARD FOR FINANCIAL PERFORMANCE AND PREDICTION IN HEALTHCARE DISCIPLINES.

PRIYADHARSHNI A/P MOHAN NATHAN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

AN ANALYTIC DASHBOARD FOR FINANCIAL PERFORMANCE AND PREDICTION IN HEALTHCARE DISCIPLINES.

PRIYADHARSHNI A/P MOHAN NATHAN

This report is submitted in partial fulfillment of the requirements for the Bachelor of [Computer Science (Artificial Intelligence)] with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2024

DECLARATION

I hereby declare that this project report entitled

AN ANALYTIC DASHBOARD FOR FINANCIAL PERFORMANCE AND PREDICTION IN HEALTHCARE DISCIPLINES.

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : ______________ ________________ Date : 12/06/24
PRIYADHARSHNI A/P MOHAN NATHAN

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of Bachelor of [Computer Science (Software Development)] with Honours.

SUPERVISOR : PROF MADYA DR ZERATUL IZZAH Date : 4/9/2024
([NAME OF THE SUPERVISOR])

DEDICATION

This report is wholehearted to my respectful family, who cheered me on and have been very motivational in my journey. To my beloved mother, whose strong support and love have been source of inspiration to complete the report, I extend my deepest gratitude and love. I also would like to acknowledge my sisters and friends for their precious support and understanding. Their readiness in providing listening ears and a helping hand to exchange opinions and ideas has made this journey more meaningful. Thank you all for your presence and significant contributions to my personal and professional growth.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to Professor Madya Ts. Dr. Zeratul Izzah Binti Mohd Yusoh for her invaluable guidance and support throughout the completion of my final year project. Her insights and assistance have been crucial to the successful realization of this work, and her mentorship has significantly enhanced my knowledge in this field. Additionally, I extend my deepest thanks to my parents for their unwavering support and motivation, which have been the driving force behind this project. Their encouragement and love have provided a constant source of strength and inspiration.

ABSTRACT

In the rapidly evolving healthcare industry, effective financial management is crucial to adopt innovative strategies to ensure financial viability. However, many healthcare financial administrators still rely on the conventional approaches which fails in providing the detailed insights necessary for financial planning dan decision-making. This project addresses the gap of the conventional method by developing and implementing analytics dashboards enhanced with machine learning model. The objectives were to evaluate and identify limitations of traditional financial management methods, develop machine learning model for revenue prediction and forecasting and design an interactive analytics dashboard to provide actional insights. The machine learning model are used for two purposes which are prediction using Random Forest Regression, XGBoost Gradient Boosting and Decision Trees and ARIMA model was used for forecasting next year's revenue. For prediction, Random Forest Regression demonstrated the best performance with Training Mean Squared Error (MSE) of 0.0016 and a Testing MSE of 0.0020, with Training and Testing R^2 scores of 0.6384 and 0.5419, respectively. The interactive dashboard designs provide dynamic visualization and real-time analytics facilitates the financial performance monitoring and make informed decisions. This project underscores the importance of integrating advanced predictive analytics and interactive data visualization tools in the financial management practices of healthcare facilities.

ABSTRAK

Dalam industri kesihatan yang berkembang pesat kini, pengurusan kewangan menggunakan strategi inovatif dan berkesan amat penting untuk memastikan industri ini terus mempunyai potensi yang tinggi. Walaubagaimanapun, masih banyak pentadbir kewangan industri kesihatan masih bergantung pada pendekatan kovensional yang gagal memberikan pandangan terperinci yang diperlukan untuk membuat perancangan kewangan dan membuat keputusan. Projek ini bertujuan untuk menangani jurang kaedah kovensional dengan membina papan pemuka analitik yang dipertingkatkan dengan model pembelajaran mesin. Objektifnya adalah untuk mengkajidan mengenal pasti bidang utama dalam pengurusan kewangan dalam industri kesihatan, membangunkan model pembelajaran mesin untuk ramalan dan ramalan hasil sambil mereka papan pemuka interaktif untuk memberikan cerapan tindakan. Model pembelajaran mesin digunakan untuk dua tujuan iaitu ramalan menggunakan Random Forest Regression, XGBoost Gradient Boosting dan Decision Trees dan model ARIMA digunakan untuk meramalkan hasil tahun hadapan. Untuk ramalan, Random Forest Regression menunjukkan prestasi terbaik dengan Latihan Mean Squared Error (MSE) sebanyak 0.0016 dan MSE Ujian sebanyak 0.0020, dengan skor Latihan dan Ujian R^2 masing-masing 0.6384 dan 0.5419. Reka bentuk papan pemuka interaktif menyediakan visualisasi dinamik dan analisis masa nyata memudahkan pemantauan prestasi kewangan dan membuat keputusan termaklum. Projek ini menekankan kepentingan mengintegrasikan analitik ramalan lanjutan dan alat visualisasi data interaktif dalam amalan pengurusan kewangan kemudahan penjagaan kesihatan.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

LIST OF ABBREVIATIONS

FYP	**-**	**Final Year Project**
ARIMA	**-**	**AutoRegressive Integrated Moving Average**
LSTM	**-**	**Long-Short Term Memory**
AI	**-**	**Artificial Intelligence**
OP	**-**	**Outpatient**
IP	**-**	**Inpatient**
KPI	**-**	**Key Performance Metrics**
KJMC	**-**	**Kelana Jaya Medical Centre**

Chapter 1: INTRODUCTION

1.1 Introduction

In the rapidly evolving healthcare industry, effective financial management is crucial to ensure a sustainability and growth of healthcare facilities. Financial sectors need to adopt innovative strategies and utilize advances tools that provide accurate forecasts and actionable insights. However, many administrators in this industry still rely on traditional financial management methods often fall short in delivering the precision and adaptability requires for contemporary challenges. These includes recording financial transactions, maintaining spreadsheets and compiling financial reports by using basic software like Excel, using Power Point Presentation for insights presentations and forecasting are performed using expert judgment. To overcome this constraint and improve financial management in healthcare facilities, there is a need to embrace advanced technologies such involvement of AI such as predictive or forecasting analysis, and interactive dashboards.

1.2 Problem Statements

In the healthcare industry, effective financial management is important for building sustainable financial growth in the healthcare industry as the demand of this industry rapidly increasing. However, the current financial administrator practices conventional financial management, which leads to a lack in providing actionable insights and adaptability for healthcare operations.

One of the conventional methods used is Excel for revenue calculations to forecast future revenue that is incapable to capture the relationship between numerous factors in the data such as date, patient admission type, hospital disciplines or doctor's performance which indirectly influence the forecasted revenue. This approach with statistical methods does not provide important insights for accurate revenue forecasting which affects the reliability of the forecasted revenue.

Next, the conventional method further highlights the critical need of analytics dashboard development and implementation. The traditional monthly revenue report using Power Point Presentation provides limited data representation, visual and analysis. Without these capabilities, the healthcare management struggles to monitor the financial performance as they are unable to capture trends and patterns of the revenue to informed decisions.

1.3 Objective

- To study and identify the limitations of traditional financial management methods.

- To develop and implement machine learning model for revenue prediction and forecasting.

- To design and develop revenue analytics dashboard

1.4 Scope

1.4.1 Modules

The project has three main modules to be implemented as shown in the table below. The first module focuses on predicting revenue using machine learning techniques. It involves implementing models such as Random Forest Regression, XGBoost Gradient Boosting, and Decision Trees for revenue predictions based on various input factors. The second module is dedicated to forecast future revenue using the time series machine learning model using ARIMA (AutoRegressive Integrated Moving Average) model. This module analyzes historical data and forecast revenue. Then, the third module involves developing a revenue analytics dashboard using Power BI. This dashboard provides insights on the revenue based on disciplines, admission type, doctor’s performance as well insights on the predicted and forecasted revenue using machine learning models.

Table 1.1: Modules in project

Modules	Functions
Revenue Prediction Module	This module utilizes Random Forest Regression, XGBoost Gradient and Decision Tree to predict revenue based on several features in the historical data.
Revenue Forecasting Module	This module uses ARIMA (AutoRegressive Integrated Moving Average) time series machine learning model to forecast the future revenue by analyzing the historical data over date.
Analytics Dashboard	This module focuses on developing an interactive and dynamic analytics dashboard using Power BI on revenue data.

1.4.2 Target User

The target user for this project will be financial administrators who are responsible for managing the financial operations within healthcare facilities. Financial administrators can make use of this project to present the financial performance to the senior management such as CEOs or CFOs using the analytics dashboard to analyze the financial performance and make strategic decisions.

1.5 Project Significance

The project is significant with its ability to transform effective financial management within healthcare facilities by introducing new technological solutions. One of the primary benefits is the enhancement of revenue accuracy by incorporating predictive machine learning models such as Random Forest Regression, XGBoost Gradient Boosting, and Decision Trees to improve the accuracy of revenue predictions by involving several features that might influence the revenue. With this enhancement, the healthcare facilities will be able to make well-informed financial decisions that reduces the risk of budget shortage and financial mismanagement.

Another significant aspect of the project is that through the development and implementation of time series machine learning model which is ARIMA anticipate future financial trends by analyzing historical data. This forecast impacts the financial management to prepare for potential for resource allocation and strategically plan for future financial growth, ensuring that the organization financial performance is durable and adaptable over time.

The project also contributes to data-driven decision-making within healthcare organizations. The development of a revenue analytics dashboard using Power BI empowers financial administrators to analyse and to visualize the complex financial data in a user-friendly format. In addition to these benefits, the project enhances operational efficiency by automating financial processes by ensuring that financial data is always up-to-date and available for decision-making. This refinement is vital for the smooth and reliable operation of the healthcare facilities.

1.6 Expected Output

The expected output for this project includes revenue prediction with high performance machine learning performance that will be validated with performance metrics to ensure the accuracy of revenue prediction. Other than that, the expected output for this project will be to validate the reliability in using time series machine learning model, ARIMA model to forecast revenue compared to the statistical method used by the financial administrator. The final expected output for the project will be producing a Power BI dashboard that is able to visualize the revenue analysis with user-friendly features allowing users to gain real-time insights on financial performance.

1.7 Report Organisation

Chapter 1 discuss about the overview of the project including the background, objectives of the project, the project scope and its significance.

Chapter 2 discuss on the existing system, literature review, techniques and the methodology chosen for the project. This chapter also highlights on the project requirement including the software, hardware and other resources needed.

Chapter 3 focuses on analyzing the problem statement. This chapter explores the current stakeholder system and how it functions. Then dives into project's requirements including data requirements, functional requirements, non-functional requirement and software hardware requirements.

Chapter 4 discuss on the design of the project which covers the user interface and database design. Additionally, the AI components in the projects are explained detail including the techniques and how the technique is included in the project design.

Chapter 5 presents the results of the project discussing the significance of the project which involves the evaluation of the AI technique used and the functional requirement testing.

Chapter 6 focusses on the project limitation and strengths, and the way to enhance the project in the future.

1.8 Summary

Traditional financial administrators in the hospital industry play a vital role but often constrained by outdated methods and tools. To overcome these limitations and improve financial management, there is a need to adopt to predictive techniques and interactive dashboard. These tools can be a comprehensive method to enhance the financial performance in healthcare industry.

CHAPTER 2: LITERATURE REVIEW AND PROJECT METHODOLOGY

2.1 Introduction

This chapter provides a detailed examination of the domain, existing systems, and methodologies relevant to the project. It begins with an exploration of Kelana Jaya Medical Centre's current financial management practices, highlighting their reliance on traditional methods and limitations in providing actionable financial insights. The chapter then continues with the reviews on existing healthcare and financial dashboards and the research conducted on revenue performance enhancement using machine learning techniques. Then the chapter moves to outline the methodology used to develop and implement the project successfully.

2.2 Facts and Finding

2.2.1 Domain

Kelana Jaya Medical Centre is a healthcare industry that has provided a wide range of high-quality healthcare services since 1999. Despite being a leading private healthcare facility, Kelana Jaya Medical Centre still practices traditional financial management methods that lacks the precision and adaptability of the financial management in the rapidly growing and demanding healthcare industry. Currently, financial management uses a conventional approach by employing Excel and PowerPoint presentations to analyze and visualize revenue performance. This visualization and analysis are typically limited to past month data, lacking the depth and foresight necessary for proactive financial planning and strategic decision-making.

2.2.2 Existing System

There are some related existing systems in producing an interactive dashboard for the principal metrics in this healthcare industry. However, there is very limited healthcare dashboard that emphasizes on the financial performance in this field especially on providing insights on the revenue and targeted revenue. Hence, my project will be an initiative to implement real time interactive dashboard, which focuses in providing the insights on financial performances in the healthcare industry with a broader financial performance coverage.

Datapine.com is a company that provides hospital KPI dashboard offerings key metrics on various categories needed to track the patients in the hospital such as number of visitors over certain period of time, cost for treatments and waiting time in the emergency room before seeing a doctor based on division that able provide insights on the average ER waiting time, staff allocation and number of visitors. Despite these useful features, Datapine’s dashboard also has limitation where it does not provide detailed insights on the cost based on the division and admission type, nor does it compare the hospital’s target revenue versus the revenue achieved. This missing functionality restricts a comprehensive financial analysis, which is necessary for strategic financial planning and performance enhancement.

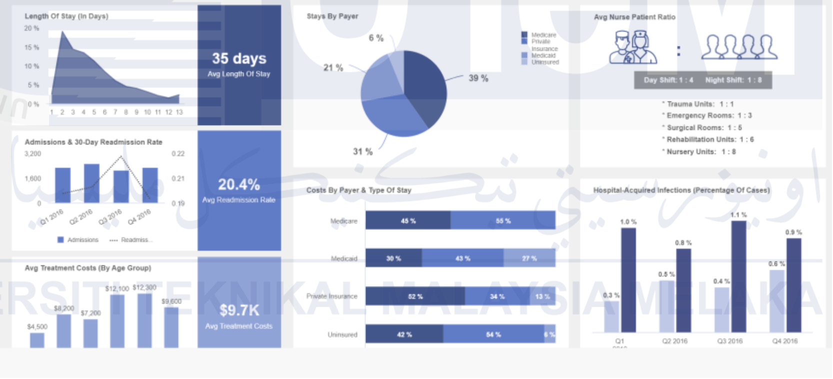
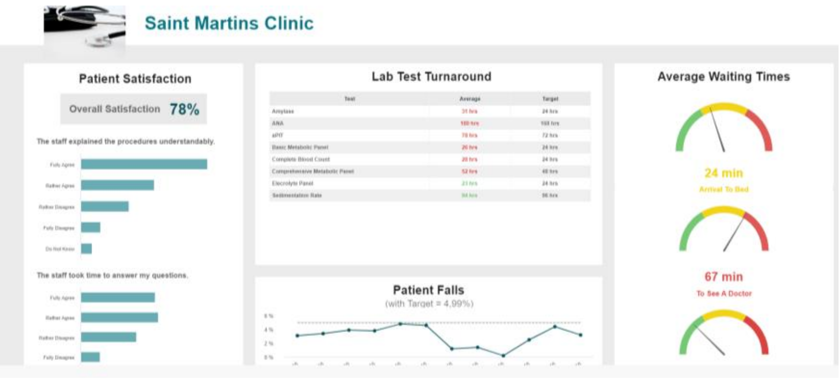


Figure 2.2: Datapine’s hospital KPI dashboard



Test	Average	Target
Amylase	31 hrs	24 hrs
ANA	180 hrs	168 hrs
aPTT	78 hrs	72 hrs
Basic Metabolic Panel	26 hrs	24 hrs
Complete Blood Count	28 hrs	24 hrs
Comprehensive Metabolic Panel	52 hrs	48 hrs
Electrolyte Panel	21 hrs	24 hrs
Sedimentation Rate	94 hrs	96 hrs

Figure 2.3: Datapine’s clinical patient satisfaction dashboard

Sofweb Solution.com is another company that provides financial dashboard for healthcare industry where the dashboard consists of functionality of providing actual versus target cost by the hospital division, average patient treatment cost, expenses revenue trend where the revenue is calculated based on the total treatment cost minus with the expense and number of patients including Inpatient and Outpatient. However, the dashboard has notable disadvantage. It does not provide an analysis of achieved versus target revenue which is crucial for gaining deeper insights into financial performance and identifying gaps between projected and actual financial outcomes. This lack of detailed revenue analysis limits the ability of healthcare facilities to make informed decision aimed at improving financial health and strategic planning.

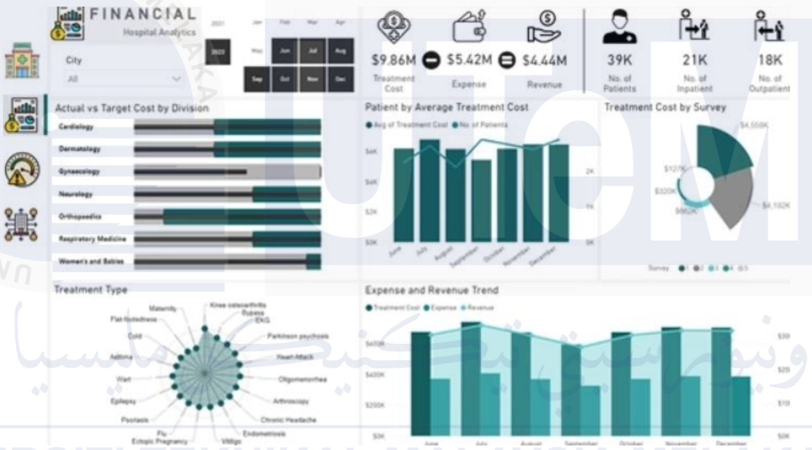


Figure 2.4: Sofweb Solution.com financial dashboard

Table 2.1: Summarize Existing System

Name of the system	Description	Advantages	Disadvantages
Hospital KPI Dashboard **(datapine.com)**	Dashboard that gives the principal metrics on the major category needed to track the patients in the hospital:	- Able to provide insights on the financial metric on the cost to reduce overwhelming cost - Able to provide insights decrease	- Does not provide detailed insights on the cost based on the division - Does not provide on the number of staff allocated on

	- Number of visitors over certain period of time - Cost for treatments - Waiting time in the emergency room before seeing a doctor based on division	the average ER wait time - Measure on the number of visitors to allocate staff more efficiently	each division for recruitment purpose/ shortage of staff in division - Does not provide average time for get a treatment and to admitted to wad
Healthcare Marketing Dashboard **(dashthis.com)**	Dashboard that provides easy to use automated marketing to spend less time on the healthcare marketing and more time focusing on the patients	- Provide insights on the website performances such as traffic and number of SEO. - Provide insights on the cost spending on the marketing - Provides insights on the website appointment forms completed	- Does not provide insight details on the appointment based on the selected division - Does not show the SEO based on the hospital division.
Orthopedic Clinical Variation Dashboard **(boldbi.com)**	Dashboard that provides insights on particular variation such as Orthopedic Clinical	- Provides insights on bed occupied and the average length of stay. - Provide insights on admissions and readmissions break down - Provide insights on the patient population break down by urgency type.	- Does not provide insights on the staff-based patients allocation or appointments

		- Provide insights on the medical costs	
Patient Management Analytics Dashboard for Hospitals **(spec-india.com)**	Dashboard that provides insights on the number of total patients, patients in ICU, patients who died, re-admit patients, average days of discharge, patients by gender & age group.	- Provides insights on patient's background such as the city based. - Provides insights on the average waiting time by division. - Provide insights into the number of patients by hospitals, division, LOS Bucket, and discharge type	- Does not provide information on the most visited division and the doctor in the division.

Clinician dashboard **(softweb Solutions .com)**	Dashboard that offers a comprehensive view of how a department manages its workload.	- Identification of imbalances in clinician distribution - Assessment of individual employees based on various factors - Assessment of individual employees based on various factors	Does not provide insight on performance of each employee by the completion rates and number of visits for those particular employees.
Patient No-Shows Dashboard **(sisence.com)**	Dashboard that increases the number of patients seen by our providers and reduce the number of patients that do not show or cancel last minute.	- Able to identify the breakdown of appointments by a variety of patient. - Able to predict what patients may be more likely to cancel so we can schedule accordingly. - Able to send appointment reminder text	- Does not provide insights on the reason behind breakdown appointments - Does not provide insights on the most appointment breakdown division and employee
Mayo Clinic Health Care System **(asquare Technologies web.com)**	Dashboard that reports hospital, specialties, doctors, staff, patients as well as Covid 19.	- Able to show insights on the bed occupancy details and patients' statistics in for every department - Provides insights on specialties and services provided for every department with	- Does not provide insights on which doctor attended which patients - Provides basic information only for both patients and doctors.

		the doctors' details. - Provides insights on covid statistics on overall and in each department	

2.2.3 Techniques

2.2.3.1 ARIMA Univariate Time Series Forecasting

ARIMA is a foundational method in time series forecasting that combines autoregression (AR), differencing (I), and moving averages (MA). The ARIMA model is particularly suited for datasets where the values are dependent on previous values and can account for trends and seasonality through differencing and seasonal adjustments (SARIMA).

Mustafa Afeef, Anjum Ihsan and Hassan Zada (2018) on their research for forecasting stock prices through Univariate ARIMA Modeling that highlights applicability of ARIMA predicting the values of a variable and to investigate which type of prediction, short-term or long-term is best provided. The studies are significant to help investors decide when to invest in each company's stock. Their finding was ARIMA has a very good prediction for a short run however long-term prediction using lagged values only makes a little sense.

Other than stock time series prediction, Mindaugas Česnavičius (2020) provides a studies on electricity market price forecasting model based on univariate time series analysis where he emphasize on short-term electricity price forecast by using average, seasonal naive and exponential smoothing methods and their findings were the monthly periods between the time series allow to make a meaningful one-year forecast without forecast repetition which is not possible to be done using hourly, daily or weekly time periods. The model used for this research are ARIMA model, SARIMA and weighted SARIMA model and the studies suggested further research of a considering models which includes external factors to predict the price.

2.2.3.2 Random Forest Regression Prediction

Random Forest Regression is a machine learning technique that leverages an ensemble of decision trees to make predictions. It operates by combining the predictions of multiple decision trees to reduce overfitting and enhance accuracy of the traditional decision tree prediction model. In this approach, the tree is built separately with different subsets with random data and the final output is derived from the average or weighted average of all individual trees' predictions.

In a 2005 study, "Predicting customer retention and profitability by using random forest and regression forest techniques," by Bart Larivie`re and Dirk Van den Poel that focuses on customer retention and profitability using random forests and regression forest model. The study analyzed data from 100 000 customers of European financial services companies and their findings was random forest model outperforms the traditional regression model in prediction accuracy. The key variable used for the prediction was the customer behavior including past behavior, customer demographics and intermediary factors.

The next research by Nikolay Lomakin et al. (2023) explores the usage of random forest and regression algorithm to model the profit forecasting for the Russian Banking Sector. His research includes macroeconomic indicators from the year 2017 to 2021to forecast the profit of the selected banking sector using random forest machine learning model and a neural network regression model. His findings show that the random forest model achieved a mean absolute error 61% lower than the linear regression model. This indicates that the random forest model is reliable in making predictions and applied in real-world scenarios.

Narayana Darapaneni, Sreelakshminarayanan Muthuraj, Prabakar K and Madhavan Sridhar in the International Conference on Communication and Signal Processing, April 4-6, 2019, focus on analyzing logistics data to predict demand and revenue using machine learning using logistics data with the features latitude, longitude, hour of the day and day of the week. In their research, they focused on implementing AI that can enhance transportation and logistics operations through effective forecasting techniques. Their findings are multivariate time series are good in forecasting revenue while random forest is good in demand prediction.

2.2.3.3 XGBoost Gradient Boosting Prediction

XGBoost is a versatile machine learning library that operates within the Gradient Boosting framework. The key feature of this model is that it builds decision trees to make predictions. However, it focuses on reducing the residual error in building better performing tree subsequently making this model to be more accurate than the traditional decision tree.

Recent research by Vikranth Udandarao and Pratyush Gupta with the tittle "Movie Revenue Prediction Using Machine Learning Models," with a particular objective which is to develop a machine learning to predict the movie earning based on various features such as movie name, genre, release year, rating, director, writer, cast, production country, budget and production company. The model was built with several machine learning models including Linear Regression, Decision Trees, Random Forest and XGBoost Gradient Boosting Regression. The research findings show that Gradient Boosting achieved highest accuracy with 91.58% for training data and 82.42% for the testing data. The model was also enhanced using hyperparameter tuning and cross validation. The hyperparameter tuning used was GridSeachCV to maximize the XGBoost model performance and a 5-fold cross-validation method is used to generalize the data and avoid overfitting.

The thesis by Sagar Maan Shrestha and Aman Shakya, titled "A Customer Churn Prediction Model suing XGBoost for the Telecommunication Industry in Nepal," focuses highlighting the effectiveness of XGBoost in addressing class imbalances in churn predictions and providing a competitive edge in customer retention while utilizing two types of datasets with total amount of data nearly 56 000 records. The effectiveness of the XGBoost model can be seen by achieving accuracy of 97% and F1-score of 88% which outperformed other models such as Random Forest and traditional decision tree algorithms. The study further highlights the performance of XGBoost can be improved by using more quality data and additional features.

2.2.3.4 Decision Tree Prediction

Decision Tree is a supervised learning algorithm that can be used for regression analysis where it structures a tree with subsets where the node in the subsets represents the features and continue to expand until the outcome of the branch is represented in each leaf node. The straightforward nature of this model has the capability to handle numerical and categorical data however the model cannot perform complex task. The final output is based on summation of each leaf node and the number of leaf nodes produced.

The research paper titled "Profit Driven Decision Trees for Churn Prediction" by Sebastiaan Höppner, Eugen Stripling, Bart Baesens, Seppe vanden Broucke and Tim Verdonck, published in European Journal of Operational Research, 2020, explores on the development of ProfTree, a profit-driven decision tree model for customer churn prediction shows that decision tree model outperforms classifier in predicting the customer churn. However, the decision tree model used was a model that was specifically designed for churn prediction that emphasis on profitability while other models emphasize on accuracy or statistical significance.

Other than that, in a paper titled " Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression" by Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar and Prof. V. A. Kanade, the authors explore the application of decision trees, logistic regression, and neural networks in making effective churn predictions to improve revenue management while utilizing data mining techniques to identify patterns in historical data. They found out that decision tree models specifically C5.0 and CART outperformed other models with clear decision rules and handle both numerical and categorical data.

2.3 Methodology

The Waterfall model is one of the earliest and most popular software development life cycle models, characterized by its linear-sequential approach. Despite its popularity, in this project that involves predictive modeling and data analysis, the traditional Waterfall model limits the ability to make necessary changes after each phase it completed. To address these limitations, the iterative waterfall model is used where this model feedback paths to the previous phases allowing for improvements based on findings and outcomes such as the prediction model performance, model validation or even refining the data preparation.

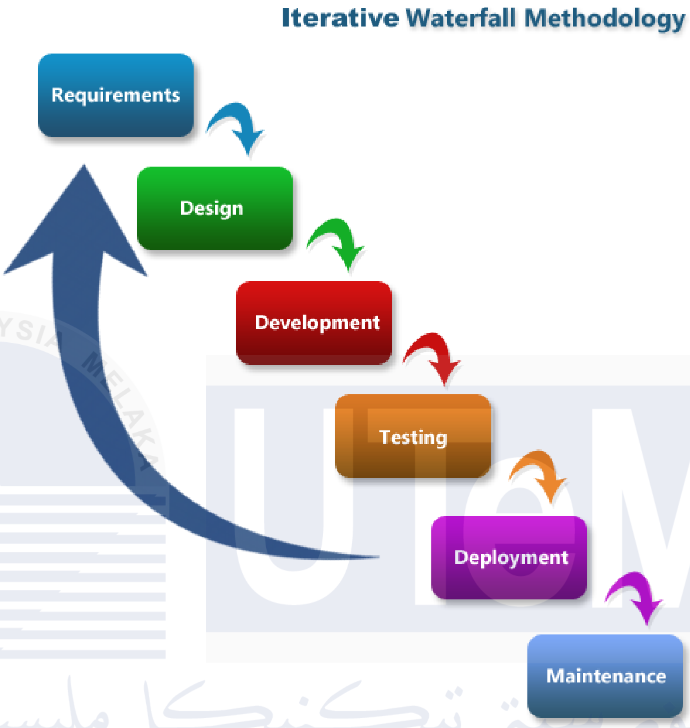


Figure 2.5: Iterative Waterfall methodology

2.3.1 Requirement Analysis

In the requirement analysis phase, all the requirement for this project was captured. This process involves studies on existing systems, understanding the respective research papers techniques and finding, a discussion with stakeholder on the respective expected output of this project and studies on building interactive analytics dashboard using Power BI. The identified software requirement for this project is Power BI, SharePoint and Jupyter Notebook while the hardware utilized for this project was 13th Gen Intel® Core™ i5-1335U with 12.0 GB.

2.3.2 System Design

In the system design phase, the design of the system was prepared based on the requirement analysis made in the previous phase. This includes the development of system architecture, dashboard design and algorithm design. Additionally, the design phase involved detailed sketching and discussions to ensure all aspects of the system were thoroughly considered and aligned with the project's goals.

2.3.3 Development

In the development phase, coding writing and configuring the selected software and hardware was carried to build the designed system. The development phase for this project started by analyzing the Excel dataset, followed by cleaning the data and implementing Random Forest Regression, XGBoost, Decision Tree for prediction purpose while using ARIMA model as the forecasting model. This phase includes the development of Power BI dashboard using historical data.

2.3.4 Testing

After development of the machine learning prediction and forecasting model, testing was conducted on the models using error metrics to evaluate the model's performance. The evaluation included calculating the Mean Squared Error (MSE) and the R-squared (R^2) values for both training and testing datasets. Other than the machine learning prediction and forecasting testing, the dashboard functionality and usability test was conducted to ensure the dashboard is free of error and met the stakeholder expectations. This phase was crucial in validating the accuracy and reliability of the system.

2.3.5 Deployment

In the deployment phase, the best prediction machine learning model and the forecasting model was deployed in the dashboard. This phase ensures that the dashboard is equipped with all the prediction, forecasting and analysis of the historical data with filters that helps to drill the report in a smooth and full-scale operation. Then, this phase includes presenting the final product to the stakeholder and validation on the requested requirements and discussion on the dashboard features and improvements.

2.3.6 Maintenance

In the maintenance phase, the revenue analytics dashboard was continuously monitored and updated to ensure the functionality of the dashboard. The maintenance included refining the machine learning models to incorporate the new data thereby improving the predictive and forecasting accuracy over time.

2.4 Project Requirements

For the successful implementation and operation of the project there are certain software and hardware requirement.

2.4.1 Software Requirement

For this project, Python with libraries such as Scikit-learn, and additional tools is used to develop and deploy machine learning models, including ARIMA, Decision Tree, XGBoost, and Random Forest regression. Jupyter Notebook provides the development environment for coding and model iteration. SharePoint works as a database hub that connects to Power BI dashboard and provides real-time data. Power BI Desktop is utilized to create and present the revenue analytics dashboard, converting model outputs into actionable insights for stakeholders.

2.4.2 Hardware Requirement

The hardware required will be minimum 8GH RAM memory, high resolution display such as 1280x800 for better visualization, dual-core processor such as i5 or equivalent and Microsoft Windows 8/10/11.

2.5 Project Schedule and Milestones

For FYP 1, from week two to week five, the focus is to conduct intensive research and study on existing systems and AI techniques used for similar case study to get an overview and ideas on the project. At end of week 5, the goal was to gather detailed requirements and define the scope, including functional and non-functional requirements. Then the process continues with analysis of the data and understanding the features in the given data and how impactful are they for the project. At the same time, the system architecture was designed to have an overview on the system flow, dashboard wireframe designs and flow of machine learning models that will be used. The development phase, from weeks eleven to thirteen, involved building machine learning model and building analytics dashboard. The final one week was used to test the performance of the machine learning model and get feedback to improve the system.

FYP 2 continued with focusing on hyperparameters tuning for the prediction algorithms to improve the accuracy of predictions and dashboard development. The timeline for FYP 2, from week one to week seven, emphasizes the structure of and refining the prediction algorithms and preparing the product which is the dashboard. From week one to week five, the goal was to optimize the prediction algorithms, followed by extensive testing and fine-tuning. Next, the focus was on developing the dashboard and testing the functionality of the dashboard. The final phase involves preparing the final report and presenting the completed system.

Table 2.2: Gantt Chart for FYP 1

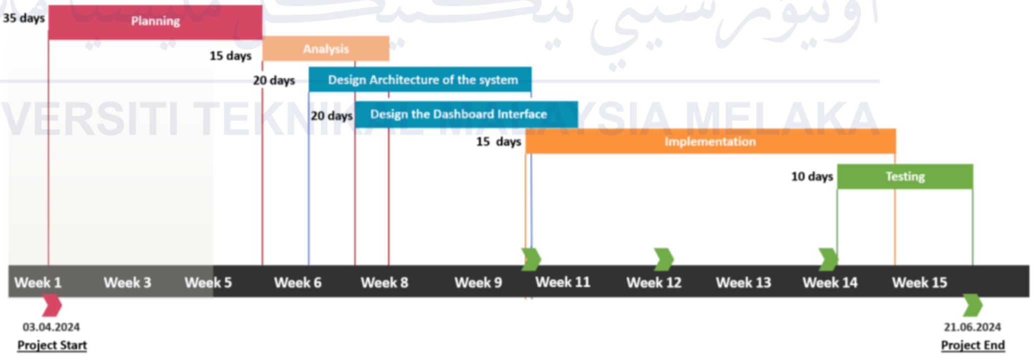
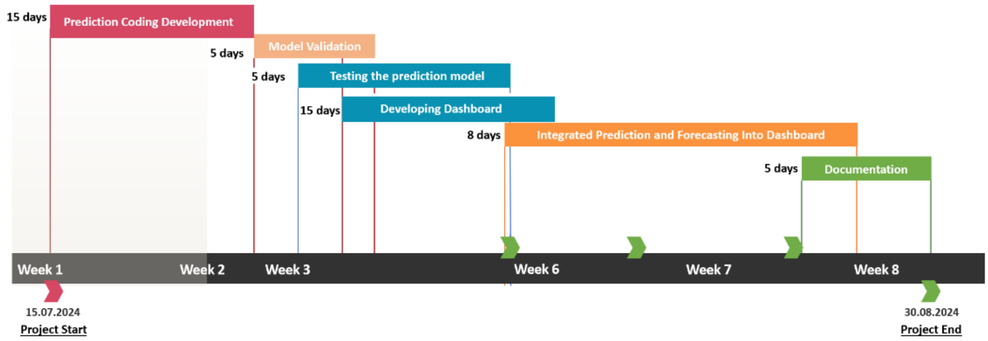


Table 2.3: Gantt Chart for FYP 2



2.6 Summary

In conclusion, this chapter covers the domain of the project and includes a comprehensive overview of the existing systems and related AI techniques. The iterative waterfall model was structured in detail according to the project schedule and milestones to ensure a structured process in achieving the project objectives.

CHAPTER 3: ANALYSIS

3.1 Introduction

In this chapter, detailed problem analysis will be conducted to refine the problem statement and requirements needed to address the financial management practices at Kelana Jaya Medical Centre. By thoroughly analyzing the problem and defining the requirements, this chapter aims to plan the structured methodology to address the problem.

3.2 Problem Analysis

The financial management practices at Kelana Jaya Medical Centre currently rely heavily on basic software tools to generate financial performance reports and forecasting. Hence, this section will provide a detailed analysis that highlights their limitations and the potential benefit of adopting new reliable technologies into their financial management practices.

3.2.1 Current Practices and Limitations

Firstly, the Kelana Jaya Medical Centre utilize Microsoft Excel to perform calculations and forecast revenue using statistical methods. The data for this calculation are collected from various departments within the hospital where the financial team performs percentage contributions of several features in the patient listing data with the targeted revenue to generate the forecast revenue. These practices comes with uncertainty of the forecast revenue as it does not take the seasonality, trends, and other external factors into account that might impact the future revenue. As a result, the forecasts may be oversimplified and not accurate.

Besides, the financial performance and targeted data are compiled into static reports software using Microsoft PowerPoint. These PowerPoint presentations are used during monthly review meetings. PowerPoint points out lack of visual and data analysis that is needed in order to have an interactive and drill down report to have an overview on the hospital’s financial performance and any deviations from the projected figures.

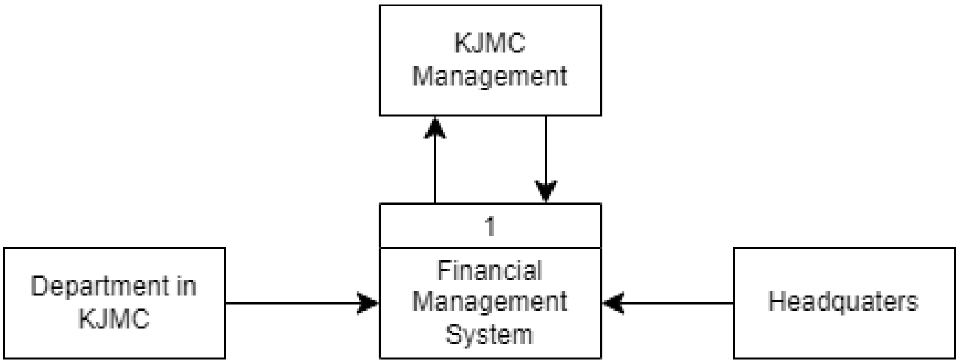


Figure 3.1: Context diagram of the current system

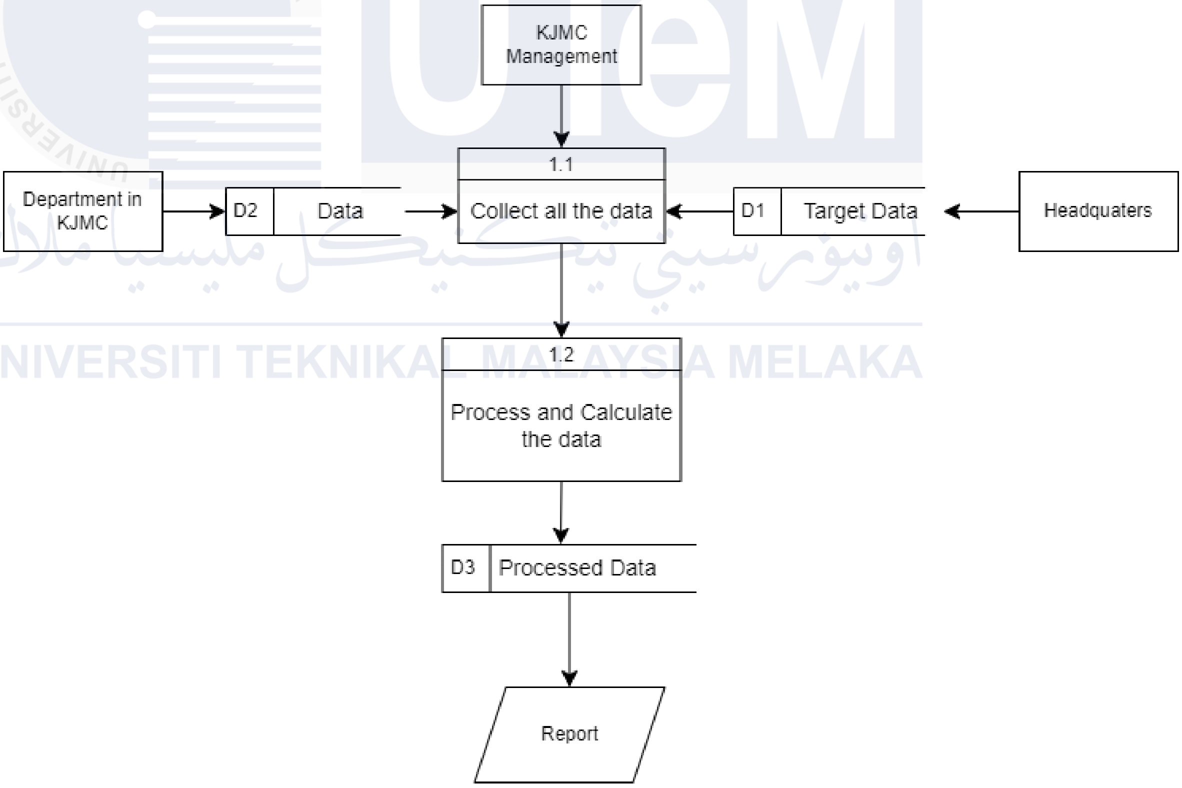


Figure 3.2: Level 1 data flow diagram of the current system

3.2.2 Detailed Methodology Using Iterative Waterfall Model

In an effort to enhance the financial management and improve the financial strategic planning capabilities of Kelana Jaya Medical Centre, the machine learning techniques for prediction and forecasting the revenue while developing an interactive analytics dashboard is very crucial. The implementation of these solutions will be guided by the iterative waterfall model to ensures a structured yet agile approach.

The first phase starts by gathering and analyzing all necessary requirements to build the new system. This begins with conducting comprehensive meetings with the hospital financial administrator to understand the features in the data and specific needs and expectations for the system. During these discussions, the understanding of limitations of the current system and practices was identified, and the expectation of the system was assured. We determine the sources and formats of the data and how often this data needs to be collected and updated. Additionally, hardware and software requirements are defined to support the machine learning models and analytics dashboard. Detailed user requirements for the dashboard, including desired visualizations and filtering options, are also collected to ensure the final product meets user needs.

In the system design phase, the architecture of the system is meticulously planned, covering data flow, machine learning model integration, and dashboard layout. The data architecture is designed to outline how data will be collected, stored, and processed. The selection and design of machine learning models involve choosing ARIMA for linear and short-term trends for forecasting the revenue while Random Forest Regression, Decision Tree and XGBoost are choose to predict the revenue based on several features, along with designing their integration. Wireframes and mockups of the interactive dashboard are created to specify the layout and functionality of visualizations. A comprehensive system integration plan is also developed, detailing how the data collection system, machine learning models, and dashboard will interact.

Next, in the development phase, coding, configuration of selected software and hardware, and building the designed software were key activities. This phase began with the analysis and cleaning of the Excel dataset, followed by the implementation of

ARIMA for forecasting, along with Random Forest Regression, Decision Tree, and XGBoost for revenue prediction. Concurrently, a Power BI dashboard was developed using this historical data to offer interactive and real-time visualizations based on the modify wireframe and mockups made previously.

After developing the machine learning prediction models and dashboard, extensive testing is conducted to ensure accuracy and usability. The performance of the models were tested using error metrics to ensure that the predictions and forecasting are reliable. Usability testing will be conducted to ensure that the dashboard is functioning well, user-friendly and meets all the stakeholder needs and the dashboard are free from significant errors.

In the deployment phase, the best performance machine learning prediction model and ARIMA time series machine learning model were integrated into the analytics dashboard. In this phase, the dashboard was fully equipped with historical data analysis features, including filters to facilitate the detailed report, prediction data and forecasting data. This dashboard then is presented to the stakeholder to ensure it provides all the necessary information is visualize and discuss on the potential improvements to be carried out.

The final phase involves maintaining the system, which involves continuous monitoring and updating the revenue analytics and forecasting dashboard to ensure the ongoing functionality. This phase including refining the machine learning with more data to enhance accuracy of predictions and forecast over time.

3.3 Requirement Analysis

3.3.1 Data Requirement

The first data provided is patient listing data consist of unique IDs, doctor’s information, admission and discharge date, location that refers the department in the hospital, payment plan of the patients, branch of the hospital, patient admission type and amount paid for past several years. The details of the patient listing data was simplified by selecting columns of features that are required in order to build the analytics dashboard that meets the stakeholder expectations. With that, the patient listing data consist of unique ID, admission and discharge date, location, patient admission type and amount paid shown in Table 3.1. Next, the data given is targeted revenue for every location, patient admission type and targeted patient visit for both admission type shown in Table 3.2. Additionally, two general information data was collected which are location data consisting of location and respective discipline as shown in Table 3.3 and doctor's data consisting of doctor's name and the location they are working as shown in Table 3.4. These data are stored in several excels files in a SharePoint.

Table 3.1: Data Dictionary for Patient Listing Data

Columns	Data Type	Description
Episode No	Whole Number	UniqueID for each patient
Admission Date	Date	The patient admission date
Discharge Date	Date	The patient discharge date
Location	String	The department the patients visited
Doctor	String	The doctor that attended the patient
Adm Type	String	Type of patient either Inpatient (IP) or Outpatient (OP)
Pat Amount	Decimal	Amount paid by the patient
Payor Amount	Decimal	Amount paid by third party such as insurance company

		or government health program
Amount	Decimal	The sum of bill from the Pat Amount and Payor Amount for a patient.

Table 3.2: Data Dictionary for Targeted Revenue Data

Columns	Data Type	Description
Month-Year	Date	The month and year for the targeted revenue
Location	String	The disciplines in the hospital
Adm Type	String	Type of patient either Inpatient (IP) or Outpatient (OP)
Count	Whole Number	Targeted number of patients based on Adm Type and Location.
Amount	Decimal	The targeted revenue based on the month and year.

Table 3.3: Data Dictionary for Location Data

Columns	Data Type	Description
Location	String	The location name which is unique
Disciplines	String	The disciplines comes under the location

Table 3.4: Data Dictionary for Doctor Data

Columns	Data Type	Description
Doctor	String	The doctor's name which is unique
Location	String	The location the doctor is working based on their disciplines

3.3.2 Functional Requirement

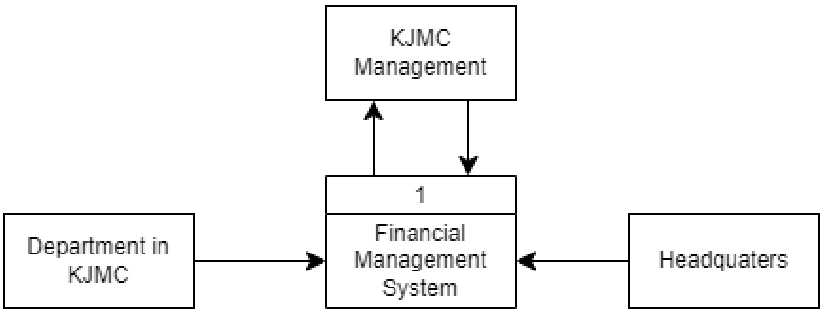


Figure 3.3: Context diagram of the system

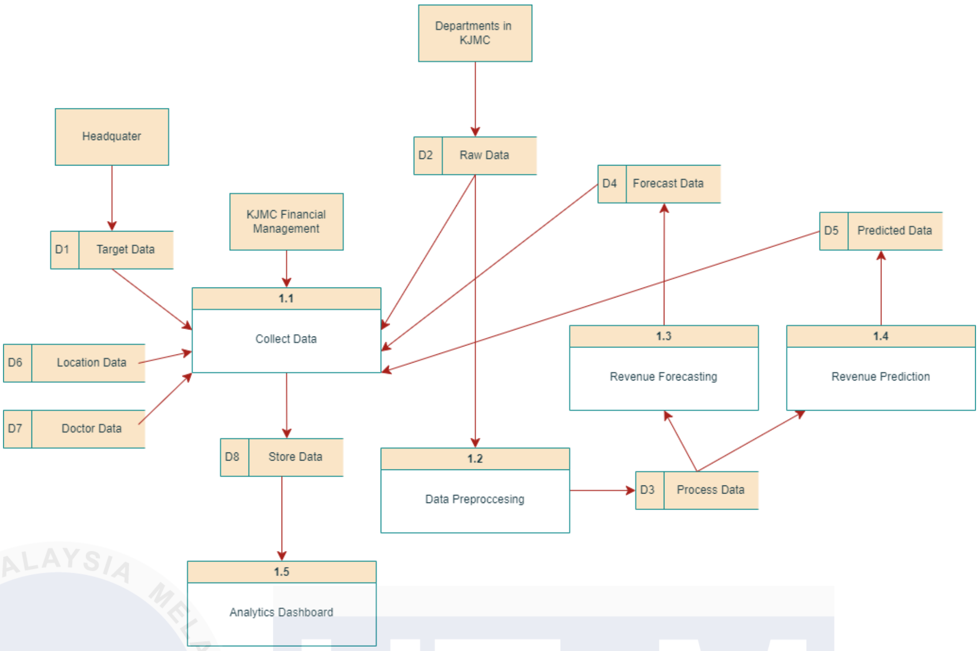


Figure 3.4: Level 1 data flow diagram of the system

The process starts with the collection of targeted data from the headquarter and patient raw data from different departments within KJMC. These inputs are labelled as D2 (Data from departments) and D1 (Target Data from headquarters).

Once the data is collected, the next step is data preprocessing which are cleaning and organizing the raw data to ensure it is fit for analysis. Preprocessing tasks includes handling missing values and standardizing data formats. Other than that, this process involves selecting the features needed for analysis to make the data looks more organized and easier to understand. The pre-processed data (denoted as D4) is then ready for analysis. This step is crucial to ensure the data quality.

The main function of the system is to predict and forecast the revenue using machine learning models like Random Forest Regression, Decision Tree, XGBoost and ARIMA models. These models are train with the pre-processed data. The output of the predicted revenue is denoted as D5, which provides insights into features roles in predicted the revenue while D4 the forecasted data provides insights into the expected financial performance.

The final step involves creating an analytics dashboard. All the data from departments, headquarter, doctors' data, location data, predicted and forecasted data are collected in the store data denoted as D8 to build the analytics dashboard. This dashboard includes interactive elements such as line charts, bar charts, and pie charts, with filtering options by month and year.

3.3.3 Non-Functional Requirement

The prediction model should provide less error metric values which should be less than 5% to ensure the reliability of the predictions and aids in effective financial planning. Besides, the analytics dashboard should be user-friendly and consist all the filter requirement needed to provide comprehensive report, allowing stakeholders to easily navigate and interpret the data. It should include clear visualizations and intuitive filtering options. The system can be accessed 24 hours a day.

3.3.4 Other Requirement

The system will utilize several key software tools to ensure efficient data handling, analysis, and visualization. Microsoft Excel will be employed for data collection, cleaning, and preliminary analysis. Its widespread use in the healthcare industry and familiarity among staff make it an ideal choice for organizing all the data at the same platform, which is SharePoint. Python libraries such as Pandas, NumPy, Scikit-learn and Statsmodels will be used. Power BI will be the primary tool for creating analytics dashboards.

The system requires a computer processor for data entry and analysis with at least i5 processor, 8 GB of RAM, and 256 GB of SSD storage to ensure efficient performance since the volume of the data collected is high. A highly secure and speed network infrastructure is needed to support data transfer between departments and store all the data in one database hub so that Power BI will be able to retrieve and visual the data without any delay while maintaining the data integrity and security.

3.4 Summary

This chapter outlines the current limitations of the Kelana Jaya Medical Centre and provide a structured methodology for developing a system to overcome the limitations. Key steps include gathering the required data, identifying the features in the data, visualizing the overview of the new system that will be implemented, functional requirements and nonfunctional requirements to ensure the quality of the system. Besides, in this chapter, the requirement of the hardware and software for this system is explained which is vital in developing the system to ensure the system able to run smoothly without and disruption.

CHAPTER 4: DESIGN

4.1 Introduction

In this chapter, the machine learning models used such as the ARIMA model, Random Forest Regression, Decision Tree and XGBoost model for prediction and forecasting revenue will be explained step by step on how the model works with the data and produce output. Other than that, the chapter will provide abstract on the high-level design, including system architecture, user interface design, and database design.

4.2 High-Level Design

4.2.1 System Architecture

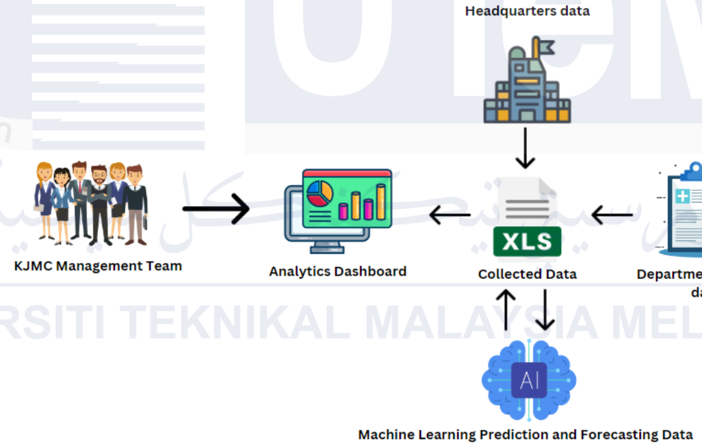


Figure 4.1: System Flow and Architecture

The system flow and architecture of the solution system at Kelana Jaya Medical Centre is designed as shown in Figure 4.1. The process start with collecting data from various sources, performing analysis on the data, and visualizing the data with the integration of all the data into the analytics dashboard. The data will be stored in Excel (XLS) format. Using the data fed and setting of visualizations in the Power BI dashboard, the dashboard will provide various visualizations such as line charts, bar charts, pie charts and tables to display the analysis of the data. The filter tool to drill the data by certain criteria such as the date hierarchy allows to provide more detailed information on the data analysis. The KJMC Management Team will interact with the system primarily through the analytics dashboard to interpret the financial performance.

4.2.2 User Interface Design

The user interface design which is the dashboard is designed to be intuitive and user-friendly to facilitate the interaction for the KJMC Management Team. The dashboard serves as the main interface where the management can access and interact with the financial data developed using Power BI. The dashboard consists of four pages. Figure 4.2 illustrates the first page in the dashboard which is the revenue overview. In this page, the management will be able to see the monthly basis performance based on the filtered year and month where is consist of total revenue, revenue by department and patient type, count of patient type versus the targeted revenue for each department and patient type as well as count of targeted inpatient and outpatient count. Figure 4.3 illustrates the second page of the dashboard which is the annual report page where on this page the management will be able to filter by year and get an insight of the yearly financial performance of each department and patient type. Next, the figure 4.4 shows the third page in the dashboard that is responsible to show the doctor's performance by providing insights on the doctor's name, department, revenue, total inpatient and outpatient attended and the revenue generated for each type of patient type. The fourth page show in figure 4.5 shows the machine learning model prediction and forecasting page.

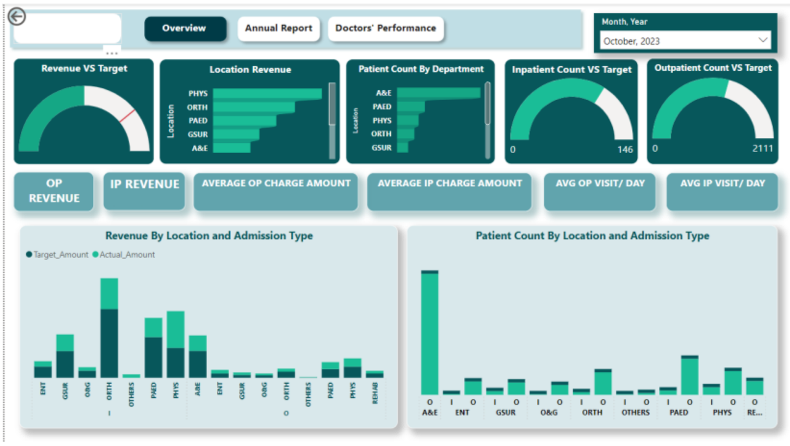


Figure 4.2: GUI above shows the Revenue Overview page in the dashboard

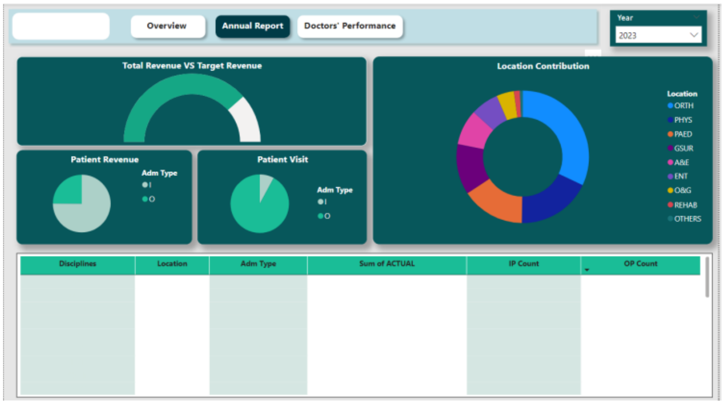


Figure 4.3: GUI above shows the Annual Report page in the dashboard

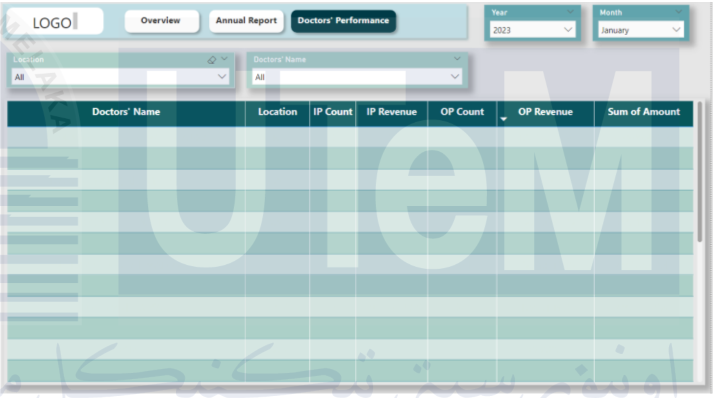


Figure 4.4: GUI above shows the Doctor’s Performance page in the dashboard

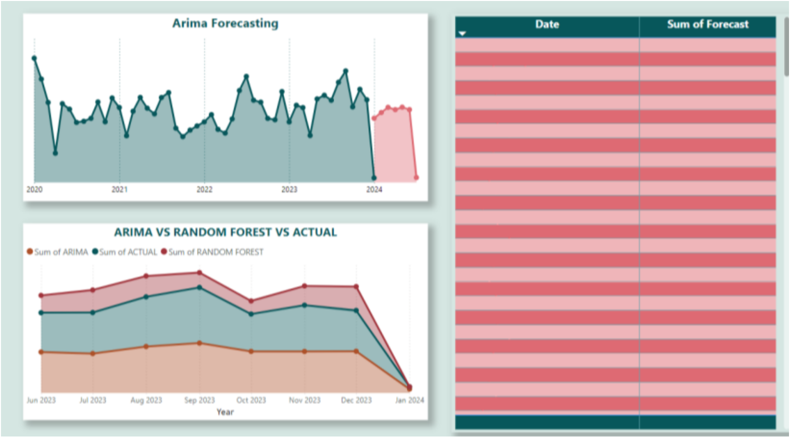


Figure 4.5: GUI above shows the Machine Learning Models in the dashboard

4.2.3 Database Design

Entity relationship diagram (ERD) is used to represent how a table of data related to another table of data. The business rule is that one doctor will have only one location, but one location can have one or many doctors. One doctor can have one or many patients, but one patient can have one doctor only. One location can have only one targeted revenue, but targeted revenue can have for one or more location. One patient can have one location only, but one location can have one or many patients. The predicted data and forecast data stands alone since it is extracted data from patient listing to perform machine learning prediction and forecasting outside the Power BI tool. All relationships are weak relationships.

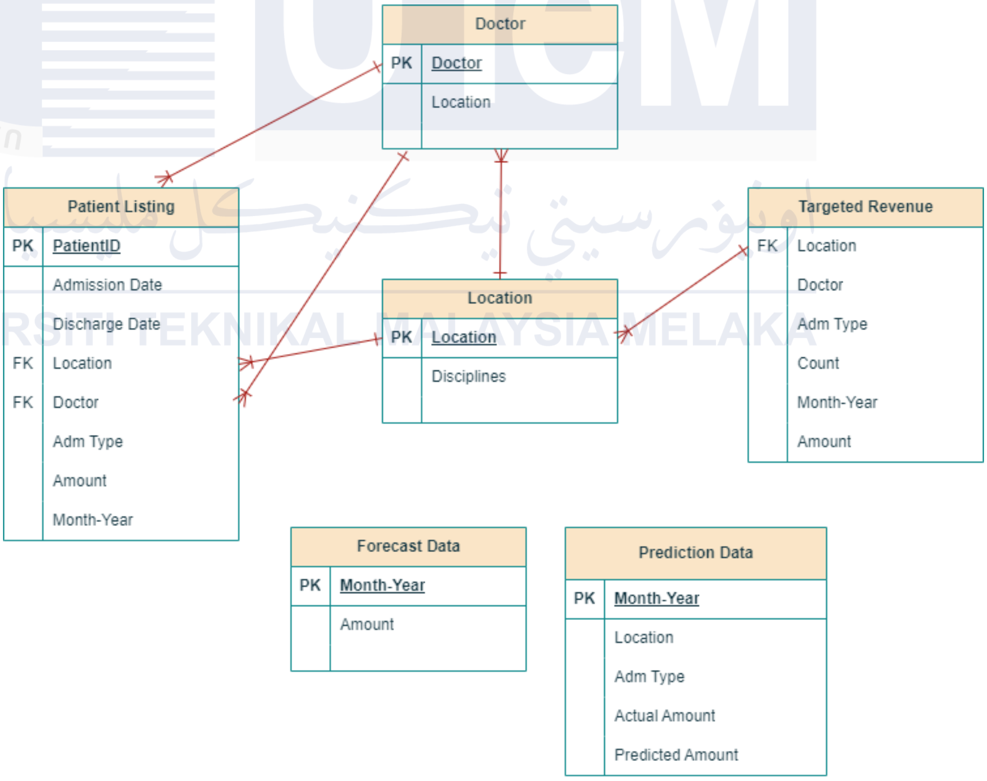


Figure 4.6: ERD diagram of the excel data

Table 4.1: Data Dictionary for Patient Listing

Column Name	Data Type	Constraint	Description
PatientID	Whole number	Primary Key	The patient unique ID
Admission Date	Date	Not Null	The patient admission date
Discharge Date	Date	Not Null	The patient discharge date
Doctor	String	Foreign Key	The doctor's name who attended the patient
Location	String	Foreign Key	The departments in the hospital
Adm Type	String	Not Null	The patient admission type
Amount	Double	Not Null	The amount paid by the patient

Table 4.2 Data Dictionary for Target Revenue Data

Column Name	Data Type	Constraint	Description

Month-Year	Date	Not Null	The month and year in date format
Adm Type	String	Not Null	The patient admission type
Location	String	Foreign Key	The departments in the hospital
Amount	Double	Not Null	The targeted revenue
Count	Decimal	Not Null	The targeted count of patient visit

Table 4.3 Data Dictionary for Prediction Revenue Data

Column Name	**Data Type**	**Constraint**	**Description**
Month-Year	Date	Not Null	The month and year in date format
Adm Type	String	Not Null	The patient admission type
Location	String	Foreign Key	The departments in the hospital
Amount	Double	Not Null	The predicted revenue

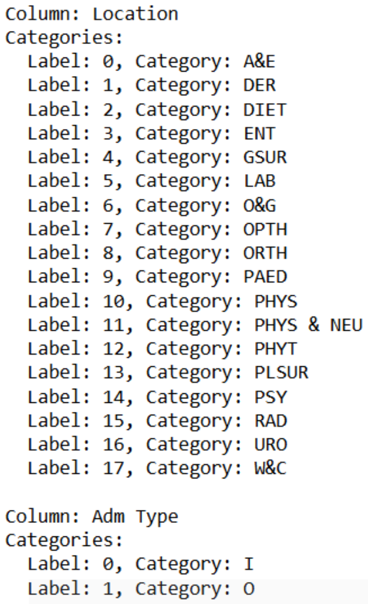
Table 4.4 Data Dictionary for Forecast Revenue

Column Name	Data Type	Constraint	Description
Month-Year	Date	Not Null	The month and year in date format
Amount	Double	Not Null	The forecast revenue

4.3 AI Component Design

4.3.1 Dataset Description

The dataset consist of the patients records from the year 2020 to 2023 which includes the columns Admission Date, Discharge Date, Location, Admission Type and Amount. The Admission Date is excluded from the analysis as the length of patient stay does not count as a feature that contributes to revenue. The data is aggregated by summing the Amount for each unique combination of Location and Admission Type based on the Discharge Date. Categorical data such as the Location and Admission Type are encoded using LabelEncoder like shown in the Figure 4.7 to convert them into unique numerical values to ease the machine learning model prediction process. Then the Discharge Date is converted into datetime format and new features such as Year, Month and Day are extracted. The Amount column is normalized using MinMaxScaler to standardize the amount range between 0 to 1. Then finally, the dataset is divided into training and testing sets where 80% of the data is used as the training data.



```
Column: Location
Categories:
  Label: 0, Category: A&E
  Label: 1, Category: DER
  Label: 2, Category: DIET
  Label: 3, Category: ENT
  Label: 4, Category: GSUR
  Label: 5, Category: LAB
  Label: 6, Category: O&G
  Label: 7, Category: OPTH
  Label: 8, Category: ORTH
  Label: 9, Category: PAED
  Label: 10, Category: PHYS
  Label: 11, Category: PHYS & NEU
  Label: 12, Category: PHYT
  Label: 13, Category: PLSUR
  Label: 14, Category: PSY
  Label: 15, Category: RAD
  Label: 16, Category: URO
  Label: 17, Category: W&C

Column: Adm Type
Categories:
  Label: 0, Category: I
  Label: 1, Category: O
```

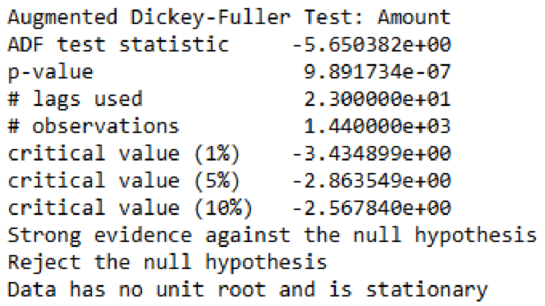
Figure 4.7 shows Encoded Categorical Data

4.3.2 Proposed Techniques

4.3.2.1 Revenue Time Series Forecasting Model

4.3.2.1.1 ARIMA Model

For this project, the ARIMA model is used to perform the time series forecasting. The ARIMA model was chosen for its ability to work with seasonal and non-seasonal data. The process begins with using processed data from patient listing data which includes daily amount from January 2020. The first step to perform the ARIMA model is by checking the data to identify whether the data is stationary or not. Stationary data are data that does not have predictable patterns in the long term. It is checked using Augmented Dickey-Fuller (ADF). ADF conducts a hypothesis test on targeted data which is the Amount and if the p-value is below threshold (0.5), hence the model suggests that there is no unit root (stable mean of time series without by eliminating the trend/seasonality) and the is stationary. ADF test is a modified version of the old Dickey-Full (DF) test which is more efficient in handling larger and complex data. Based on the ADF performance, the test statistic and p value shows a negative number indicating strong evidence against the null hypothesis that the data is stationary.



```
Augmented Dickey-Fuller Test: Amount
ADF test statistic      -5.650382e+00
p-value                  9.891734e-07
# lags used              2.300000e+01
# observations           1.440000e+03
critical value (1%)     -3.434899e+00
critical value (5%)     -2.863549e+00
critical value (10%)    -2.567840e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Figure 4.8 shows the ADF test value for the process data.

Once the stationary test is done, a decomposition of the targeted data will be plot into four figures which are Observed (Original data), Trend (Long Term Movement), Seasonal (Repetitive Pattern) and Residual (Noise of the data after the trend and seasonal components are removed). The purpose of this visualization is to identify the seasonality and trend of the data.

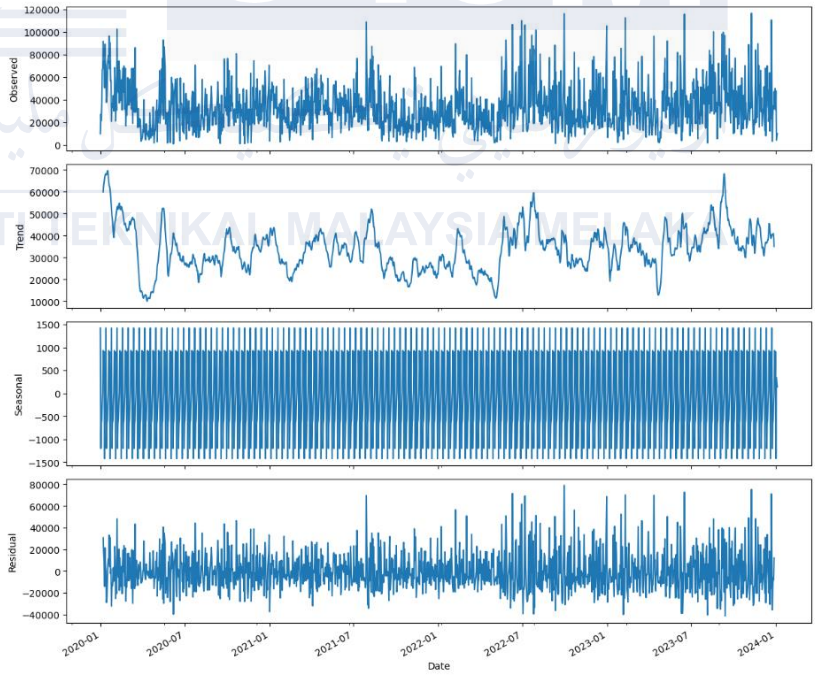


Figure 4.9 shows the decomposition of the targeted variable in the data

From the decomposition above, it can be clearly seen that the trend of the data is unpredictable over the time period, there is sudden low and high fluctuations in the graph with no clear linear direction and the seasonal components captures repetitive short-term patterns in the data. Most likely as the date cycle is repetitive with the date, month and year. The residual error or noise in the data resembles that these data are resembling white noise in which these leftover data does not shows and pattern or decomposition structures.

Then, the next step will be identifying the parameters involved in building the ARIMA model. The ARIMA model consists of three parameters which are the Autoregression (p), Integration/Differentiation (d) and Moving Average (q). The p values represents the lag of the regression model that forecast the target variable using linear combination of the past values. The Partial Autocorrelation Function (PACF) is used to plot the lags in the data.

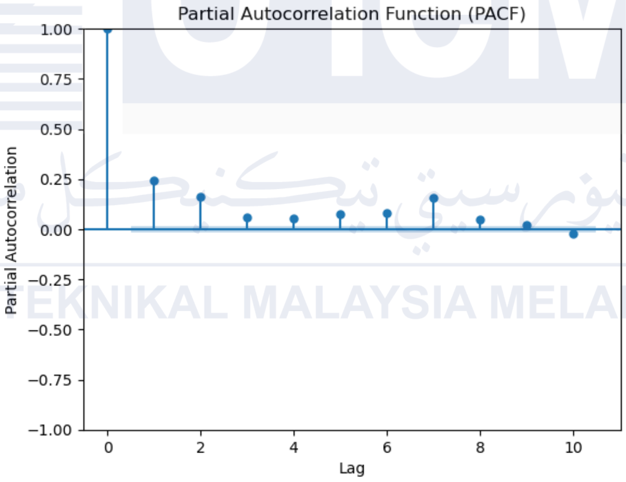


Figure 4.10 shows the PACF plot graph

Based on figure 4.10, the most significant lags is Lag 1 hence the p value is set to be 1. Next, the d value represent the approach of making the data stationary. Since the data is stationary hence the d value is 0. The q values works similarly like the p value however instead of past forecast value lag, it uses past forecast error lag using regression model as well. This can be seen using Autocorrelation Function (ACF) plot in figure 4.10 where the spike that shows the error lag nearest the linear regression plot will be spike 2 and spike 3. By considering the significant spike that is further than the linear regression line, the q value is considered as 2.

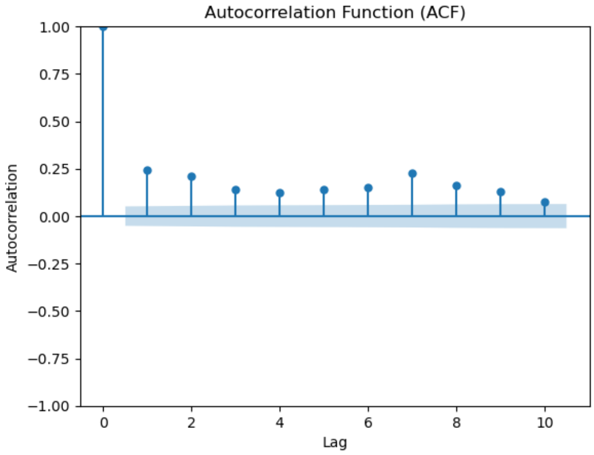


Figure 4.11 shows the ACF plot

Once all the parameters are identified, the model are fitted to the parameters (p, d, q) which are equal to (1, 0, 2) and perform prediction on the data. The performance of this data is evaluated using the Mean Squared Error value. Then, the model continues with forecasting the following 6 months data from the given data. This forecast revenue by the model will be saved as forecast data into the database system.

4.3.2.2 Revenue Prediction Machine Learning Model

4.3.2.2.1 Random Forest Regression Model

For this project, the Random Forest Regression model is used for predicting the revenue using several features which are Discharge Date, Admission Type, Location and Amount. The initial step starts with data preparation which includes processing the Discharge Date into Year, Month and Date, using encoded categorical data such as location and admission type as shown in Figure 4.7 and column Amount. Then step two is that the features and target variable are defined. In this case, the target variable is Amount and others are considered as features (Year, Month, Day, Location, Adm Type). The target variable is then normalized using min max scaler to avoid bias by the highest amount to dominate the learning process overshadow the features with lower amount value. Then the data is fed to the Random Forest Regression library which is ‘RandomForestRegressor’ which works by first creating multiple subsets of tree of the training data through the bootstrapping process. For each subset, the model randomly selects based on the constructed decision trees. Then the final prediction is made by calculating average outputs from the best leaf node of each tree to ensure smooth and accurate prediction like shown in figure 4.12. Then, the model performance is accessed through metrics such as Mean Squared Error (MSE) and R-squared (R^2) score.

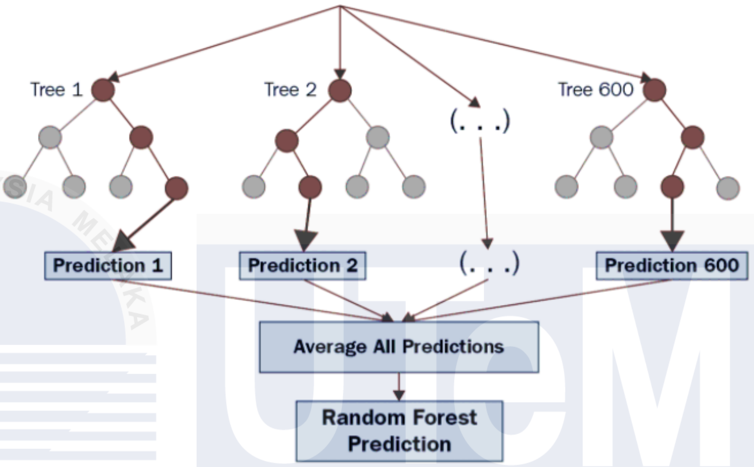
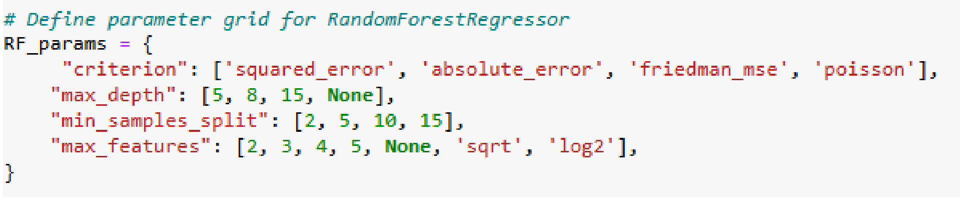


Figure 4.12 shows Architecture of Random Forest Regression Model

Then the model undergoes hyperparameter tuning, with tree-specify hyperparameters defined randomly. The parameters used are criterions which checks the quality of split in a tree, and the selected feature for this parameter is based on the scikit-learn API documentations. Next, the parameter used max_depth which defines the maximum depth of the tree using small scale like 5, bigger scale like 15 and None which will expand until the leaves are pure/ less than min_sample_split. Next, the min_samples_split represents the split point at any depth of the tree which can only be considered if the left branch of the tree at minimum sample in each left and right branch of the tree. These are also scale from smaller value, which is 2 ,5 ten bigger values such as 10, 15. Moving on to the next parameter is the number of features to consider when looking for the best split in the tree (max_features). Since the number of features in my data is 5, the numbers are from 2 features to 5 features with square root (sqrt) and log 2 of the total features.



```
# Define parameter grid for RandomForestRegressor
RF_params = {
     "criterion": ['squared_error', 'absolute_error', 'friedman_mse', 'poisson'],
    "max_depth": [5, 8, 15, None],
    "min_samples_split": [2, 5, 10, 15],
    "max_features": [2, 3, 4, 5, None, 'sqrt', 'log2'],
}
```

Figure 4.13 shows the defined parameters for RandomForestRegressor

Then, GridSeachCV is used to find the best and optimal parameters where it will perform exhaustive search over specified parameters values in grid (column x row). The parameters of the GridSearchCV is specifies with estimator which is the 'RandomRegressorModel' model, parameter grid referring to the predefined parameter earlier, scoring using neg_mean_squared_error which is similar to mean-squared errors (MSE) the advantage of using negative values are that the algorithm can minimize the MSE values. Next the n_jobs which refers to the maximizing the CPU resources is defined at -1 to ensure multiple cores. The verbose (detailed output information of the search) is 2 and cross validation, cv = 5 the length of my total features. Then, the model's performance is once again measured with error metrics such as Mean Squared Error (MSE) and R-squared (R^2) score. Then prediction is made best performance model on the test data and the predicted amount with its features are compiled in the Excel file. The best parameters for this model are criterion: 'friedman_mse', max_depth : None, max_features: log2 and 'min_samples_split: 15.

4.3.2.2.2 XGBoost Regression Model

For this project, the second revenue prediction model used is XGBoost Regression. The initial step starts with data preparation which includes processing the Discharge Date into Year, Month and Date, using encoded categorical data such as location and admission type as shown in Figure 4.7 and column Amount. Then step two is that the features and target variable are defined. In this case, the target variable is Amount and others are considered as features (Year, Month, Day, Location, Adm Type). The target variable is then normalized using min max scaler to avoid bias by the highest amount to dominate the learning process overshadow the features with lower amount value.

Then, the process begins by initializing the XGBoost Regressor in which the model constructs the decision tree in a way that the tress tries build the next tree which the purpose to minimize the error from the previous tree by using gradient of the loss function. Each tree makes prediction by a weighted sum of all the trees outputs like shown in Figure 4.14. Then the model make predictions with test and train data and the performance of both these data are measured using MSE and R^2 values.

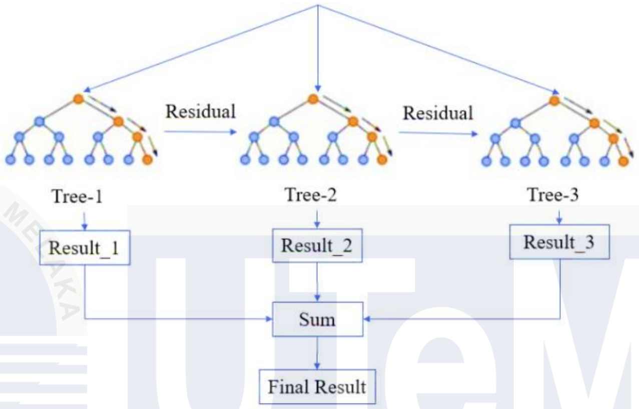
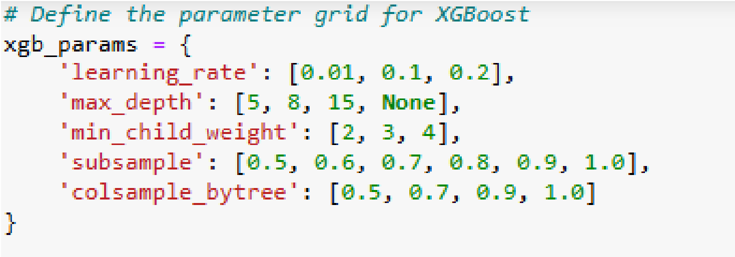


Figure 4.14 shows on the Architecture of the XGBoost Regression Model

Then the model undergoes hyperparameter tuning for tree-specify hyperparameters which are shown in figure 4.15. max_depth parameters values is used similar to the random forest regression model, and four more new parameters are used in this model which are min_child_weight which helps to control the complexity of the decision tree to ensure that the leaves are created not too small because of the nature of this model that produces tree with less error function values. Then the subsamples shows how many percent of the rows must be used for each tree construction which at least 0.5 to 1 represents all the rows in the data to be used. The final parameter is the colsample_bytree which represents the columns to be used which uses the same parameter values at subsamples. Then GridSearchCV is with the same parameter values for random forest is used to find the best paramaters for XGBoost Regression model. Then, the model's performance is once again measured with error metrics such as Mean Squared Error (MSE) and R-squared (R2) score. Then prediction is made best performance model on the test data and the predicted amount with its features are compiled in the Excel file. The best paramater for this model are colsample_bytree: 0.7, learning_rate: 0.1, max_depth: 5, min_child_weight: 3 and subsample: 0.7.



```
# Define the parameter grid for XGBoost
xgb_params = {
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [5, 8, 15, None],
    'min_child_weight': [2, 3, 4],
    'subsample': [0.5, 0.6, 0.7, 0.8, 0.9, 1.0],
    'colsample_bytree': [0.5, 0.7, 0.9, 1.0]
}
```

Figure 4.15 shows the defined parameters for XGBoost Regression Model

4.3.2.2.3 Decision Tree Model

For this project, the third prediction model used is Decision Tree model. The initial step starts with data preparation which includes processing the Discharge Date into Year, Month and Date, using encoded categorical data such as location and admission type as shown in Figure 4.7 and column Amount. Then step two is that the features and target variable are defined. In this case, the target variable is Amount and others are considered as features (Year, Month, Day, Location, Adm Type). The target variable is then normalized using min max scaler to avoid bias by the highest amount to dominate the learning process overshadow the features with lower amount value. The decision tree simply works by expanding trees with the features until it reaches the leaf nodes that cannot be expanded more. The final output of this model is the summation of the targeted variable divided by the total number leaf node produced by the tree. Then, the process begins by initializing the model Decision Tree Regressor with a random state of split 42. This value is the most common seed value. Then the model makes predictions on the test and train data and the performance are evaluated using the MSE and R^2 values.

The model undergoes the same parameters that focuses on the tree-specify hyperparameters tuning as the random forest regression model and uses the similar GridSeachCV parameters to evaluate and select the best parameters that works with the decision tree. Then, the model's performance is once again measured with error metrics such as Mean Squared Error (MSE) and R-squared (R2) score. Then prediction is made best performance model on the test data and the predicted amount with its features are compiled in the Excel file. The best parameters for this model are criterio': squared_error, max_depth: 5, max_features: 5 and min_samples_split: 15.

4.4 Summary

In this chapter, the architecture of the entire system is designed which includes designing the user interface which is the dashboard which consist of three pages for revenue overview, annual reports, and doctor performance. Other than that, the database design is also included in the chapter using Entity Relationship Diagram (ERD) and data dictionaries for all the different Excel files used. The AI techniques used was designed starting from the data preparation, modeling, fitting the model, hyperparameters tuning of the model, performing grid search to find the bets parameters and select the best parameters to train the model and evaluate the performance. The ARIMA model structure was also explain in the detail in this chapter on the way of selecting the parameters and how the model is fitted to the data.

CHAPTER 5: RESULTS AND DISCUSSION

5.1 Introduction

In this chapter, the testing phase for the machine learning models used and the analytics dashboard will be evaluated to ensure accuracy, reliability and usability. The testing phase for the machine learning models are evaluated using test data, using error metrics which are Mean Squared Error (MSE) and R-squared (R^2) while the analytics dashboard will be evaluated using test case.

5.2 Evaluation of AI Techniques used in the project

Firstly, the performance of the machine learning forecast model, ARIMA was analyzed. The performance metrics used was Mean Squared Error (MSE) to measure the average squared difference between the actual and predicted values. The model was trained and tested with 1464 rows of data and the MSE value obtained is very high which is 366238073.34. This shows that the model is unable to predict the revenue despite having seasonality and trend. Figure 5.1 shows the actual amount and predicted amount of the ARIMA model over date and amount where the blue line shows the actual amount, and the red lines shows the predicted amount.

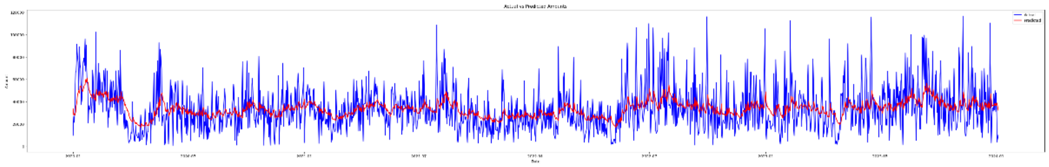


Figure 5.1 shows the Actual versus Predicted Amount of the ARIMA model

Next moving on to the performance of each predictive machine learning model with and without using the hyperparameters tuning with GridSearch. The performance metric used are Mean Square Error and R^2 or the coefficient of determination score is a statistical method used in regression model to see on the proportion of the variance in the dependent variable in which the higher the value the better the data is fit to the model. Table 5.1 below shows the MSE and R^2 of the three models without hyperparameter tuning on the test data and Table 5.2 show the MSE and R^2 for the three models with hyperparameter tuning on the test data. The test data is used to evaluate the accuracy to know on how well the model learn based on the data train and able to perform on unseen data.

Model	Decision Tree	Random Forest Regression	XGBoost Regression
MSE Error	0.00386	0.00219	0.00231
R^2 Score	0.13659	0.50780	0.48173

Table 5.1 shows the performance of the model without hyperparameter tuning

Model	Decision Tree	Random Forest Regression	XGBoost Regression
MSE Error	0.00215	0.00204	0.00180
R^2 Score	0.51640	0.54193	0.54131

Table 5.2 shows the performance of the model with hyperparameter tuning

The Random Forest Regression model has the lowest MSE of 0.00219 compared to other models before the hyperparameter tuning and highest R² score of 0.541934. The low MSE indicates that the model's predictions are close to the actual values. The R² value of approximately 0.542 means that the model can explain 50.7% of the variance in the amount variable with the other input features. This suggests that the model outperforms other models even without hyperparameter tuning. After the hyperparameter tuning occurs, the MSE value reduce by 0.0005 and the R^2 score increase the variance of the data. This indicates that with hyperparameter tuning enhance the performance of the model slightly.

Next, XGBoost Regression shows an MSE of 0.00321 and an R^2 scoreof 48.2% before the hyper parameter tuning. The MSE is very close to the Random Forest model, indicating likely accurate predictions. However, after the hyperparameter tuning, the MSE values is the lowest compared to other models the R^2 value is marginally increase by 6% , meaning XGBoost explains 54.1% of the variance in the target variable. While the difference is minimal with the Random Forest Regression, it suggests that XGBoost performs almost good like Random Forest but might slightly underperform in capturing the variability of the target variable.

The Decision Tree model records highest MSE of 0.00368 and lowest R^2 score of 13.7%. The higher MSE indicates that the model's predictions is less accurate compared toother two models. After applying hyperparameter tuning, the MSE values reduces however the model was able to explains 51.6% of the variance in the target variable, which is higher than the other two models but better than the previous score. This lower R^2 indicates that the Decision Tree model might be more prone to overfitting or underfitting.

Based on the result, it can be concluded that Random Forest has the best performance compared to other two models with the smaller MSE value and higher R^2 score value better followed by XGBoost Regression with a small difference in MSE and R^2 value and followed by Decision Tree. Hence, for this project Random Forest has the most accurate revenue prediction as the model is more effective in capturing the underlying patterns in the data.

5.3 Testing of Functional Requirement

Test case is conducted on the analytics dashboard to ensure the functionality and usability of the dashboard to ensure the accuracy of data, features like filtering and handles errors.

Table 5.2: Test Case for Analytics Dashboard

Test Case	Results from Dashboard	OK / Failed
Dashboard loads correctly within 5 seconds where all the components are initialized.	Dashboard loads within 5 seconds with all the widgets and charts are displayed correctly.	**OK**
Use data filtering by month, year, department, patient type and doctor's name.	Dashboard accurately filters and displays data based on the selected filter.	**OK**
Validate the data display for both revenue and patient visit count same as in the database.	The value shown in dashboard is similar to values of the data in database.	**OK**
Verify the forecasting data is displayed in the dashboard and the value is similar to the database.	Forecasting data is displayed and the value is similar to the database.	**OK**
Verify the prediction data is displayed in the dashboard and the value is similar to the database.	Predicted data is displayed and the value is similar to the database.	**OK**
Verify that the Doctor's Performance page correctly displays doctors' names, departments, revenue, inpatient/outpatient counts and the revenue by patient type.	All the data is accurately presented with clear distinctions between doctors and patient types.	**OK**

Clicking on a chart to see if the page updates to the relevant data.	Charts and graphs are responding and displaying the relevant data based on the click.	**OK**
User interface is consistent in all pages maintaining design, color scheme and layout.	All pages maintain consistent look with uniform buttons, fonts and navigation,	**OK**

5.4 Summary

This chapter covers the testing phase of the machine learning models and the analytics dashboard. The chapter begins by evaluating the machine learning prediction model performance using two error metrics which are MSE and R² to select the most accurate prediction model. Besides, the chapter also covers on the functionality test on the analytics dashboard ensuring the reliability, accuracy and effectiveness of the dashboard.

CHAPTER 6: CONCLUSION

6.1 Observation on Weaknesses and Strength

The strength of the project is demonstrated on the successful integration of machine learning models such as like Random Forest, Decision Tree, and XGBoost for accurate revenue prediction by choosing the best model performance using error metrics Besides the machine learning models, the strength of this project is the development of Power BI dashboard that provides insightful visualizations and filtering options that make the financial report accessible anytime and anywhere.

Weakness of the project is that the forecasting model does not provide a good performance in forecasting the revenue suggest that ARIMA model is not the best model to forecast the revenue for the used dataset. Consequently, this potentially limits the projects overall effectiveness in forecasting future revenue trends /patterns

6.2 Propositions for Improvement

The future improvement for the project will be increasing the historical data span to include more years especially data before the Covid 19 and after the Covid 19 which allows the forecasting and predictive machine learning model to capture wider range of time and complex patterns over time since there are changes in patient demographics, treatment patterns and healthcare practices. Other than that, the future improvement for this project will me exploring and implementing additional time series forecasting techniques such as LSTM, Prophet, SARIMA model and other deep learning models to provide more accurate forecasting to capture non-linear relationships and long-term dependencies.

6.3 Project Contribution

The project advances academic knowledge by demonstrating practical applications of machine learning in healthcare revenue forecasting. It provides a real-world case study that enriches the curriculum and offers a valuable reference for future research and student projects. The integration of various models such as ARIMA, Random Forest Regression, XGBoost, and Decision Tree enhances the understanding of their comparative performance and real-world applicability.

For Kelana Jaya Medical Centre, the project will be able to provide prediction system that supports data-driven decision-making by using analytics dashboard which helps to enhances financial planning and resource allocation, ultimately contributing to improved operational efficiency and strategic planning. The project provides an opportunity to experience in designing, developing, and implementing machine learning models and analytics dashboards using Power BI which indirectly provides knowledge and experience to handle complex datasets, way to optimize algorithms, and deliver user-friendly solutions. This practical experience is beneficial for professional development and positions the individual as a capable contributor in the field of data science and healthcare analytics.

6.4 Summary

In conclusion, the revenue forecasting and prediction system for Kelana Jaya Medical Centre did not meet the project's objectives in using machine learning forecasting model however implementing other machine learning models for accurate revenue predictions was successfully delivered together with the analytics database. This project highlights on expanding forecasting models beyond ARIMA to capture longer-term trends and complex patterns, to effectively performs its core functions. The project's design and implementation demonstrate significant potential for enhancing financial analysis and decision-making within healthcare settings especially with industry that uses conventional method. With further refinement and adaptation, this system holds promise for broader application and impact in the field of healthcare analytics.

REFERENCES

Nandi, B. K., Chaudhury, M., & Hasan, G. Q. (2014). Univariate time series forecasting: A study of monthly tax revenue of Bangladesh. East West Journal of

Česnavičius, M. (2020). Lithuanian electricity market price forecasting model based on univariate time series analysis. Energetika, 66(1), 39–46

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS International Journal of Geo-Information, 7(5), 168. https://doi.org/10.3390/ijgi7050168

Lomakin, N., Kulachinskaya, A., Naumova, S., Ibrahim, M., Fedorovskaya, E., & Lomakin, I. (2023). Modelling profits forecasts for the Russian banking sector using random forest and regression algorithms. Sustainable Development and Engineering Economics, 3. https://doi.org/10.48554/SDEE.2023.3.1

Udandarao, V., & Gupta, P. (2024). *Movie revenue prediction using machine learning models* [arXiv:2405.11651v1]. https://arxiv.org/abs/2405.11651

Dankorpho, P. (2024). Sales Forecasting for Retail Business using XGBoost Algorithm. *Journal of Computer Science and Technology Studies*, *6*(2), 136–141. https://doi.org/10.32996/jcsts.2024.6.2.15

Nahar, N., & Ara, F. (2018). Liver disease prediction by using different decision tree techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 8*(2), 1-14. https://doi.org/10.5121/ijdkp.2018.8201

Hossain, M. R., & Timmer, D. (2021). Machine learning model optimization with hyperparameter tuning approach. *Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence, 21*(2), 1-11. Global Journals. https://doi.org/10.34257/GJCSTD21.2.1

Weerts, H. J. P., Müller, A. C., & Vanschoren, J. (2020). *Importance of tuning hyperparameters of machine learning algorithms* [arXiv:2007.07588v1]. https://arxiv.org/abs/2007.07588

Choudhary, J. (2023). Mastering XGBoost: A technical guide for machine learning practitioners. Retrieved from https://medium.com/@jyotsna.a.choudhary/mastering-xgboost-a-technical-guide-for-intermediate-machine-learning-practitioners-f7ad167c6865#:~:text=Step%201%3A%20Initialize%20with%20a,approximation%20for%20the%20target%20variable.

Muhajir, D., Akbar, M., Bagaskara, A., & Vinarti, R. (2022). Improving classification algorithm on education dataset using hyperparameter tuning. *Procedia Computer Science, 197*, 538–544. https://doi.org/10.1016/j.procs.2021.12.023

Segal, M. R. (2003). *Machine learning benchmarks and random forest regression*. Division of Biostatistics, University of California, San Francisco, CA 94143-0560. https://doi.org/10.2139/ssrn.442342

Shrestha, S. M., & Shakya, A. (2023). A customer churn prediction model using XGBoost for the telecommunication industry in Nepal. *Procedia Computer Science, 215*, 652–661. https://doi.org/10.1016/j.procs.2022.12.067

Wang, C.-C., Kuo, P.-H., & Chen, G.-Y. (2022). Machine learning prediction of turning precision using optimized XGBoost model. *Applied Sciences, 12*(15), 7739. https://doi.org/10.3390/app12157739

Lyashenko, V., & Jha, A. (2024, April 12). *Cross-validation in machine learning: How to do it right*. Neptune Blog. https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right

Great Learning Team. (2024, April 30). Hyperparameter tuning with GridSearchCV. Great Learning. https://www.mygreatlearning.com/blog/gridsearchcv/

Gutierrez, D. (2018, September 14). Gradient boosting and XGBoost. ODSC. https://opendatascience.com/gradient-boosting-and-xgboost/

Botchkarev, A. (n.d.). *Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology*. GS Research & Consulting, Department of Computer Science, Ryerson University.

Taylor, J. W., de Menezes, L. M., & McSharry, P. E. (n.d.). *A comparison of univariate methods for forecasting electricity demand up to a day ahead*. Saïd Business School, University of Oxford; Cass Business School, City University; Department of Engineering, University of Oxford.