CHAPTER 1 : INTRODUCTION

1.1 Rationale

Indonesia, the fourth most populous country, is transitioning to a single-payer national health insurance system and is anticipated to become the largest in the world [1]. This healthcare insurance is administered by the Indonesia Health Insurance Agency/ *Badan Penyelenggara Jaminan Sosial* (*BPJS)*. BPJS plans to cover the Indonesian population, with one of its key initiatives being the National Health Insurance/ *Jaminan Kesehatan Nasional (JKN).* As of May 30, 2022, nearly 86% (240,315,474 people) of the Indonesian population participates BPJS Health insurance [2]. According to BPJS Health data, most diseases suffered by JKN-KIS participants in the age group of 44-94 years are cases of Diabetes Mellitus, with a prevalence rate of 88.51% of the total cases from 2015 to 2018, the number of Diabetes Mellitus patients continues to increase yearly. The International Diabetes Federation (IDF) reported that in 2021, the number of people affected by Diabetes Mellitus reached 537 million. It is predicted to increase to 643 million by 2030 and further to 783 million by 2045. The mortality impact of this disease is significant, causing 6.7 million deaths. It is estimated that one Diabetes Mellitus patient dies every 5 seconds [3]. This situation underscores the urgency of addressing this disease more efficiently and swiftly. However, an issue has arisen concerning the hospital patient density, attributed to the Length of Stay (LOS) [4]. Consequently, Indonesia can conduct a more in-depth analysis using the national health insurance data samples provided by BPJS [5].

The prediction of Length of Stay (LOS) has been a crucial measure of a hospital's quality of care and effectiveness. Additionally, LOS is recognized as a significant indicator in determining the success of diabetes patient therapy [6]. Prolonged LOS significantly impacts waiting times and imposes a substantial financial burden on patients' families, especially in the cases of diabetes mellitus [7]. This problem raises the need for a system predicting the LOS for Diabetes Mellitus patients to analyze and recommend improvements in the quality of medical care [7]. Machine learning for predicting the length of stay can be developed to estimate the length of stay, aiming to save costs and reduce the duration of inpatient care for patients. There are several

machine learning and deep learning methods commonly used for predicting LOS [7] [8][9], including ANN, SVM, Classification Tree, KNN, Bayesian Network, and Logistic Regression. A study about the length of stay prediction was conducted in study [9] using King Abdulaziz Cardiac Center (KACC) from 2008-2016: and it showed that the highest accuracy was obtained by Random Forest, followed by SVM. Another study [7] using dengue patient dataset in India from 2012 – 2017 obtained the best result using Logistic Regression with the elastic net. The similar study [10] utilized the ischemic stroke dataset from Cangzhou China Central Hospital from 2017-2018, demonstrating that XGBoost outperformed other models with an accuracy of 0.89.

Process Mining, a relatively new and evolving field, involves analyzing process flows using event logs generated by information systems [11] [12]. A critical application of Process Mining is in healthcare [13]. Understanding the healthcare data structures is challenging, necessitating structured methods for explaining and visualization [14]. Several studies have applied Process Mining in healthcare, including the study by Baek et.al [15]. In the study [15], Process Mining and Data Mining approaches were used to predict LOS using Regression. They utilized log data from a General Hospital from January to December 2013. The study also analyzed factors influencing patient LOS, such as the impact of surgical procedures on lengthening patient LOS. In the medical field, the factors influencing patient LOS vary significantly and depend on the type of disease the patient.

This study develops a system capable of predicting the length of stay (LOS) for Diabetes Mellitus patients by comparing three machine learning methods: Logistic Regression, Random Forest, and XGBoost. The study also seeks to identify the process factors affecting patient's length of stay (LOS) using process mining. Logistic Regression has been demonstrated to be effective, sensitive, and capable of handling imbalanced datasets. [7][8][16]. While Random Forest is chosen for this research because it overcomes the problem of overfitting commonly associated with decision trees [9]. Additionally, the XGBoost method is employed in this study due to its characteristics of high accuracy, difficulty in overfitting, and scalability [10].

Meanwhile, Process Mining can assist in detecting the patient process flow. By using Process Mining, healthcare organizations can effectively identify process elements that significantly impact on the duration of the process. By integrating of Machine Learning and Process Mining, BPJS can enhance its processes by leveraging accurate predictions (built through ML) and analyzing process paths (via Process Mining). Through this combination, BPJS can make smarter decisions and enhance the efficiency and quality of their services. Based on the LOS prediction, the study analyses the factors affecting LOS, to provide uses data from BPJS Advanced Referral Health Facility Level *(FKRTL),* Non-Capitated Primary Health Facility *(FKTP),* and Membership dataset for 2015-2018. However, due to limited BPJS data, there is no information regarding the hospital name for each patient. This study focuses on providing recommendations to BPJS Indonesia and patients only.

1.2 Statement of the Problem

Based on the above background regarding the need for diagnosing people with diabetes, because88.51% of BPJS patients are people with Diabetes Mellitus, emphasizing the urgent requirement for more efficient and swift disease management. Furthermore, Length of Stay (LOS) serves as a pivotal indicator in assessing the success of diabetes patient therapy [6]. Previous studies have highlighted the effectiveness, sensitivity, and capacity of Logistic Regression in handling imbalanced datasets [7][8][16]. On the other hand, Random Forest was opted for this investigation because it mitigates the issue of overfitting commonly associated with decision trees [9]. XGBoost technique in study [10] is attributed to its ability to achieve high levels of accuracy, its resilience against overfitting, and its scalability. Healthcare organizations can effectively pinpoint process elements that have a substantial impact on process duration by leveraging Process Mining [11] [12]. So that the formulation of the problem in this study is as follows.

1. Which one is more superior among machine learning model among Logistic Regression, Random Forest, and XGBoost demonstrates superior predictive performance in estimating the Length Of Stay (LOS) for Diabetes Mellitus patients, considering accuracy, precision, F1 score, recall, and prediction time?

2. What are the key factors that contributing to variations in LOS among Diabetes Mellitus patients in the BPJS data for 2015 – 2018?

3. What recommendations can be provided to BPJS Health for efficiency using the process mining method ?

1.3 Objective and Hypothesis

Objective

Considering the problem statement that has been described in the previous point, the objectives of the research carried out are as follows:

1. To identify the machine learning model that provides the most accurate and efficient estimation of the length of stay (LOS) for Diabetes Mellitus patients among Logistic Regression, Random Forest, and XGBoost, considering metrics such as accuracy, precision, F1 score, recall, and prediction time.

2. To identify and analyze the factors that contributing to variations in LOS among Diabetes Mellitus patients in the BPJS data for 2015 – 2018.

3. To provide recommendations to the length of stay (LOS) efficiency within BPJS healthcare service processes using Process Mining techniques.

Hypotesis

Based on the problem statement, the hypothesis of this study posits that at least one of the machine learning instances among Logistic Regression, Random Forest, and XGBoost displays superior predictive performance in estimating the Length Of Stay (LOS) for Diabetes Mellitus patients. This consideration includes metrics such as accuracy, precision, F1 score, recall, and prediction time. Additionally, specific factors are anticipated to exert a notable influence on LOS. Process Mining is projected to yield valuable insights into the patient flow within the healthcare system, unveiling critical process elements that affect LOS. Ultimately, through the analysis of LOS prediction and the impact of factors on LOS, the study aims to provide actionable

recommendations to both BPJS Indonesia and patients, to enhancing the quality and efficiency of healthcare services for Diabetes Mellitus patients.

1.4 Scope and Limitation

The scope of this study concentrates on Indonesia, particularly utilizing the Indonesia Health Insurance Agency/ *Badan Penyelenggara Jaminan Sosial* (*BPJS)* with a focus on Diabetes Mellitus in Advanced Referral Health Facility Level (FKRTL) and Non-Capitated Primary Health Facility (FKTP). The research employs Logistic Regression, Random Forest and XGBoost for predicting Length of Stay (LOS) and utilizes Process Mining techniques to identify process factors influencing LOS. Comparative analysis valuate the proposed model against alternative methods commonly used in LOS prediction for diabetes datasets. The study covers the time frame from 2015 to 2018, utilizing BPJS Membership datasets.

The limitation of this study is the limited BPJS data, the study lacks detailed hospital name-specific information for each patient, restricting the analysis to broader trends within the healthcare system. Recommendations primarily target BPJS Indonesia and patients, excluding a broader healthcare system perspective. The exclusive focus is on Diabetes Mellitus patients, potentially excluding insights into LOS prediction and factors for patients with other medical conditions. Acknowledging potential data limitations, particularly regarding comprehensiveness of BPJS datasets and the absence of certain hospital name-specific variables, it is important to note that the study is temporally constrained to the specified time frame (2015-2018). Therefore, the findings may not necessarily reflect the current healthcare landscape or recent developments.

1.5 Importance of the Study

The study focuses on providing actionable recommendations specifically tailored to BPJS Indonesia. These recommendations can potentially guide policy decisions, resource allocation strategies, and overall healthcare management practices within the organization. The findings could empower healthcare providers and patients with valuable information for making informed treatment plans and expectations decisions.

The study contributes to the existing body of knowledge by combining the predictive capabilities of Logistic Regression, Random Forest and XGBoost with the insights derived from Process Mining in the context of Diabetes Mellitus patient LOS. This interdisciplinary approach could pave the way for further research and advancements in healthcare analytics.