CHAPTER 1: THE PROBLEM

1.1 Background

Quoted from the 2016 WHO data, 70% of total deaths in the world are caused by diabetes, and 90% - 95% of diabetes cases are type 2 diabetes, which is mainly preventable because it is caused by an unhealthy lifestyle [1]. Diabetes mellitus is a chronic metabolic disorder caused by the pancreas not producing enough insulin or the body unable to use the insulin effectively [2]. In Indonesia, according to Basic Health Research (RisKesDas) in 2018 [2], people with diabetes from 2013 to 2018 increased gradually, where 6.9% of Indonesia population is diabetic. 69.6% of those with diabetes were undiagnosed, and 30.4% diagnosed. Meanwhile, in 2013, 5.7% were diabetic. As many as 73.7% of these people with diabetes were undiagnosed, and 26.3% were diagnosed. This data shows that diabetes mellitus is a dangerous disease since it can lead to various complications of other diseases, such as heart disease, kidney failure, stroke, and even paralysis and death [2].

The prevalence of diabetes mellitus (DM), based on a doctor's diagnosis in the population aged ≥ 15 years, is increased to 2% based on the report of RisKesDas 2018 [2]. The largest DM sufferers are in the age range of 55-64 years and 65-74 years [2]. In 2018, the percentage of DM sufferers for female (1.8%) and male (1.2%) [2]. As for domicile areas, the percentage of DM sufferers in urban areas (1.9%) than in rural areas (1.0%) [2]. The highest estimated number of DM cases in Indonesia will occur in 2030, with a total population of 21.3 million [2]. Based on RisKesDas diabetes data [2], undiagnosed patients can be detected beforehand. Diabetes detection could be performed by a doctor based on blood sugar and insulin levels or conducted automatically based on individual medical checkup data.

Prediction of diabetes diagnosis using data can determine whether the patients have diabetes or not. Several studies discussed diabetes diagnosis prediction based on data [3-21]. Besides the Pima Indian dataset [3-17], there are also several diabetes datasets used for research such as hospital data in Luzhou China [4], repository California University in Irvine [18], laboratory diagnosis in Kashmir [19,20], Dr. Schorling [10,21] and online questionnaire [21]. There are various classification methods on diabetes diagnosis prediction like random forest [4,6,10,13], J48 [5,6,9], Naïve Bayes [8,9,10,12,17], support vector machine [8,12,13,15,16], logistic regression [10], neural network [5,10,15], and K-Nearest Neighbors [9,12,14,15].

In previous studies, the classification and prediction of DM with Pima Indian data have been carried out using several machine learning methods. However, only a few studies discussed preprocessing on Pima Indian dataset. The problem of missing value is discussed in a limited number of papers [9,11,12,15,16]. The problem of imbalanced data [11,13,14] and feature selection [5,10,13,17] have been discussed too. Several models have been used in data preprocessing, such as missing values using median [9], Interquartile Range [15,16], mean [12], and Naive Bayes [11]. In imbalanced data, there is Synthetic Minority Over-sampling [13,14], Random Oversampling [13], and Adaptive Synthetic Sampling [11]. Meanwhile, in feature selection, there is Principal Component Analysis [5,10], Maximum Relevance and Minimum Redundancy [5], Fisher Discriminant Ratio [10], Analysis of Variance [10], Information Gain [13], and forward-backward [17] models.

According to several prior studies on diabetes prediction, important factors that contribute to classification accuracy are imbalanced data, the presence or absence of missing values, and features that affect the results [4,7,8,11,12,14,16,17,19-22]. In addition, the paper explains that data augmentation can improve the accuracy of diabetes prediction by solving the data imbalanced problem [23]. Data augmentation is an algorithm used to augment the observed X data with a quantity of Y, referred to as latent data [24]. In the Pima Indian dataset, imbalanced data occurs in the class label. Imbalanced data is a problem related to the performance of learning algorithms faced with underrepresented data, and the slope of the class distribution is severe [25]. The missing value is a problem that replaces the null value in a variable [10]. The maximum limit for missing value varies from 5-10% and 50% [26]. Feature selection is an important problem in machine learning since it gets the most informative features [10].

In this study, we propose an approach to dataset preprocessing, which is applied to diabetes prediction. The preprocessing approach consists of the following process: missing value process, imbalanced data process, feature importance process, and data augmentation process. This study aims to improve the precision and recall outcomes in diabetes prediction using data preprocessing.

1.2 Problem Identification

This study will use data from Pima Indian and Karya Medika which contains diabetes-related parameters. There are 3 problems with the data used to be discussed in this research, namely, first about imbalanced data [10,11,13,14,16,19,20,22] with the number of diabetics or not

diabetes. The second problem is about missing values [9,11,12,15,16,22] where there are still values that have no value or are presented with a value of 0/NaN/{}. The third problem about feature selection [4,9,10,17,22] where there are features that are not very influential in producing this prediction is seen when trying initial experiments between variables having different values sometimes too high and too low. The third processes are considered important because they can affect the results of the classification of diabetes or no diabetes, if the third processes are not carried out then what should be detected by diabetes, but the result is not diabetes or who should not have diabetes but detected diabetes. This is certainly very dangerous for the application of the results later.

1.3 Scope of Work

The scope in this study is dealing with missing values, imbalanced data, and feature selection for the classification of diabetes data. These problems need to be resolved because they affect the results of the classification of the dataset.

1.4 Objective

The contribution of this study is to improve the detection of type 2 diabetes disease by using a means in the process of preprocessing data. The system that will be built in this research is to handle the problems of missing value, imbalanced data, and feature importance to the dataset. This system will be able to fill in the values in the dataset that is still null, be able to sort the features that affect diabetes, be able to balance the amount of diabetes data between those with diabetes and those without diabetes, balance the amount of data for men and women, and balance the amount of data used fasting and not fasting. Besides that, for imbalanced problems, data augmentation techniques based on a distribution and related parameters can be used. After the problem process is resolved, the system is also able to classify the dataset which aims to improve precision and recall results in classification.

1.5 Hypothesis

To improve the precision and recall of diabetes classification it will use the median to overcome missing values in the data, Gini random forest is used as the feature importance to order the most influential features, oversampling method to overcome imbalanced class, gender, and fasting or not fasting on data, and then the data augmentation technique is used to combine the two datasets. The data classification will use entropy random forest.