CHAPTER 1
INTRODUCTION

1.1. Background

Medical workforce fundamentally is among the most important assets that a nation should have. Indonesia, as a large country by population, should naturally possess a high amount of medical workforce, but this is far from the case. In 2021, there are merely 567.910 listed medical workforce in Indonesia. According to Direktorat Jenderal Kependudukan dan Pencatatan Sipil (Dukcapil), only 0.2% of Indonesia population are medical practitioners, which he stated as concerning.

Medical workforce is very valuable, even more during the pandemic times. Shortage of medical workforce in a certain location might lead to patient abandonments, even casualties. It is important for the government to distribute medical workforce evenly and to act preemptively before a shortage can happen. Data prediction algorithm is a modern mathematical approach that can be utilized to provide future medical workforce number within a certain region. Prediction algorithms require training data for them to do the prediction. The problem is data availability of annual medical workforce in Indonesia is very limited, not enough to train the prediction model.

Data synthetic technique called Synthetic Data Vault or SDV is capable of solving the data scarcity by generating more data. This technique uses deep learning method to generate high quality data based on existing data. The research utilizes two algorithms the author deemed the most fitting based on previous research. Random Forest is selected because it has high resistance against overfitting problem which SDV generated data notoriously possess. Support Vector Regression is used because it is memory efficient, generally accurate, and effective for smaller datasets. The prediction results of implementing SDV with aforementioned algorithms are evaluated using MSE and MAPE to find their accuracy values. The output of this research will serve as a consideration material for the government or any institute to improve medical workforce management through machine learning.

1.2. Problem Formulation

Based on previously explained background, the problems can be described as follow:

1. How is the real data compared to SDV generated data?
2. Based on its accuracy, is SDV technique eligible to substitute a real medical workforce dataset?
3. Which algorithm is the best to be used with SDV?

1.3. Scope

The author's project is done using Python 3.0 and Spyder IDE. The dataset used in this research is taken from Central Statistics Bureau (BPS). Data used in the research are medical workforce data before the pandemic. The reason behind this is the fact that workforce trend changes significantly during the pandemic situation. The research covers implementation analysis of SDV, SVR, and RF. They are examined in three different perspectives, accuracy test, processing time, and data statistics, such as mean and deviation standard.

1.4. Objectives

The purpose of this project is to analyze SDV eligibility in increasing training data availability and achieving higher accuracy numbers when used with SVR and RF algorithm. This analysis provides effectivity, advantages, and disadvantages of aforementioned techniques in predicting future medical workforce number. Information and data discovered from this analysis can be used in future researches as a base to explore the implementations of prediction algorithm and synthetic data.