CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

This paper will be looking at a car price prediction model using k nearest neighbour (KNN) and random forest. The K-Nearest Neighbors (KNN) algorithm is a supervised learning algorithm which uses the training data to predict new observations. Random forest is an ensemble which creates multiple decision trees using the data. The features used in this model are the car's ID, price, manufacturer, model, production year, fuel type, and drive wheels. We will train our model on historical prices of cars from 1939 to 2020.

"The k-nearest neighbors algorithm, usually referred to as KNN or k-NN, is a supervised learning classifier that use proximity to produce classifications or predictions about the grouping of a single data point. Although it can be applied to classification or regression issues, it is generally implemented as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another."[1]

A random forest, as its name suggests, is a collection of numerous diverse decision trees. Our approach predicts the class that receives the most votes. Each tree in the random forest produced a class prediction. Random forest is built on the simple yet powerful premise of the wisdom of crowds. The widely used machine learning technique known as random forest, which mixes the output of several decision trees to obtain a single outcome, was developed by Leo Breiman and Adele Cutler. Its versatility and usability, which it uses to address classification and regression problems, are what lead to its widespread use.

This project will use RMSE to find out which algorithm is the best. RMSE is used to measure the error rate of a model in predicting a numerical value. The lower the RMSE value, the better the model's prediction accuracy.

1 https://www.ibm.com/id-en/topics/knn

4.2 Desain

The data in this project indicates what data will be used for predictions. This project's dataset was obtained from Kaggle. Dataset Car Price was owned by Deep Contractor 5 month ago. This datasets has 19237 rows and 18 columns. Dataset used in this model are the car's ID, price, manufacturer, model, production year, fuel type, and drive wheels. In the orange application, select data, then select CSV file import. After that choose Select columns select, then select Data transform. Following that, target the attribute price in select columns. Then create data sampler. In the data sampler, determine the fixed proportion of data. Then connect the CSV file import to the data sampler to select columns.

Transform in this project means targeting which column will be used for predicting and how much data will be used. In the orange application, select data, then select CSV file import. After that choose Select columns, then select Data transform. Following that, target the attribute price in select columns. Then create data sampler. In the data sampler, determine the fixed proportion of data. Then connect the CSV file import to select columns to the data sampler.

Model in this project mean what algorithm will be used to predict the dataset. Here, K-Nearest Neighbors (KNN) and Random Forest algorithms are employed (RF). The algorithm receives sample data to produce predictions. Choose kNN and Random Forest in the orange program. Establish a correlation between training data and datasets from the Data sampler.

Evaluation + reports in this project means that after the dataset is entered into the model, prediction results will come out. Then compare the results of both algorithms to see which produces better predictions. In the orange application, select evaluate then select Prediction. Then correlate Data sampler to Prediction. In edit links, change the connected link from data sample to remaining data.

Figure 4.1: Flowchart Orange

The RMSE method will be used in this project to determine which algorithm is the best. The root mean square error (RMSE) is used to calculate a model's error rate when predicting a numerical value. The lower the RMSE value, the more accurate the model's prediction.

CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1 Implementation

First, csv file inserted into orange. The contents of raw dataset columns can be viewed in the import option. CSV File Import is linked to Select Columns.

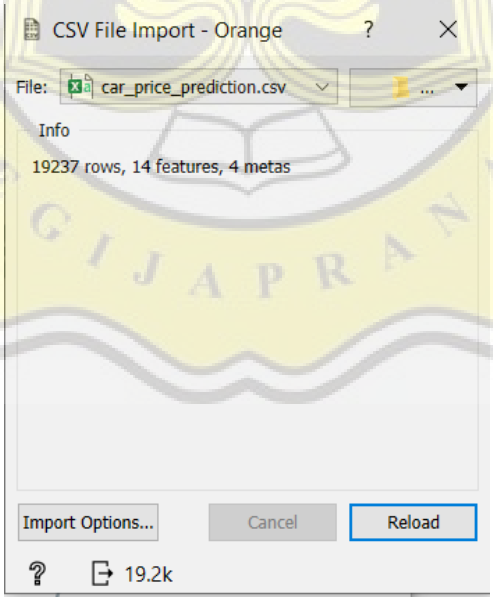

Figure 5.1: CSV file import