Chapter 5
Classification

ABSTRACT

In the world of data mining, classification reigns supreme as a popular technique for supervised learning. Its ability to identify patterns in data by dividing it into training sets and utilizing machine learning makes it an essential tool in answering critical questions related to data. For instance, classification can aid businesses in identifying customers with high purchasing potential. One of the standout features of classification is k-nearest neighbors (k-NN), which allows data to be classified according to the training data set. Decision trees are also commonly used to support decision making by producing easily interpretable diagrams. RapidMiner is an outstanding data mining tool that can employ a range of classification techniques, including k-NN, decision trees, and naïve Bayes. In this book, readers can follow a step-by-step guide to using these techniques with RapidMiner to achieve effective data classification.

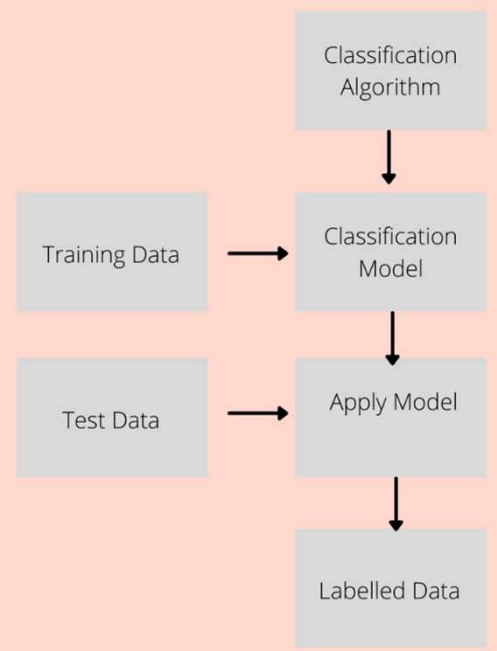
INTRODUCTION

Data classification is the intension to classify or identify the data such as the classification of customers who are likely to change mobile phone companies. The results obtained from the analysis are the Discreate or Categorial Data, which indicates the cluster or the type of data (Vichi, Ritter & Giusti, 2013). In data science, this group of data is referred as a Class Label. To have a variable target is to classify data using the principle of Supervised-Learning in data processing by dividing the data into 2 parts (Mishra, & Vats, 2021). Part 1 is for teaching machines to learn, and part 2 is to test the performance of the model. The data classification techniques are processed as follows.

As seen in the figure, data scientists can use Classification Algorithm such as K-Nearest Neighbor (k-NN) or Decision Tree to create Classification Model using Training Data to teach the machines to learn in order to obtain the needed results from data classification (Liu, 2021: Mladenova, 2021). Then, Test Data is used to apply the model to test the model's accuracy performance. The test data and the results obtained from the classification are compared.

DOI: 10.4018/978-1-6684-4730-7.ch005

Figure 1. Data analysis with classification



GENERATING TRAINING AND TEST DATA SET

The main principle of Supervised-Learning is to divide the data into 2 parts, consisting of Training Data Set and Test Data Set.

Holdout Method

To divide the data set by Holdout Method, the data is divided into 2 parts including Training Data Set 70% and Test Data Set 30%. However, if the Holdout method is used in cases where the data set is small and is still allocated for model testing, the model lacks the opportunity to learn the nature of the data and ultimately reduces the processing accuracy.

Cross Validation

Cross Validation method is to determine the number of rounds of division into k cycles by dividing the data into 2 parts in every cycle (Mnich et al., 2020). For example, the number of data division is determined and k is equal to 4. Therefore, in Round 1, Cross Validation will identify the part 1 data as a Test Data Set, and parts 2 – 4 as Training Data Set. In the second round, Cross Validation will indicate the part 2 data as Test Data Set, and parts 1 and 3 – 4 as Training Data Set. In the third round, Cross Validation will determine the part 3 as Test Data Set, and Parts 1 - 2 and 4 as Training Data Set. And finally in the fourth round, Cross Validation will set the part 4 data as Test Data Set, and Part 1 - 3 as Training Data Set, as seen below.