CHAPTER 3
RESEARCH METHODOLOGY

3. 1 Data Collection

The dataset used in this project is a downloadable dataset from MovieLens.org itself but I got this dataset from Kaggle. This dataset itself has 2 sheets, namely master movies and master ratings. In the master movies, there are several variables such as movieId, title, and genre. Then the master rating itself has userId, movieId, rating, and timestamp variables. With a rating value range from 0 to 5.

From the data above we can take several factors that can support this project. We can do collaborative filtering as well as nave Bayes by using ratings from item by item, then we can also use user's watch data to find similarities between one user and another.

3.2 Algorithms

The algorithm that will be used in this research is item-based collaborative filtering and also naive Bayes. Both are used because both algorithms are quite widely used in making recommendation engines. Then the research at this time will be the main factor is the rating. Here we can predict a film that we will recommend to users by calculating the similarity of one item to another based on its rating using the naive Bayes method, and can also recommend films using the item-based collaborative filtering method.

3.3 Design

The data that will be processed is data that comes from the master rating, with the movieId, we simply need to manage from the movieId first, then when we display the recommendation results we will take data from the master movie CSV which contains information on the title of the film and also the genre. The amount of the data will be calculated by comparing the amount of all data, then from the accumulated data, we will calculate the percentage as a benchmark for similarity. For collaborative filtering, we will use an algorithmic approach with the help of Cosine Similarity to find the closest distance. Then for the nave bayes itself, we will try to calculate the distance based on the percentage of each film.

3.4 Coding

Here the programming language that I will use is python 3.8 because python can process large amounts of data. In python, some libraries are quite adequate in the process of working on this research, but in the algorithm that I will be working on I will not use any libraries. The library that I will use is like pandas to read CSV files, then there is NumPy to perform the calculation process and also to change some forms of matrices, then I will use matplotlib to display some graphs as a form of visualization, and other libraries.

3.5 Analysis

This project aims to measure the accuracy of the two algorithms, namely item-based collaborative filtering with nave Bayes in making a film recommendation engine. With the dataset obtained from MovieLens, there are 10,000 data ratings from users and about 149,000 movie title data. We will test the level of accuracy and efficiency. The results of this analysis can be seen using the MSE and RMSE methods as a benchmark for the comparison of these two algorithms. The one who gets the smallest value from the test results will have the best performance.