# Data Analytics with Orange

This analysis predicts customer spending scores based on various attributes presented in Stella Pty. Ltd dataset using various prediction models such as Linear Regression, KNN, SVM, and others. The dataset contains 1000 instances with 7 attributes in total. The dataset contained several categorical data that were converted into numerical data for a more accurate prediction, such as "satisfaction," which had values of "high," "medium," and "low" was converted into three numerical columns called "Dummylow," "Dummymedium," and "Dummyhigh" (figure 3).

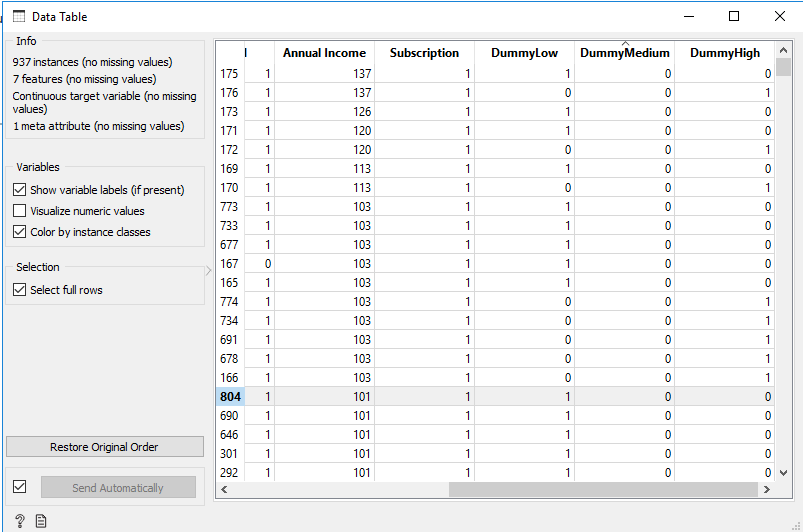


Figure 3: Data table showing dummy variable created for categorical column

## 3.1 Missing value treatment

The original dataset contains some missing values, which may have an impact on the result. It is critical to use the correct approach for missing values in order to maintain the data set's accuracy and dependability. The original dataset illustrates that the missing values are arbitrary and account for only 6.3% of the total data set, or 63 out of 1000 cases. As a result, the absence of the value may have no effect on the analysis. Thus, the missing data were excluded because replacing it with the average was affecting the data quality. The IMPUTE (Figure 23) widget was used to remove the instances from the dataset that had unknown values. After the missing values were removed, the dataset was saved for analysis and prediction (figure 8).

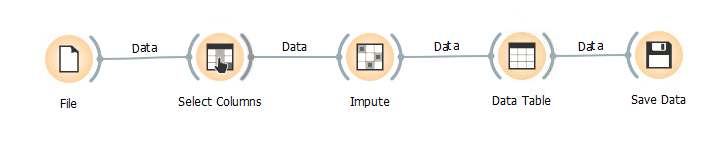


Figure 4: Data cleansing process using impute

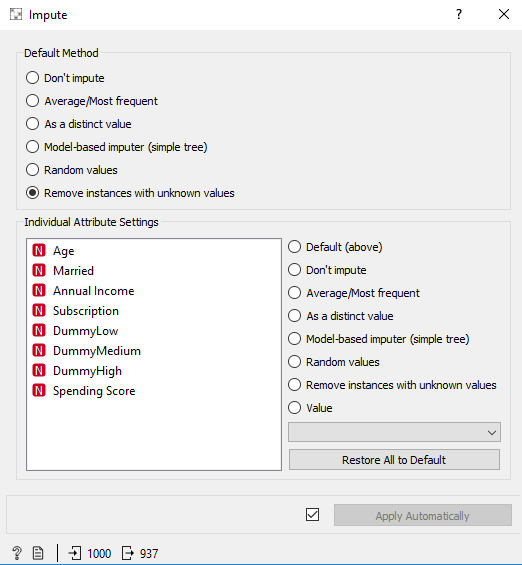


Figure 5: Impute widget

## 3.2 Prediction Model(s)

The goal of this study is to forecast the spending score of various Stella Pvt. Ltd clients, therefore many prediction models may be utilized to select the most appropriate and accurate model. To begin with the prediction of the spending score, the 'spending score' variable is set as the target variable, and the 'customer ID' variable is placed in the meta-attributes of the select columns widget.

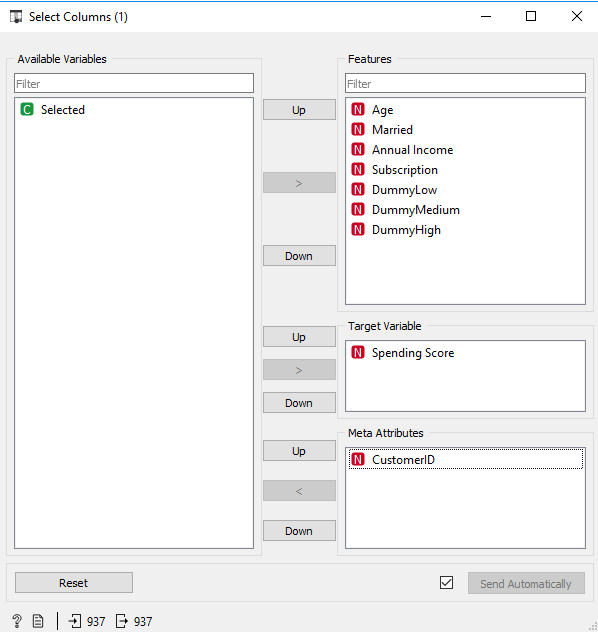


Figure 6: Selecting columns

Following the selection of suitable columns, several percentages of testing data were used during the prediction to discover which model, at what percentage of testing data, best fits the dataset. Using the Data Sampler widget in Orange, 10%, 20%, 30%, and 40% of the testing data in this case study were selected for comparison.

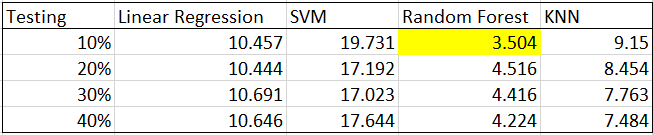


Figure 7: Result from different percentage of testing data

It can be observed that Random Forest model exceeds the Linear Regression, KNN and SVM in terms of RMSE findings. According to the results, Random Forest model’s predicted results are closest to the actual values.

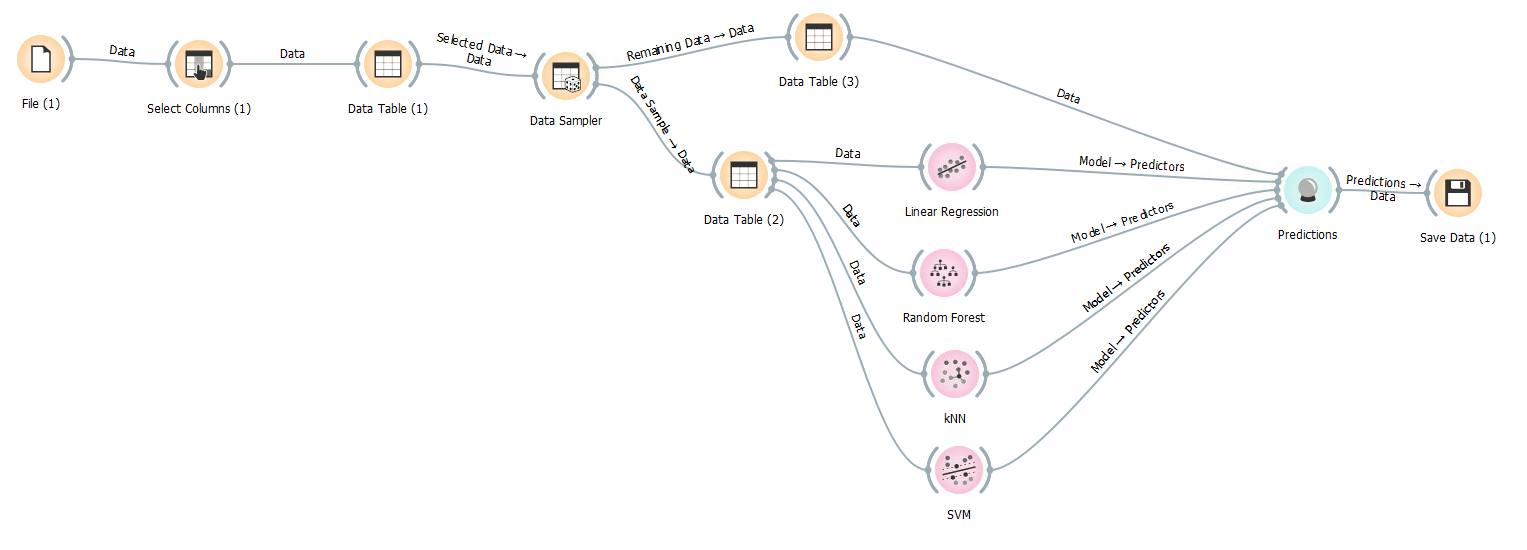


Figure 8: Complete widget path

## 3.3 Data Visualization

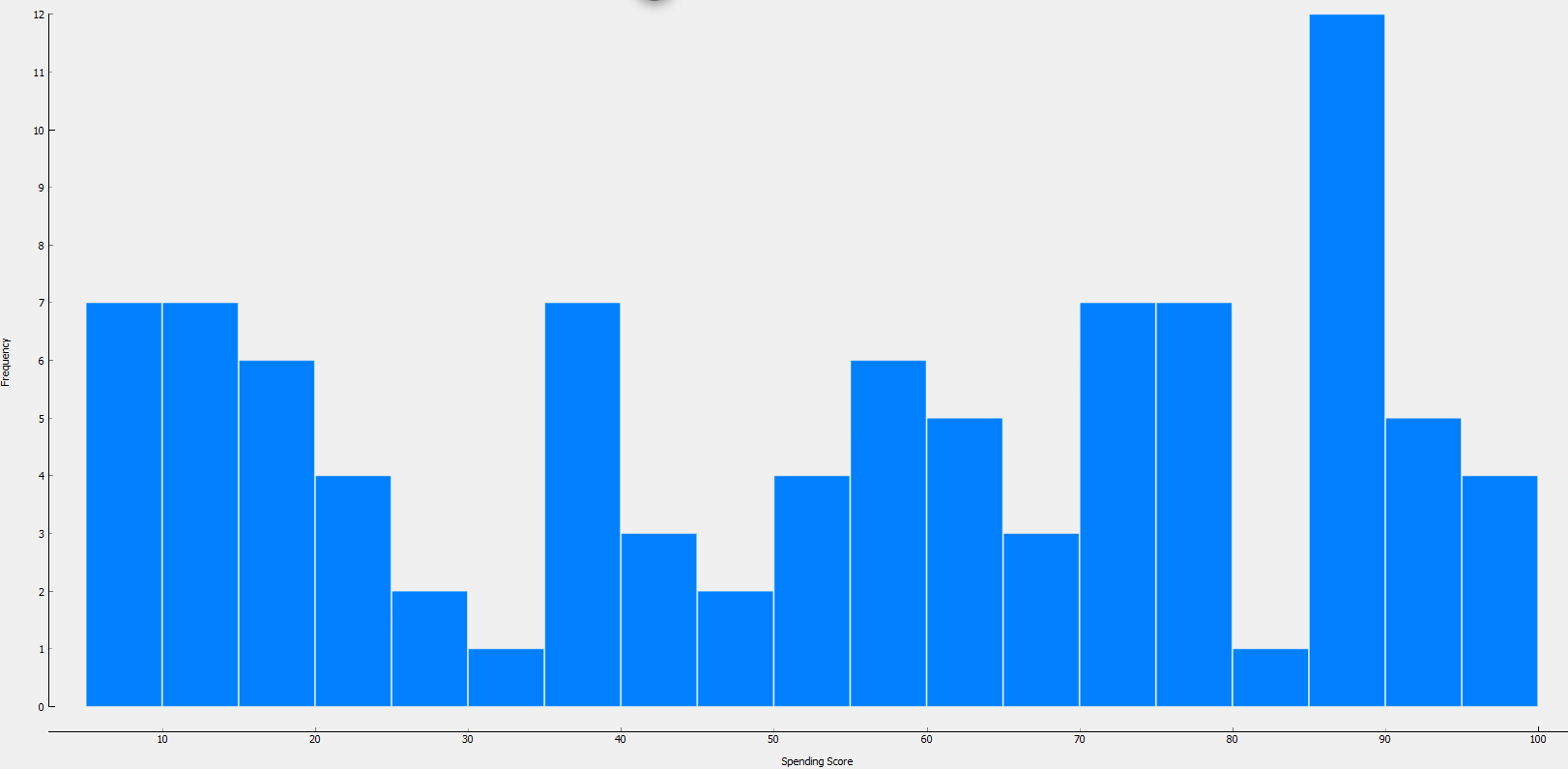


Figure 9:Spending score

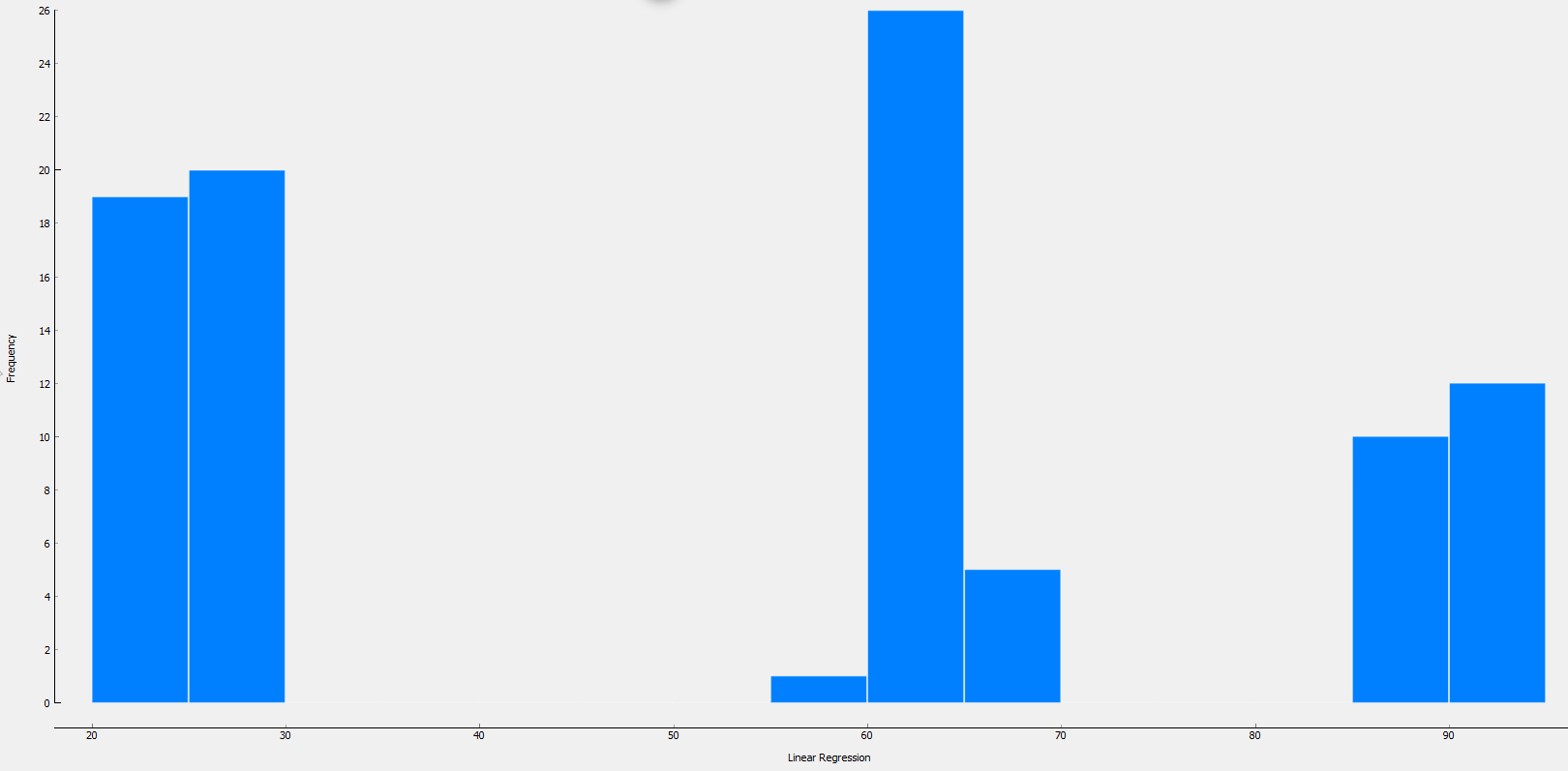
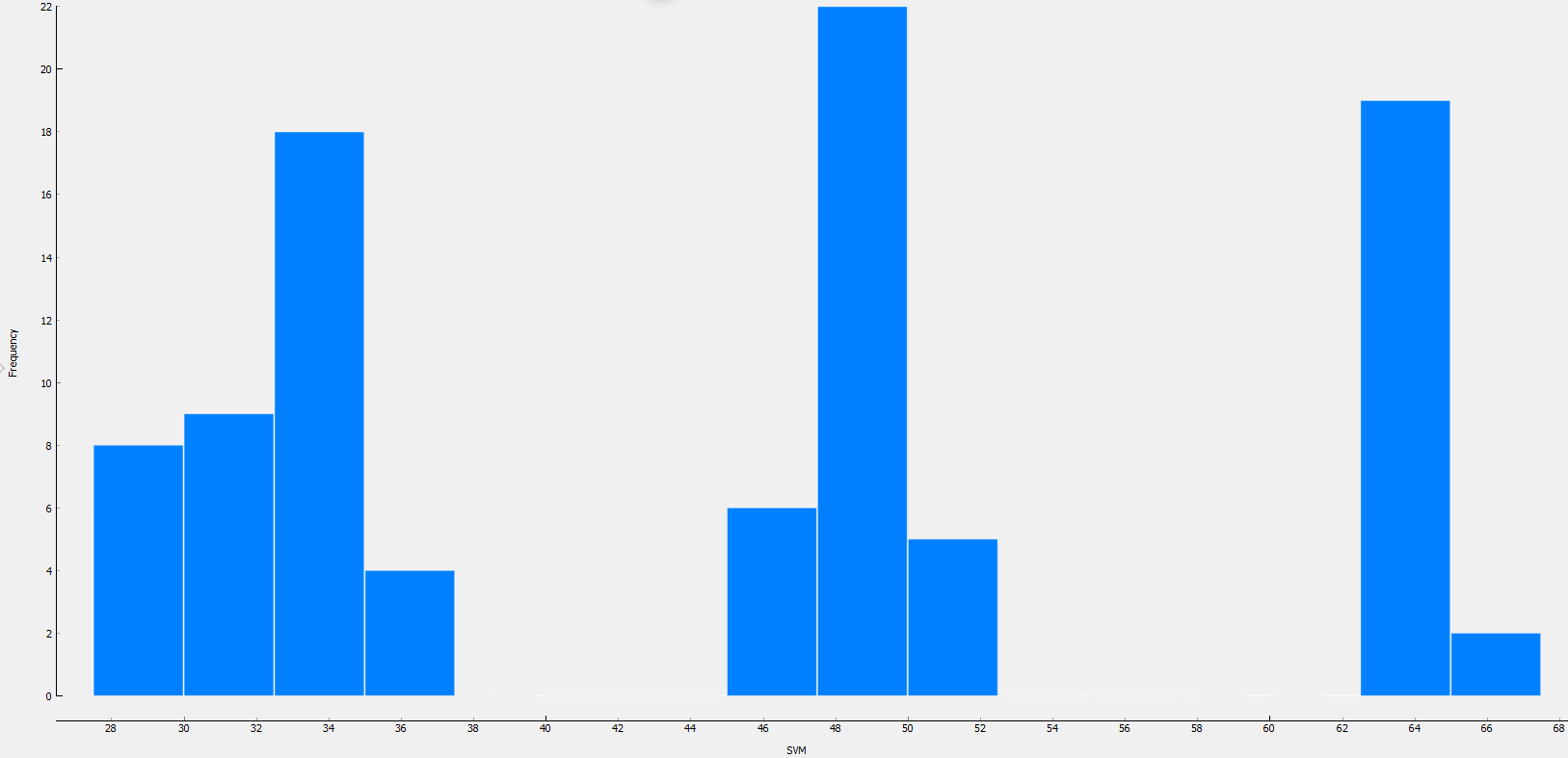
 

Figure 10: Linear Regression Figure 11: SVM

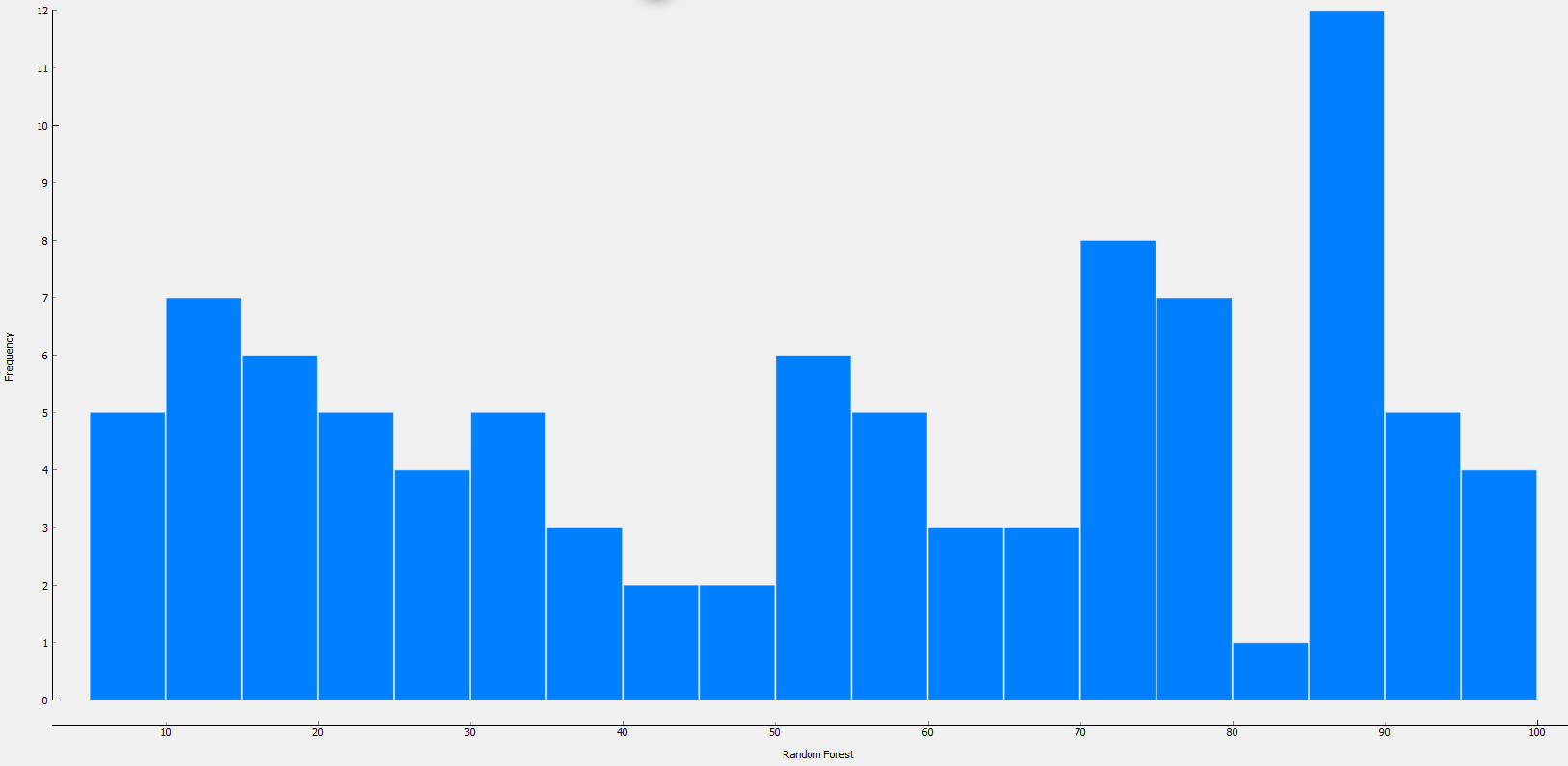


Figure 12: Random Forest

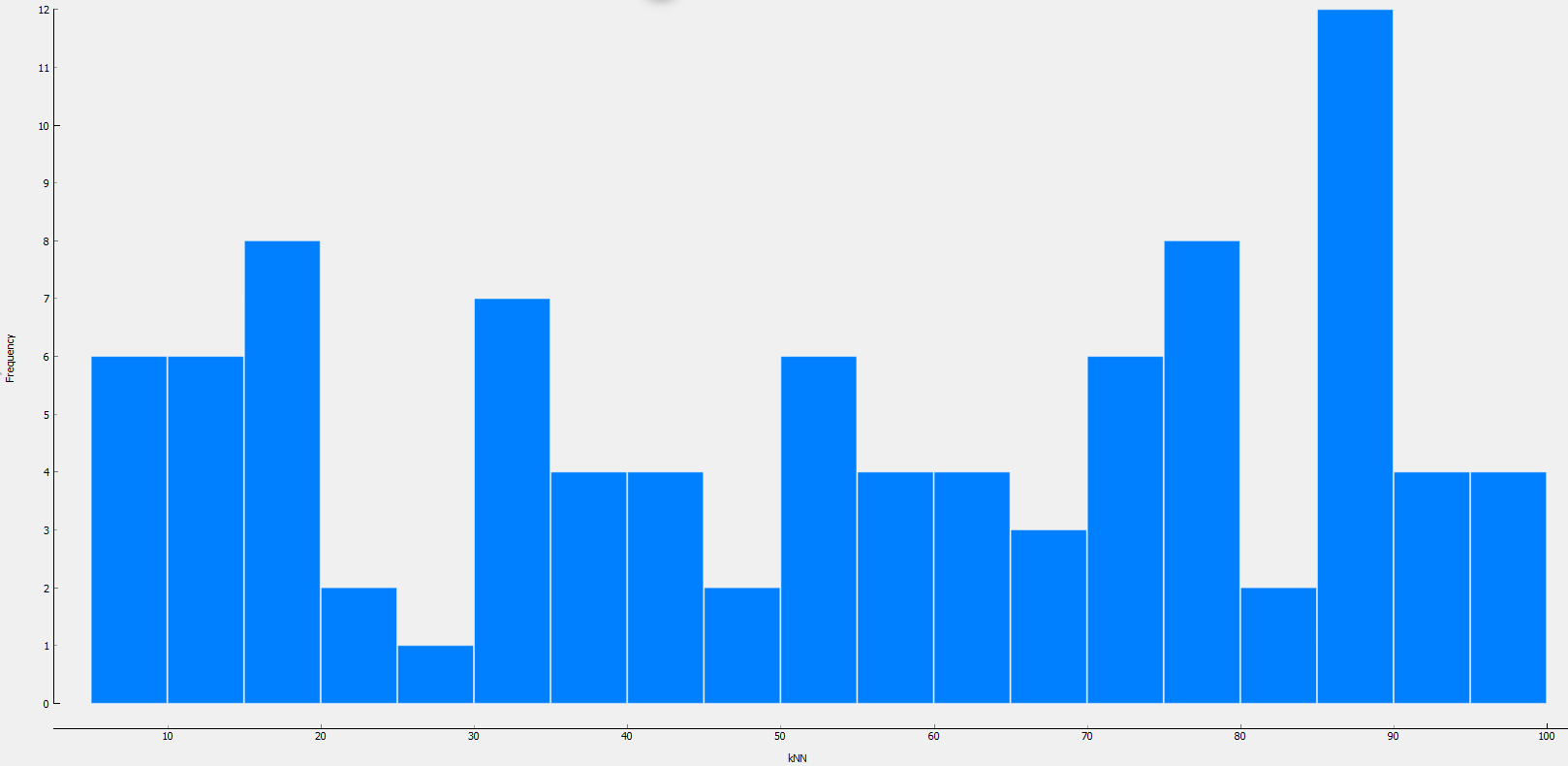


Figure 13: KNN

## 3.4 Business value of prediction models and visualizations

The prediction models and visualisations determine how a specific attribute can be determined using other corporate attributes. In this case, the expenditure score is determined by factors such as satisfaction, age, annual income, subscription, and marital status. As can be seen from the visualisations, Random Forest demonstrates that the frequency of spending score is close to 90, which is similar to actual values. Similarly, Stella Pvt. Ltd. can extend its business by providing many more firm attributes as input to these models to find the important attributes. For example, profit generation from various products based on categories, price, discount, and so on.

# APPENDIX

**Data Modelling**

1. Different type of clients and employees have been shown in the data modelling using the EER approach.
2. It is assumed that the client places an order which creates an entity called places. After the order has been placed, the order line will be created which stores the quantity of products in the order.
3. An order may or may not have shipment as the customer can buy the product from the store as well.
4. An order can have many shipments just in case the previous shipment was lost.
5. An order can also have many invoices if there is change in the amount to be paid by the client.