|  |  |
| --- | --- |
|  | Analysis of ABC Beverage manufacturing process. |
|  |  |
|  | Maryluz Cruz, Bill Stepniak, Sherranette Tinapunan  Data 624 – Predictive Analytics – Project # 2  5/23/21 |

Table of Contents

Executive Summary …………………………………………………………………………………………..2

Problem Statement …………………………………………………………………………………………..2

Data Analysis …………………………………………………………………………………………………….3

Model Building ……..……………………………………………………………………………….. 4

Important Variables …………………………………………………………………………………4

Model Evaluation …………………………………………………………………………………….5

Prediction Result ………………………………………………………………………………………………5

**Executive Summary**

ABC Beverage company is a manufacturer that make beverages that seem to produce mostly alkaline beverages. Our team was given factual data on the company’s manufacturing process for the different beverages that they make. The data was given to us so that we can determine the pH Levels of the beverages. Different models were built to see which model can improve the accuracy of the prediction. The model that had the highest predictions was chosen to be evaluated with the evaluation data to see the prediction levels of the pH. There are three processes that can cause a shift in the levels of pH which are: **Mnf.Flow**, **Usage.cont**, and **Filler.Level**. When it comes to manufacturing in the future the manufacturer should pay close attention to these processes while they are evaluating the currently produced pH levels.

**Problem Statement**

New regulations have been set to ABC Beverage, the person in charge of the production is now required to understand their manufacturing process. Along with the manufacturing process they have to understand the predictive factors, and how it relates to pH of the companies beverages. For this project we were required to find the most optimal predictive variables that are related to the pH levels of the beverages and evaluate its accuracy by testing different models.

**Data Analysis**

In order to do a data analysis we must first explore the data, and see if there are any missing values (Figure 1), outliers, see how the data is distributed, and see if any of the values that may be correlated. Figure 2 shows the distribution of the pH variable which is the Target Variable. After the data is explored, the data then needs to be prepared. In data preparation the missing data points are removed, the distribution of data is normalized, and any other preparation method that is necessary. After that is done the data that has been prepared gets split into train and test datasets, and those datasets are used to build and evaluate the models.

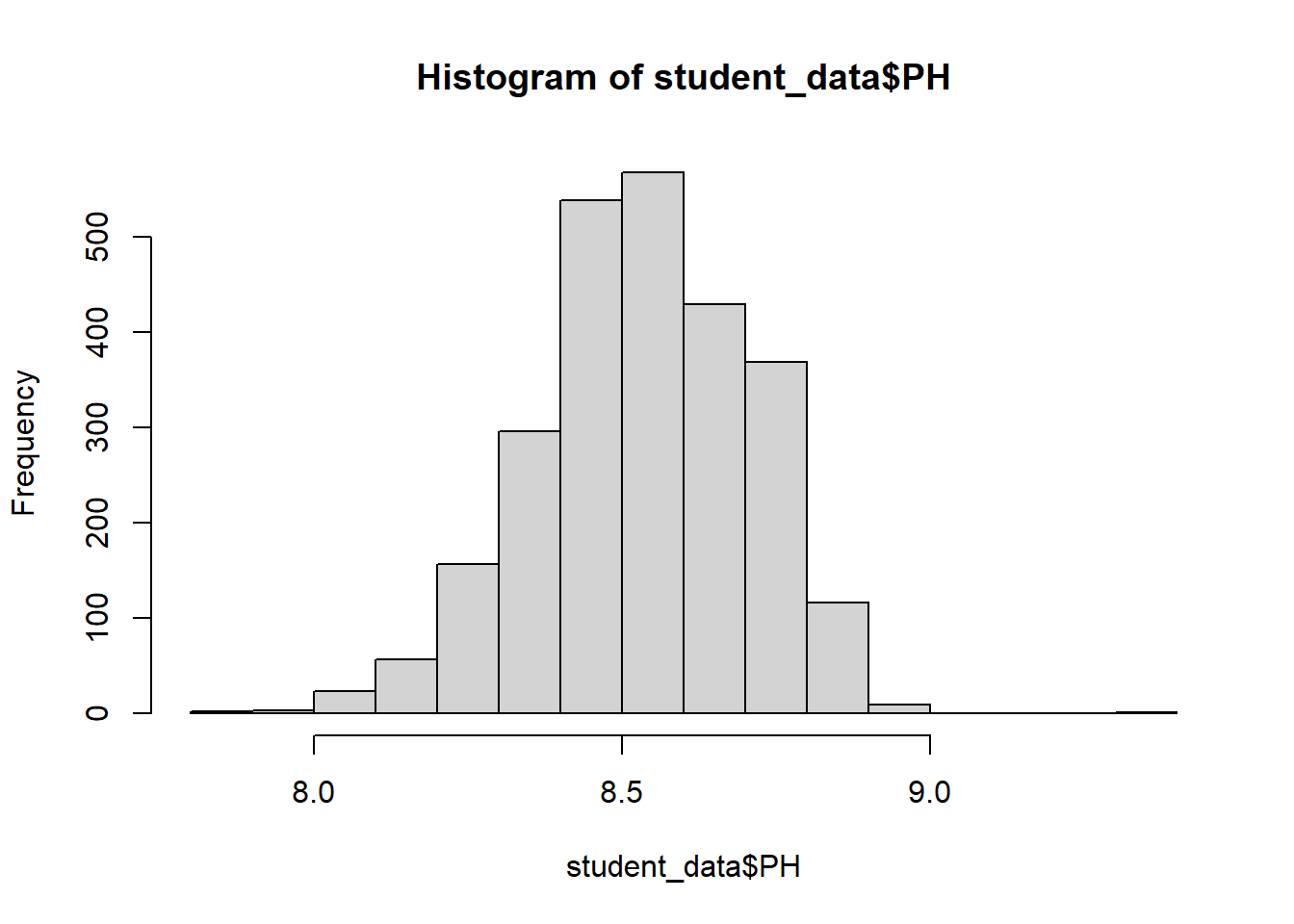
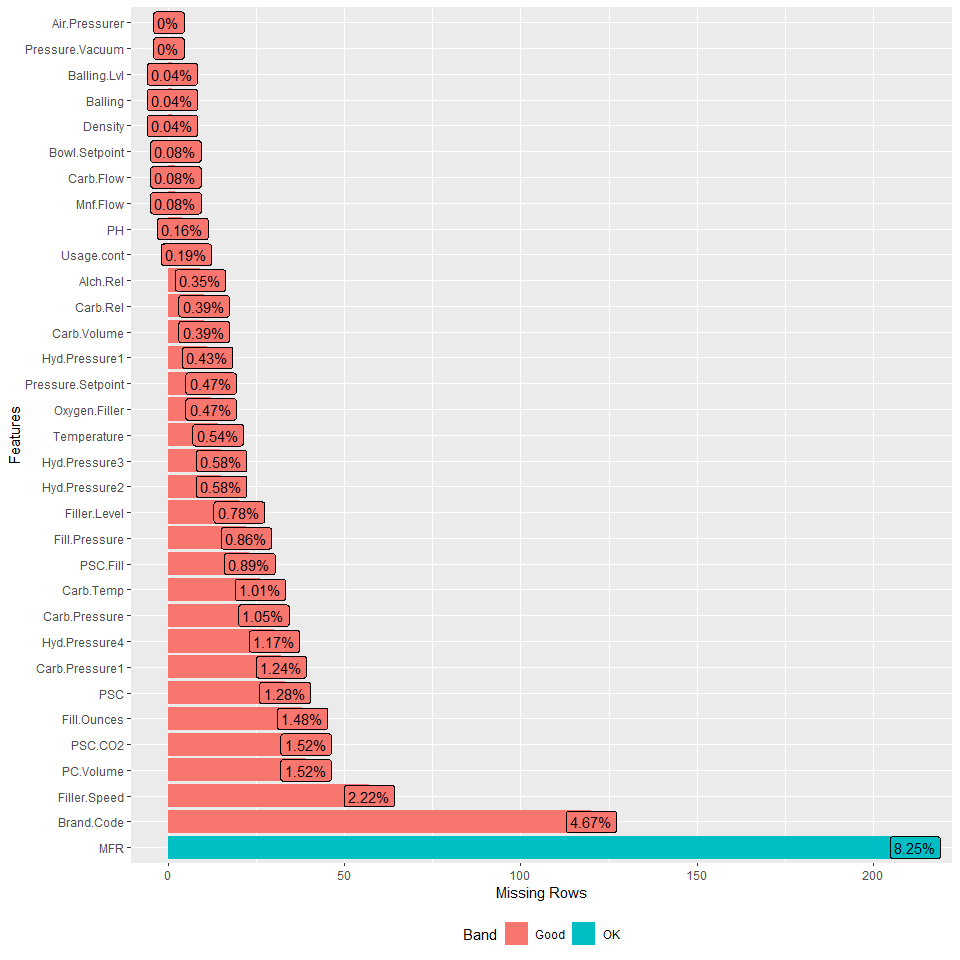


Figure 1: Missing Data Figure 2: pH Target Variable

**Model Building**

Using the training dataset different models are built, and the models built are either Linear, Non-Linear and Trees. For the Linear models: Partial Least Squares, Elastic Net, Ridge and Lasso. For the Non-Linear models: K-nearest Neighbors, Multivariate Adaptive Regression Splines, and Support Vector Machines. For the Tree models: Random Forest and Gradient Boosting Machine. After testing the data it was discovered that the Random Forest gives the lowest RMSE, and the highest RSquared which determines the optimal model.

**Important Variables**

After finding the optimal model, the next step is to find out the important features of the model. In the Random Forest(Figure 3), the top 3 important features are **Mnf.Flow, Usage.cont, Filler.Level. Min.Flow** happens to be the most important feature with all the models. **Mnf.Flow** can greatly affect the pH value of the beverages the most. **Usage.cont**. and **Filler.level** can affect the pH values as well but not as greatly as **Mnf.Flow**.

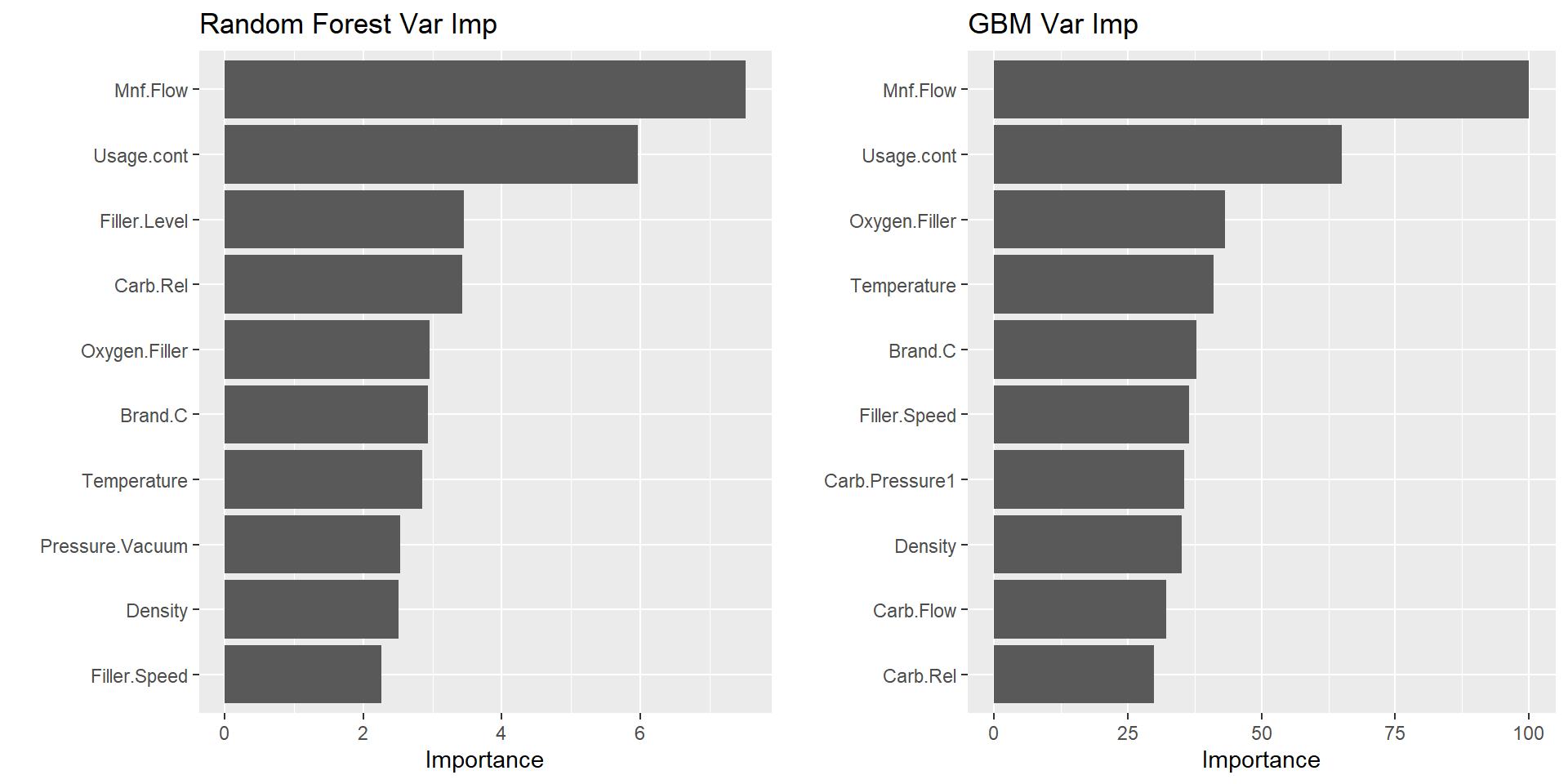


Figure 3: Random Forest Importance

**Model Evaluation**

All the models are evaluated and all of the results are examined to see which one gives the lowest RMSE and the highest RSquared. The optimal model selected was the Random Forest The optimal model is then evaluated against the evaluation data. The same process as the first dataset must be done to evaluation dataset. The data is tested against the optimal model Random Forest in order to obtain the predicted pH values. The predicted values obtained are then converted into a .csv and we are able to see the predicted values.

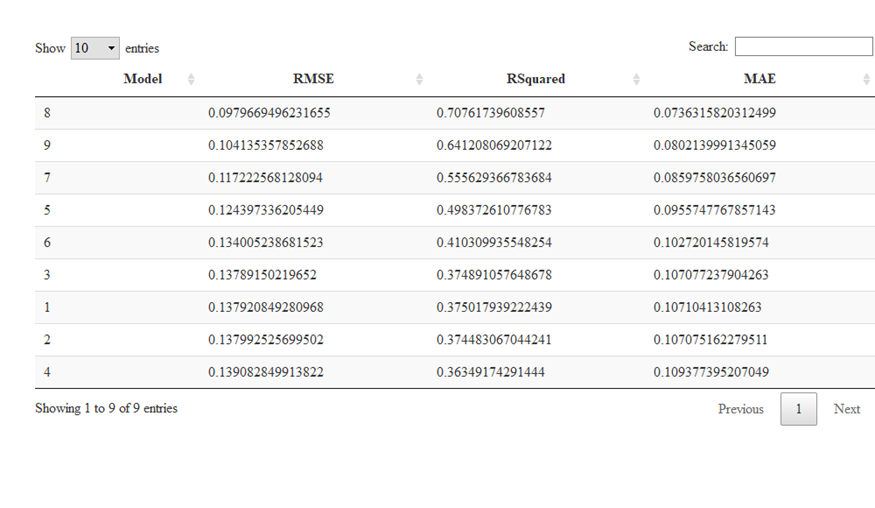


Figure 4: Evaluation of Models

**Prediction Result**

In the model evaluation the optimal model chosen was the Random Forest. With the optimal model predictions were obtained. Looking at Figure 5, you can see that the pH Levels appear to be alkaline since that are all above 8. Based on that we can clearly state that the manufacturing process of ABC Beverage is currently producing alkaline beverages.

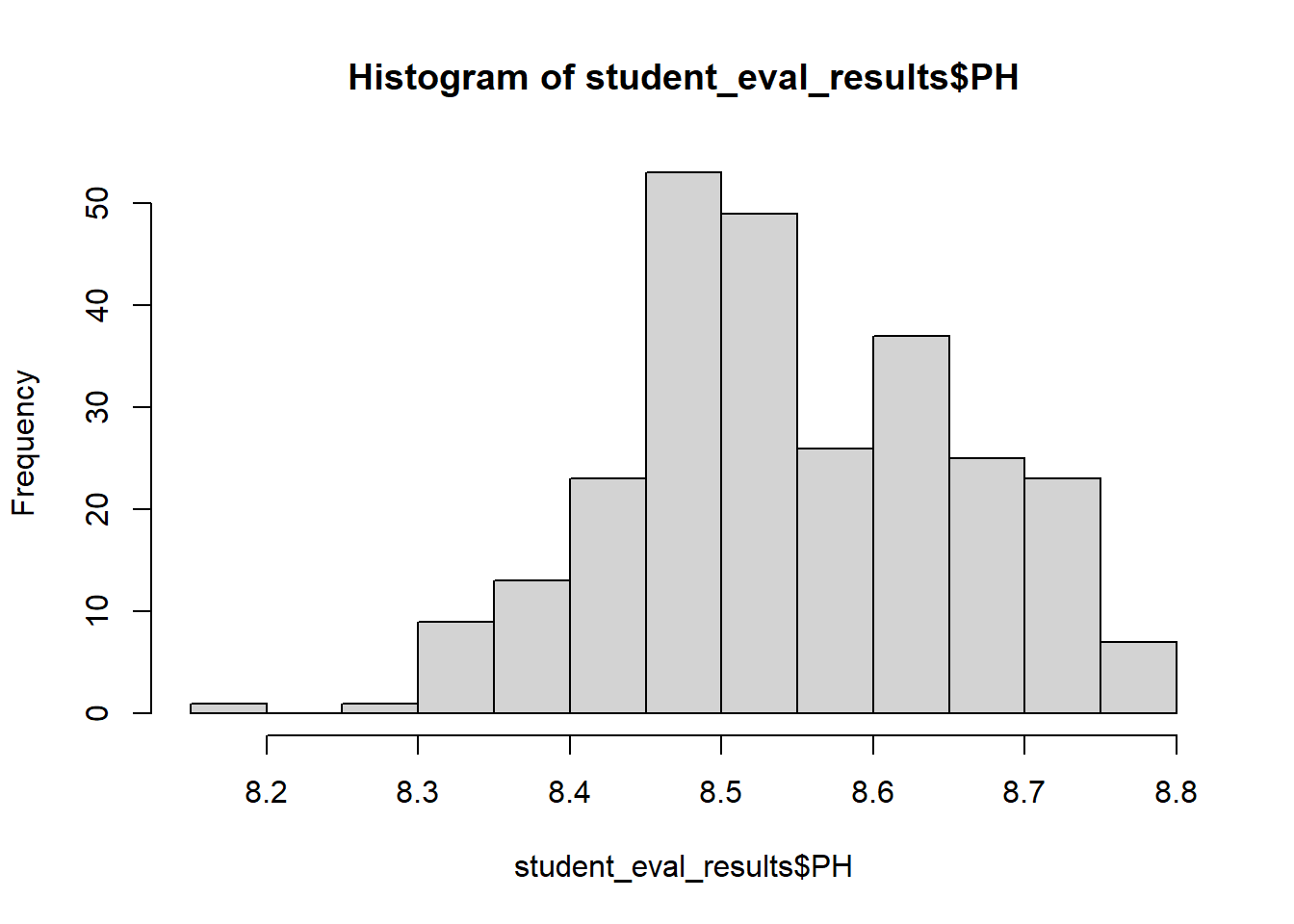


Figure 5: pH Level Predictions