|  |  |
| --- | --- |
| Predictive models of ph  Data Science Department, ABC Beverage | Abstract  Impact of PH levels in our beverages with the manufacturing process  Abdelmalek Hajjam, Monu Chacko, Paul Perez  DATA 624 |

Contents

[Overview 2](#_Toc72663136)

[Modelling 2](#_Toc72663137)

[High Impact Variables 3](#_Toc72663138)

[Conclusion 4](#_Toc72663139)

Predictive model of PH

# Overview

In this report we highlight the impact of PH levels in our beverages with the manufacturing process. We use the past manufacturing data to predict the future and stay ahead in our market segment.

PH levels historically vary from 7.88 and 9.36. It has a mean value of 8.55. In order to determine the best results we use the industry standard indicators – Root Mean Square Error (RMSE) and R-Squared values.

# Modelling

We used various modelling techniques and evaluated their performance. We used models like - Generalized Linear model (GLM), Partial Least Squares model (PSL), Ridge model, ENET model, MARS Model, SVM Model, KNN Model, Random Forest, Boosted trees, Cubist and Bagged Trees.

We used historical data and transformed them for a meaningful analysis. The data was then evaluated using all the models mentioned above to evaluate the best performance. We used *Root Mean Square Error (RMSE)* and *R-Squared values* to compare the model performance.

We found that the tree models performed better than the linear and non-linear models. The best model was Cubist. Cubist is a rule-based model that is an extension of Quinlan's M5 model tree. Random Forest model also had very good performance compared to all the other models we tuned and evaluated, so it might be a valid alternative to Cubist if for any reason an alternative is needed. The Cubist R-Squared value is 0.61, meaning that the model explains 61% of variability in the data. RMSE for Cubist is 0.11. The Random Forest R-Squared value is 0.60, meaning that the model explains 60% of variability in the data. RMSE for Random Forest is 0.11.

# High Impact Variables

Using the *Cubist* model we found the following variables to have highest impact on PH levels.



Following are the predictions from our analysis.



# Conclusion

We evaluated many linear, non-linear, tree based models using the historical data and found Cubist to be most effective. This is our initial findings based on the data we had so far. In the process or creating the model we created a system to constantly evaluate and based on the latest data available. This will make this system fine tune and make more accurate recommendations in the future.