Model of machine Learning to predict wine quality

José

Project Overview

- With a wine quality dataset based in wine chemical characteristics, try and choose a machine learning model to predict wine quality based on this characteristics.

- The characteristics used were: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.

Data Selection and Cleaning

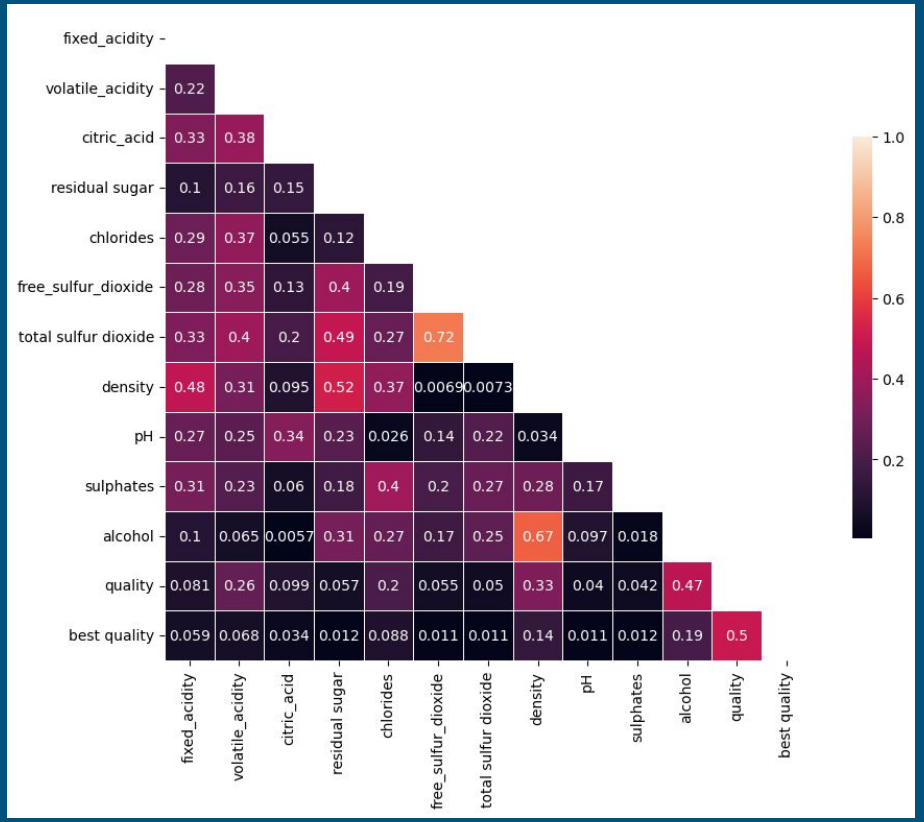
Kaggle “Wine Quality Dataset” with informations about red and white portuguese wines.

- Leading with null values
- Leading with duplicate values

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Feature engineering

- Normalization of the different values of the dataset
- Create a new column were bins were created to reshape the target
- Drop the columns with low correlation with the target



Model Building and Evaluation

1	KNN	• RMSE 0.4129534723729141 • R2 score 0.9709193245778611 • train score 0.9643276226237972
2	Linear Regression	• RMSE 0.16739002551896615 • R2 score 0.023461531529833213 • train score 0.04659232127157886
3	Logistic Regression	• RMSE 0.4129534723729141 • R2 score 0.9709193245778611 • train score 0.9640929359305327
4	Decision Tree	• RMSE 0.40958214162080514 • R2 score 0.9718574108818011 • train score 0.9652663693968552
5	Random Forest	• RMSE 0.4129534723729141 • R2 score 0.9709193245778611 • train score 1
6	Gradient Boosting	• RMSE 0.4699606921325369 • R2 score 0.9512195121951219 • train score 1

Hyperparameter Tuning and Model Optimization

For the model Optimization

- Decision Tree with:
- AdaBoost
- AdaBoost with bagging made the model perform better
- AdaBoost combined with grid or random search provided worts results

Streamlit

Bacalhoa

Álcohol: 14,5
Fixed Acidity: 6,5
Volatile acidity: 0.34
Citric acid: 0.32
Residual sugar: 1.3
Chlorides: 0.056
Density: 0.99
SO2: 0.53