Final Project Report

Prediction of Wine Quality Using Classification

Narges Goudarzi, Soudeh Nilforoushan, Anurag Bhattacharjee

ngoudar@uwo.ca, snilforo@uwo.ca, abhat23@uwo.ca

Western University · Canada

Department of Computer Science

Winter 2021

1. Abstract:

Data analysis is an important part of any business. Data analysis organises, interprets, structures and presents the data into useful information that provides context for the data. In this project we have applied 10 machine learning models on Pinot Noir wine samples to analyse objective chemical measurements and subjective sensory evaluation of the wine samples. We have tried to show the relationship between the chemicals and their impact on the quality of wine. In this experiment we have also compared between several machine learning models and how their performance varies on the same dataset. To bring more severity to the project we have also analysed all the machine learning models' performance for the same dataset with imbalance class and balanced class by oversampling the data.

2. Introduction:

Kaggle is a platform for varieties of datasets. We have picked our dataset from kaggle after analysing various datasets. We have picked the Red Wine Quality dataset because there are various factors to experiment from a diverse angle.

Analysing data is an important part of any business and understanding data is an important part for a data engineer. By understanding data and making the idea of which machine learning model to apply can give an overall insight of the data and can dictate important decisions in a business.

In this project we have experimented with 10 classifier machine learning models to predict the quality of wine from their ingredients. We have also analysed how the performance of models change along with the changes in the dataset. We experimented all these by operating several operations on the dataset like applying models on the data sets by oversampling the data or comparing the models on the same dataset with different hyper parameters. This project has given much insight on the knowledge required for analysing a dataset.

3. Exploratory Data Analysis:

In this study, we used Kaggle's Red Wine Quality dataset to build a variety of classification techniques to predict if a particular red wine is "good quality" or "bad quality. Each row in this dataset is given a "quality" score between 0 and 10. We decided to convert the "quality" column to a binary output so that each wine is either "good quality" (a score of 7 or higher) or "bad quality" (a score below 7). The quality of each wine in the dataset is defined by 11 predictors including: Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide Total sulfur dioxide, Density, pH, Sulfates, Alcohol. [1]

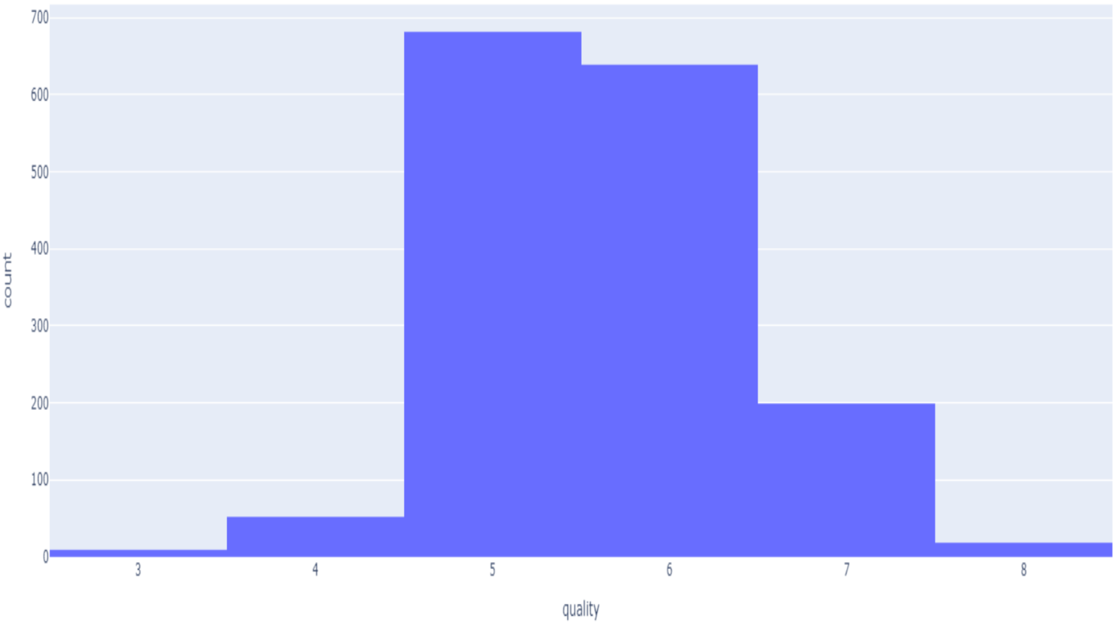


Fig. 1: Shows the distribution of the quality variable, which follows normal distribution.

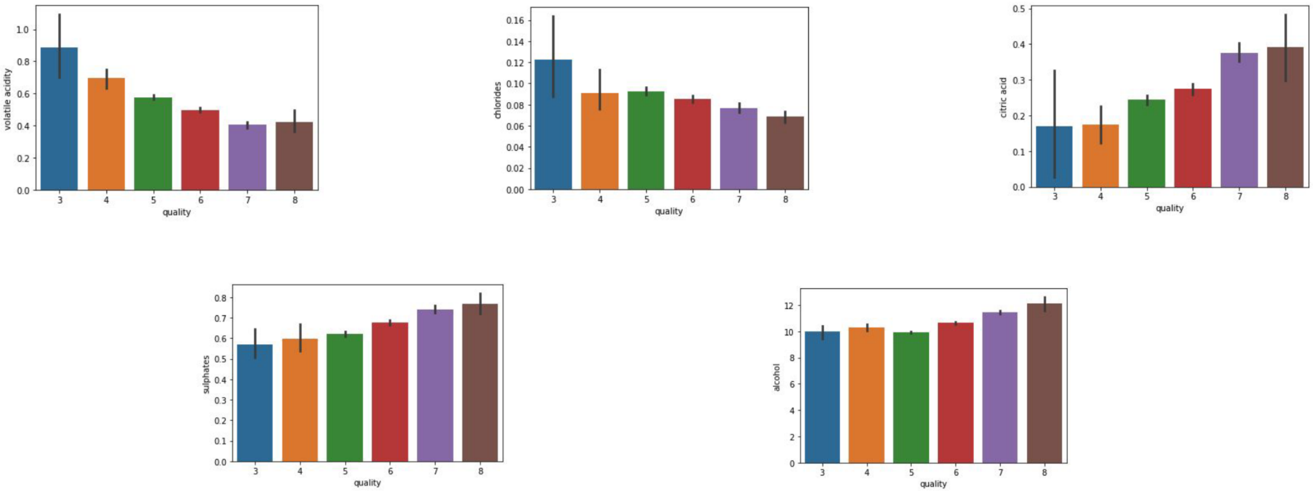
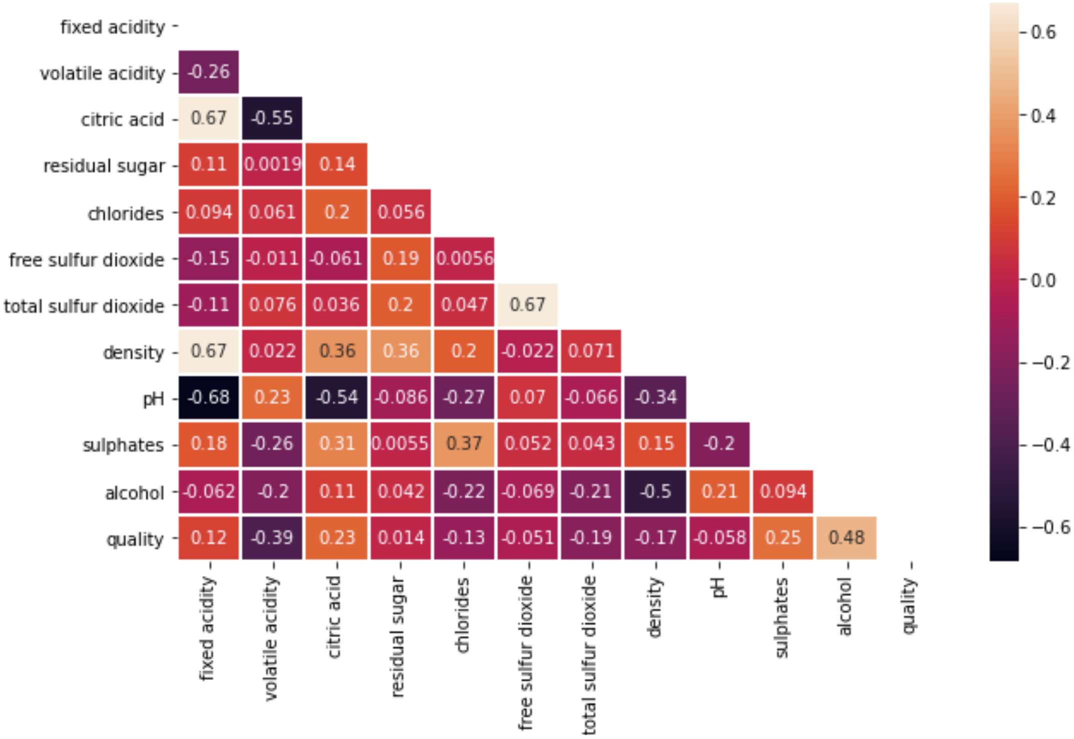


Fig. 2: Illustrates the correlations between the predictors, and also correlation between each predictor with the response variable. It will allow us to get a deeper understanding of the relationships of the variables in a glimpse.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity												
volatile acidity	-0.26											
citric acid	0.67	-0.55										
residual sugar	0.11	0.0019	0.14									
chlorides	0.094	0.061	0.2	0.056								
free sulfur dioxide	-0.15	-0.011	-0.061	0.19	0.0056							
total sulfur dioxide	-0.11	0.076	0.036	0.2	0.047	0.67						
density	0.67	0.022	0.36	0.36	0.2	-0.022	0.071					
pH	-0.68	0.23	-0.54	-0.086	-0.27	0.07	-0.066	-0.34				
sulphates	0.18	-0.26	0.31	0.0055	0.37	0.052	0.043	0.15	-0.2			
alcohol	-0.062	-0.2	0.11	0.042	-0.22	-0.069	-0.21	-0.5	0.21	0.094		
quality	0.12	-0.39	0.23	0.014	-0.13	-0.051	-0.19	-0.17	-0.058	0.25	0.48	

Fig. 3: Correlation Matrix.

We can clearly observe that there are some variables which are strongly correlated to the output. It's quite probable that these variables are the most important predictors in our machine learning model, which we will discuss later. For instance, alcohol is strongly correlated with the quality of wine. Residual sugar, and free sulfur dioxide seem to be the least important. Furthermore, Citric acid, pH, and fixed acidity have a strong correlation to each other, which is reasonable. In addition, some features like chloride, and volatile acidity are negatively correlated to the output, which shows that they are unpleasant features for the red wine.

Oversampling:

After replacing the quality column of our dataset with the CLASS column we have seen that the dataset isn't balanced. There are 1382 data samples which are labelled as bad quality wine but only 217 data samples which are qualified as good quality wine. So the dataset is imbalanced. As we are performing several models on this dataset to compare accuracy among the models we also used MLP classifier. But neural networks are sensitive to data. So, we planned to oversample the data.

Oversampling and undersampling in data analysis are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes/categories represented).

These terms are used both in statistical sampling, survey design methodology and in machine learning. We oversampled our data for CLASS label 1 and performed all the models on the oversampled data. But we also tested all the models on the imbalanced dataset as well. It helped us to compare how the performance of each model varies along with the data.

Splitting the data:

We split the data into training and test sets so that we could determine the effectiveness of our models. We set the test size to 0.3.

4. Methodology:

A variety of supervised learning techniques (Logistic Regression, Random Forest, KNN, Neural Network) were used to train classifiers on the training data.

4.1 Logistic regression:

A logistic regression is the best tool to handle categorical dependent variables. Machine Learning experts and data scientists use this method as the baseline model. However, we need to verify the technical requirements to estimate the logistic regression. For applying this technique, it is crucial to observe the presence of outliers, the occurrence of high correlation between independent variables, and also a reasonable sample size. We did not find extreme cases which are capable of harming the model's fit, and the correlation between the variables of this dataset is acceptable. For logistic regression, the size of the dataset is of crucial importance. Small dataset may produce inconsistent estimates.

On the other hand, too large datasets increase the power of statistical tests in a way that any effect can be statistically significant. Hosmer et al. (2000) suggested a minimum of 400 observations. Pedhazur (1982) recommended a ratio of 30 cases for each estimated parameter. So we can conclude that for our dataset with 1599 observations and 11 predictors this method can be a good choice. As a result, we can safely move ahead to the next step of implementation.

We chose to apply logistic regression both without penalty and with L1,L2 regularization to see whether regularization improves the performance and can reduce overfitting. They may also help us for best feature selection. When we apply L1 regularization (Lasoo), Less important predictors become zero. When we apply L2 regularization (Ridge) then coefficient shrinkage but not necessarily to zero. I did not tune the hyperparameter C for the regularization strength and just considered the default value[2,3].

4.2. Support Vector Machine(SVM)

Support Vector Machine emerged in the 1990s and it is based on convex quadratic programming[4]. It is a supervised type of Machine Learning. It has application in many classification and regression problems such as image classification, face detection and speech recognition[5]. SVM transfers the data and finds the optimal boundary between similar outputs, then it understands how to split the data based on the output of transfer. SVM has certain input hyper parameters and for all parameters we use the default, only for “kernel” we assign “linear”.

4.3. K-nearest neighbor (KNN)

K Nearest Neighbor classifier has been listed in the top 10 algorithms in machine learning [6]. KNN can be used for supervised and it can be used for regression and classification. KNN works by finding the distance between a query and all examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label. SVM has certain input hyper parameters and for all parameters we use the default, only for “n_neigbour” we assign 2.

4.4. Decision Tree (DT)

The basic idea involved in any multistage approach is to break up a complex decision into a union of several simpler decisions, hoping the final solution obtained this way would resemble the intended desired solution[7]. Decision Tree can be used for classification and prediction problems. Decision trees are simple and easy to interpret and it is valuable without requiring large amounts of hard data. Decision tree has certain input hyper parameters and for all parameters we use the default, only for “max_depth” we assign 10.

4.5. Random Forest

Random Forest is supervised machine learning that is used for classification and prediction problems, but it performs better for classification problems. Random Forest is an ensemble of decision tree algorithms. It is also an extension of bootstrap aggregation (bagging) of decision trees. Random forest makes a decision tree for each sample and returns the majority result based on the decision tree for classification and also returns the mean for the prediction problem. Random Forest has certain input hyper parameters and for all parameters we use the default, only for “n_estimators” we assign 200.

4.6. Bagging(BootStrap)

In bagging, a number of decision trees are created which each tree is created from a different bootstrap sample of the training datasets. A "bootstrap sample" is a sample of the training datasets where a sample may appear more than once in the sample, referred to as sampling with replacement. Bagging is an effective ensemble algorithm as each decision tree fits on a slightly different training dataset, and in turn, has a slightly different performance. Predictions from the trees are averaged across all decision trees resulting in better performance than any single tree in the model. Bagging has certain input hyper parameters and for all parameters we use the default, only for “random_state” we assign 1.

4.7.Stochastic Gradient Descent(SGD)

Gradient Descent is the basic form of neural network. It can be used for classification and prediction. It is often used in natural language processing. SGD is an optimization algorithm used to find the values of parameters of function that minimizes a cost function. The difference between SGD and GD comes while iterating. In Gradient descent we consider all the points in calculating the loss while in gradient descent we use a single point in loss function and its derivative randomly. SGD has certain input hyper parameters and for all parameters we use the default, only for “loss” we assign “log” and for “max_iter” we assign 1000.

4.8. MLP

The multilayer perceptron or MLP for short is a classical Neural Network. There can be multiple layers of neurons and it is a feedforward artificial neural network. These one or more hidden layers create an abstraction and the visible layer or the output layer is used to get the prediction. These layers are fully connected to the following one. The neuron, except for the input layer, consists of a nonlinear activation function. Between the input and out layer there can also be nonlinear activation functions.

MLP has certain hyper parameters

hidden_layer_sizes : It is declared as a tuple with the number of nodes in the ith position and each element representing the number of nodes. We have 1 hidden layer with 3 nodes in our model.

max_iter: the number of epochs. We ran our model for 1 epoch.

solver: weight optimization over the nodes. We used “adam” as solver.

activation: The activation function for the hidden layers. In our model we have logistic or sigmoid activation.

5. Results

5.1 Performance Comparison:

Table1. Performance comparison with oversampled data

Methods	accuracy	recall	precision	F1-Score	sensitivity	specificity
LR(No penalty)	0.81	0.86	0.77	0.81	0.86	0.76
LR(Ridge)	0.81	0.86	0.78	0.82	0.86	0.77
LR(Lasso)	0.81	0.85	0.77	0.81	0.85	0.76
SVM	0.8	0.88	0.76	0.81	0.88	0.74
KNN	0.94	0.99	0.90	0.94	0.99	0.89
Decision Tree	0.93	0.99	0.88	0.93	0.99	0.88
Random Forest	0.97	0.98	0.95	0.97	0.98	0.95
Bagging	0.95	0.98	0.93	0.95	0.98	0.93
SGD	0.96	0.98	0.95	0.96	0.98	0.95
MLP	0.767	0.774	0.754	0.764	0.774	0.761

Table 2. Performance comparison without Sampling

Methods	Accuracy	Recall	Precision	F1-Score	Sensitivity	Specificity
LR(No penalty)	0.89	0.36	0.46	0.40	0.36	0.95
LR(Ridge)	0.89	0.34	0.51	0.41	0.34	0.96
LR(Lasso)	0.90	0.36	0.54	0.43	0.36	0.96
SVM	0.89	0	0	0	0	0.11
KNN	0.89	0.18	0.45	0.25	0.18	0.97
Decision Tree	0.88	0.74	0.46	0.57	0.74	0.90
Random Forest	0.91	0.54	0.58	0.56	0.54	0.95
Bagging	0.89	0.42	0.48	0.45	0.42	0.94
SGD	0.92	0.56	0.65	0.6	0.56	0.96
MLP	0.90	0.90	0.80	0.85	-	-

5.2. AUC Comparison

Table 3. AUC comparison with oversampled data

Methods	**AUC Values**
LR(No penalty)	0.86
LR(Ridge)	0.86
LR(Lasso)	0.86
SVM	0.88
KNN	0.94
Decision Tree	0.93
Random Forest	0.99
Bagging	0.98
SGD	0.99
MLP	0.87

Table 4. AUC comparison without sampling data

Methods	**AUC values**
LR(No penalty)	0.857
LR(Ridge)	0.865
LR(Lasso)	0.853
SVM	0.84
KNN	0.76
Decision Tree	0.83
Random Forest	0.92
Bagging	0.88
SGD	0.91
MLP	0.84

5.3 AUC diagram

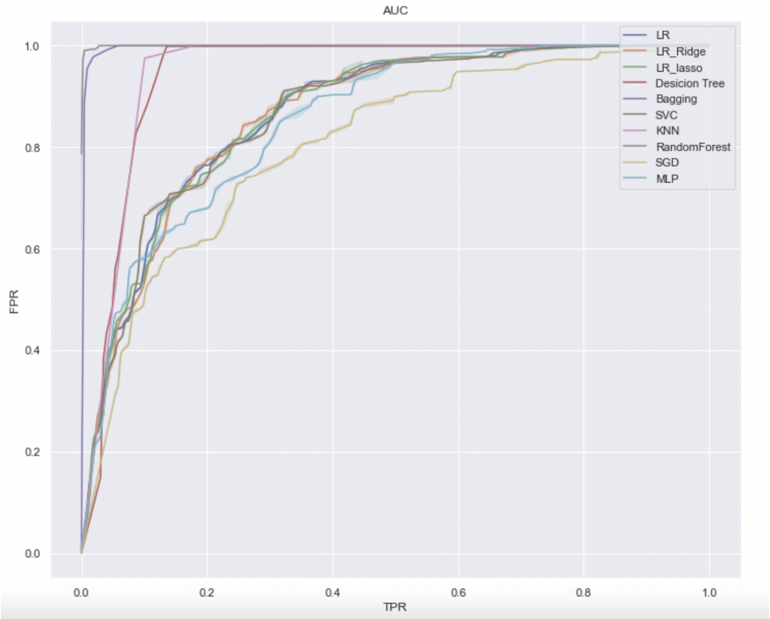


Fig4: AUC comparison without sampling data

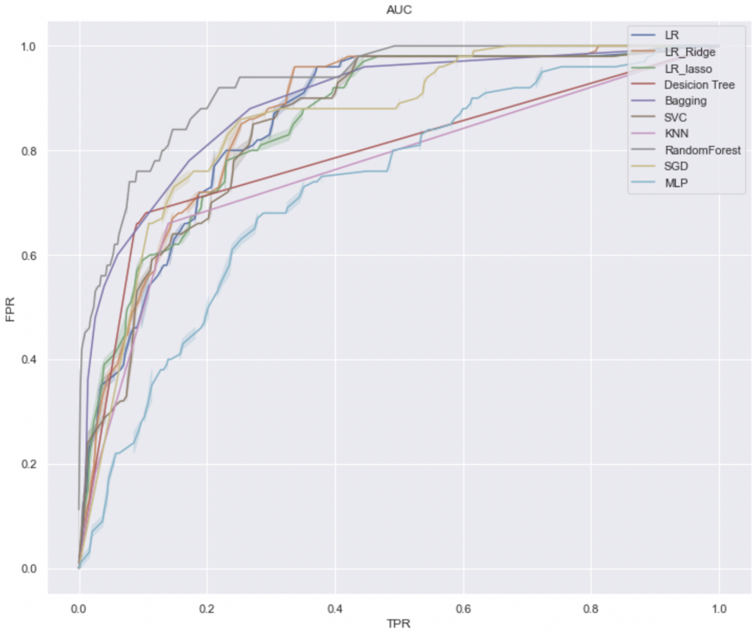


Fig 5: AUC comparison without sampling data

5.4. Feature Importance

Below, we graphed the feature importance based on the random forest and stochastic gradient descent model. While they vary, the top 3 features are the same: alcohol, volatile acidity and sulphates.

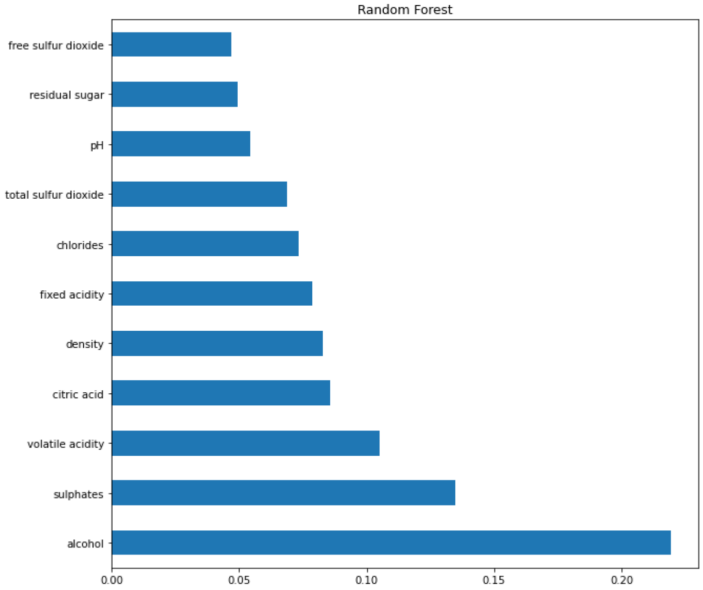


Fig 6. Random Forest Feature Importance.

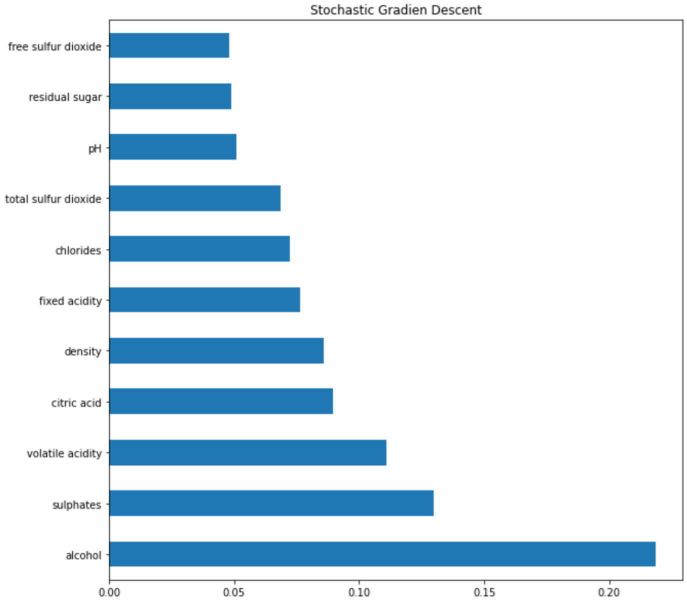
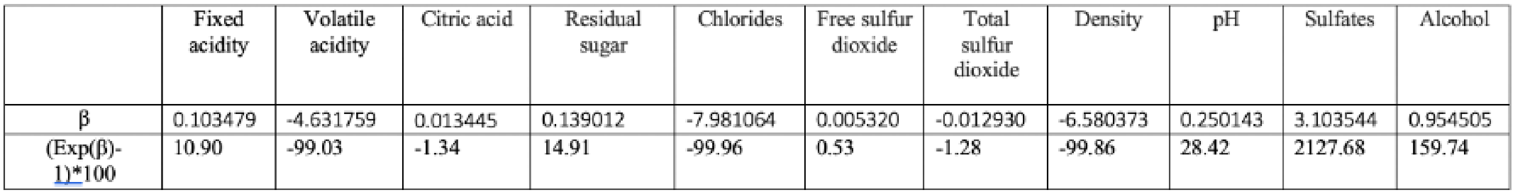


Fig 7. Stochastic Gradient Descent Feature Importance.

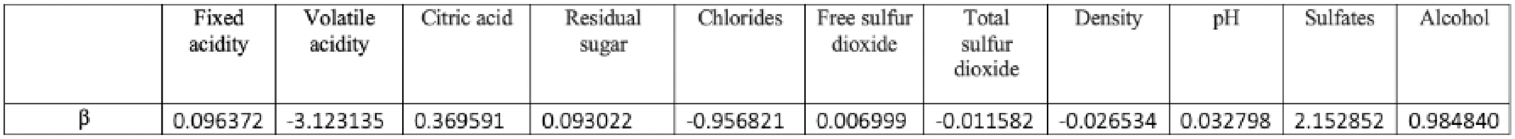
For logistic regression there are two major approaches for analysing the coefficients: 1) analyze the odds ratio and b) turn the odds ratio into a percentage. With the former, as an example we can say that higher volatile acid content reduces the chances of being elected as good quality. In terms of percentages, high values of this feature diminish in 99.03% the probability of being elected as good quality, as theoretically expected. We can also clearly see that based on the three models, alcohol content is the most important feature., while free sulfur dioxide and Ph are the less important ones. Ridge regression shrinkage the coefficients while Lasso could successfully eliminate Citric acid feature.

Table 5. Analysis of the coefficients for model 1 (No penalty)



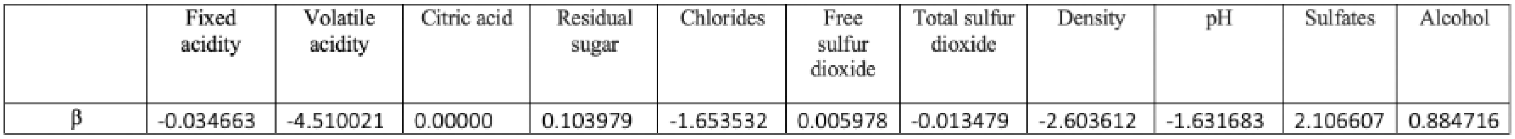
	Fixed acidity	Volatile acidity	Citric acid	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	pH	Sulfates	Alcohol
β	0.103479	-4.631759	0.013445	0.139012	-7.981064	0.005320	-0.012930	-6.580373	0.250143	3.103544	0.954505
(Exp(β)-1)*100	10.90	-99.03	-1.34	14.91	-99.96	0.53	-1.28	-99.86	28.42	2127.68	159.74

Table 6. Analysis of the coefficients for model 2 (Ridge)



	Fixed acidity	Volatile acidity	Citric acid	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	pH	Sulfates	Alcohol
β	0.096372	-3.123135	0.369591	0.093022	-0.956821	0.006999	-0.011582	-0.026534	0.032798	2.152852	0.984840

Table 7. Analysis of the coefficients for model 3 (Lasso)



	Fixed acidity	Volatile acidity	Citric acid	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	pH	Sulfates	Alcohol
β	-0.034663	-4.510021	0.00000	0.103979	-1.653532	0.005978	-0.013479	-2.603612	-1.631683	2.106607	0.884716

6. Conclusion

In this project we have projected and investigated various machine learning models and their performance on the same dataset. This is a novel method to compare among several models and to give a sense of which model performs better in a specific situation. Experimental results show that for this dataset Logistic Regression outperformed all other models. We have also observed that the performance of the models didn’t improve that much after performing them on oversampled data. This gives us an important insight on understanding data and about the machine learning models.

7. Future Work

We have experimented with diverse aspects of data analysis in this project. For the future we want to extend this project by experimenting the same sort of machine learning on different datasets and compare their performance. In that way we will be able to compare the results between each model and also acquire knowledge on the variability and usability of models for different datasets and situations.

8. References

1. **Kaggle Dataset, "Red Wine Quality, Simple and clean practice dataset for regression or classification modelling".**
2. A.A.T. Fernandes, D.B.F. Filho, E.C.d. Rocha, W.d.S. Nascimento, "Read this paper if you want to learn logistic regression" , Rev. Sociol. Polit. 28 (74), 2020.
3. L.E. Melkumovaa, S.Ya. Shatskikhb, "Comparing Ridge and LASSO estimators for data analysis ", Procedia Engineering, Volume 201, Pages 746-755, 2017.
4. G. Wang. "A survey on training algorithms for support vector machine classifiers". Fourth International Conference on Networked Computing and Advanced Information Management, volume 1, pages 123–128, 2008.
5. A. Pradhan. "Support vector machine-a survey". International Journal of Emerging Technology and Advanced Engineering, 2(8):82–85, 2012.
6. L. Jiang, Z. Cai, D.Wang, and S. Jiang." Survey of improving k-nearest-neighbor for classification". 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007),volume 1, pages 679–683, 2007.
7. S.R. Safavian and D. Landgrebe. "A survey of decision tree classifier methodology", IEEE Transactions on Systems, Man, and Cybernetics, 21(3):660–674, 1991.
8. A. Botalb, M. Moinuddin, U. M. Al-Saggaf, S. S. A. Ali, "Contrasting Convolutional Neural Network (CNN) with Multi-Layer Perceptron (MLP) for Big Data Analysis". 2018 International Conference on Intelligent and Advanced System (ICIAS).