Introduction:

In this investigation, we aim to evaluate and understand two separate datasets, detect any possible issues or challenges in implementing machine learning, and examine the effects of data preprocessing techniques. The objective is to enhance data quality and model performance by applying preprocessing techniques, identifying any missing values, visualizing feature distributions, and gaining insights into the features of the data.

The first data set consists of the drilling data gathered from one of the oil and gas operators in the Eagleford shale region in Texas (USA). The collected data consists of 17 different parameters named as Depth, Hook Load, Bit Weight, Block Height, Rate of Penetration (ROP), Top Drive RPM, Top Drive Torque, Differential Pressure, Flow In Rate, Pump SPM (Strokes per minute), Flow Out Percent, Bit Size, Gamma Ray, Mud Weight In, Pump Pressure, D-exponent and formation. The aim of this dataset is to predict the formation by using other available parameters as an input. We will be using linear regression as our prediction model.

The second dataset consists of the data of the white variant of the Portuguese "Vinho Verde" wine. It comprises of 12 different parameters named as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality. The aim of this dataset is to predict the wine quality from other available input parameters. Random forest will be used as prediction model for this dataset.

Results:

The general observations derived from the datasets mentioned above are mentioned in this section.

Results from the Drilling data: In the very first step we used the print shape command and found that the dataset consists of 8551 rows and 17 columns. Following this we used the print information command to get the information of all data types present in the dataset. This shows that all the 17 columns have 8551 data points, and we have 12 float data types and 5 integer data types. After this we used the isnull command to find out the missing values in the dataset. We observed that there we no missing values in the dataset. After this we plotted the correlation plot to find out the most correlated features. This has also been depicted in figure 1. From the correlation plot firstly, we can see that Bitsize is constant throughout the dataset hence we can remove it from our study. Secondly, we can see that Top Drive RPM is having a very high correlation with Rate of Penetration (ROP) and Top Drive Torque hence we have removed ROP and torque from our study. Pump SPM and Flow in rate also have a very strong correlation hence we have also removed pump SPM from our study. Pump pressure and differential pressure also have a very high correlation; hence we have also removed pump pressure from our analysis. Bit weight and differential pressure also have a high correlation hence we have also removed bit weight from our analysis. Finally, we also find that depth is also having a very high correlation with formations hence we have also removed depth from our study. Now we have again plotted the correlation plot with non-correlated parameters and the same has been depicted in figure 2.

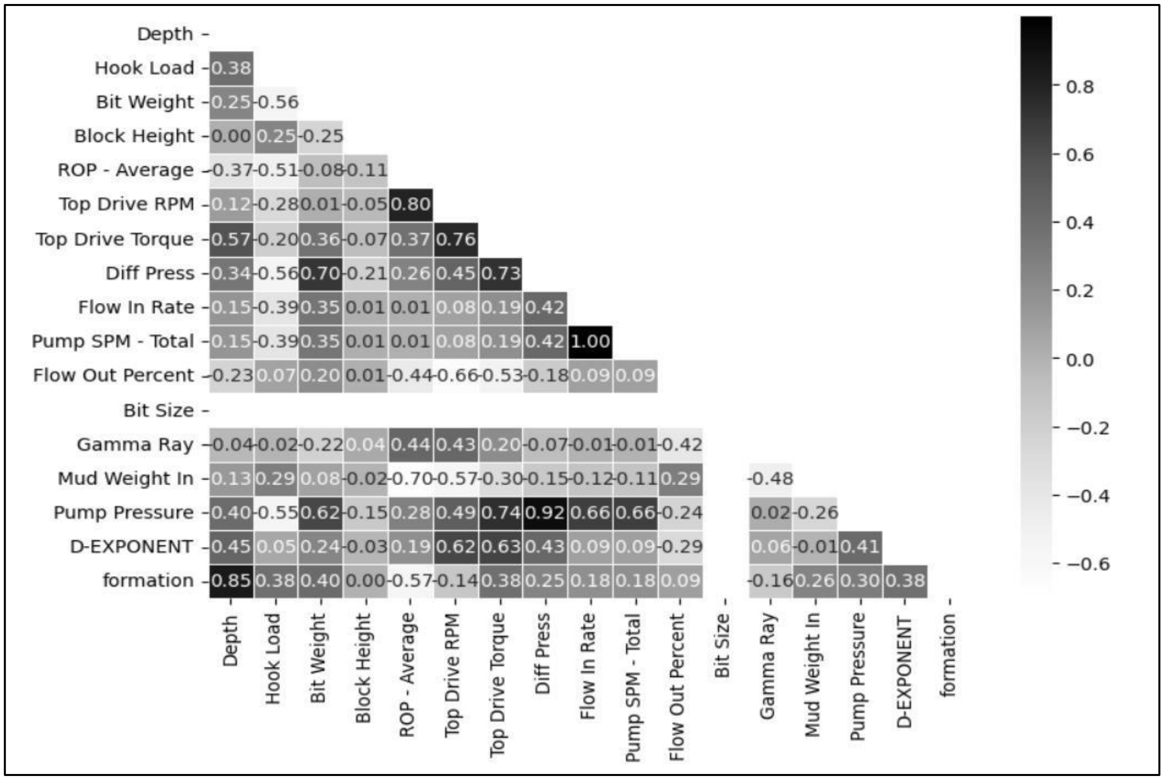


Figure 1. Correlation plot on initial drilling data

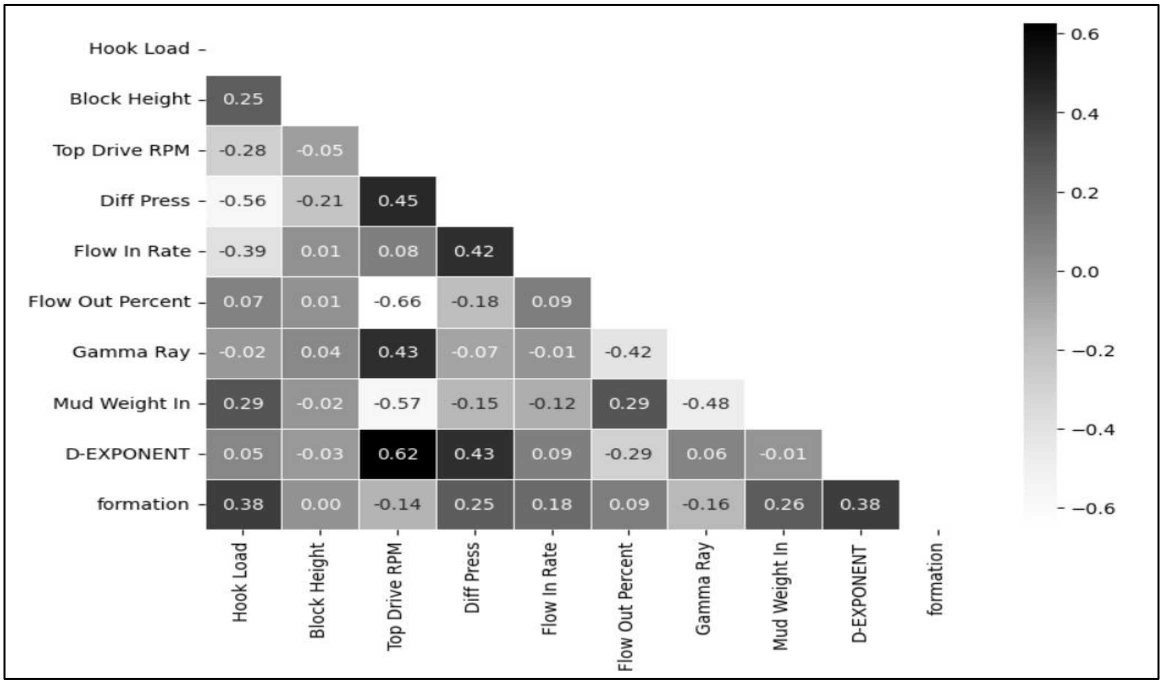


Figure 2. Correlation plot after removing correlated features in drilling data

We now used the describe command to find out the statistical features of the non-correlated features. This gives us a general idea of the range of values present under each feature. As we can see from the feature values there are some outliers present in the data hence, we plotted the box plot of each feature. We have also plotted the formation vs non-correlated feature to check how the features are varying throughout the different formations. The box plot and scatter plot has been given in appendix. We have now used z-score with a threshold of ± 3 standard deviations to remove the outliers. The new plots with removed outliers are given in appendix. As from the describe() command we can see that range of our data is varying hence there needs a necessity to normalize the data hence we are using MinMax scalar to normalize the data. After the data has been normalized, we made our prediction for classification of the formations based on input data using the random forest model. We considered it as a multi-class classification and began with importing the required modules like Random Forest classifier, Cross validation score and accuracy score from the sklearn library. After this we divided the data into input variables and target variable which is “formation” in our case. Now we have fitted the random forest model in the input and target variables. We have now performed K-fold cross-validation where K=5 for our dataset. The general working of K-fold cross validation can be explained as a single parameter called k that refers to the number of groups that a given data sample is to be split into. When a specific value for k is chosen, it may be used in place of k in reference to the model, such as K=5 becoming 5-fold cross-validation. If K=5 the dataset will be divided into 5 equal parts and the process mentioned next will run 5 times, each time with a different holdout set. Firstly, take the group as a holdout or test data set. Secondly, take the remaining groups as a training data set. Thirdly, fit a model on the training set and evaluate it on test set and finally retain the evaluation score and discard the model. At the end of the process summarize the skill of the model using the sample of model evaluation scores which is “accuracy score” in our case. As can be inferred from the process of cross validation we get 5 accuracy score values for each fold or each iteration and finally by taking a mean of these accuracy scores we get a mean accuracy value of 0.92.

Result from wine quality dataset: We begin with uploading the wine quality dataset for the white wine. After uploading the data, we found out that it cointains all the information in one column hence we are using separate function along with pd.read function to separate the data based on semicolon (i.e. “;”). Now we used the shape command to view the shape of data. The data comprises of 4898 rows and 12 columns. Following this we used the print information command to get the information of all data types present in the dataset. This shows that all the 12 columns have 4898 data points, and we have 11 float data types and 1 integer data type. After this we used the isnull command to find out the missing values in the dataset. We observed that there we no missing values in the dataset. After this we plotted the correlation plot to find out the most correlated features. This has also been depicted in figure 3.

From figure 3, density and residual sugar are having high correlation, hence we are removing density from our study. Similarly, total sulfur dioxide and free sulfur dioxide have a high correlation, so total sulfur dioxide is removed from the study. The updated parameters with non-correlative parameters are again plotted and can be seen in figure 4.

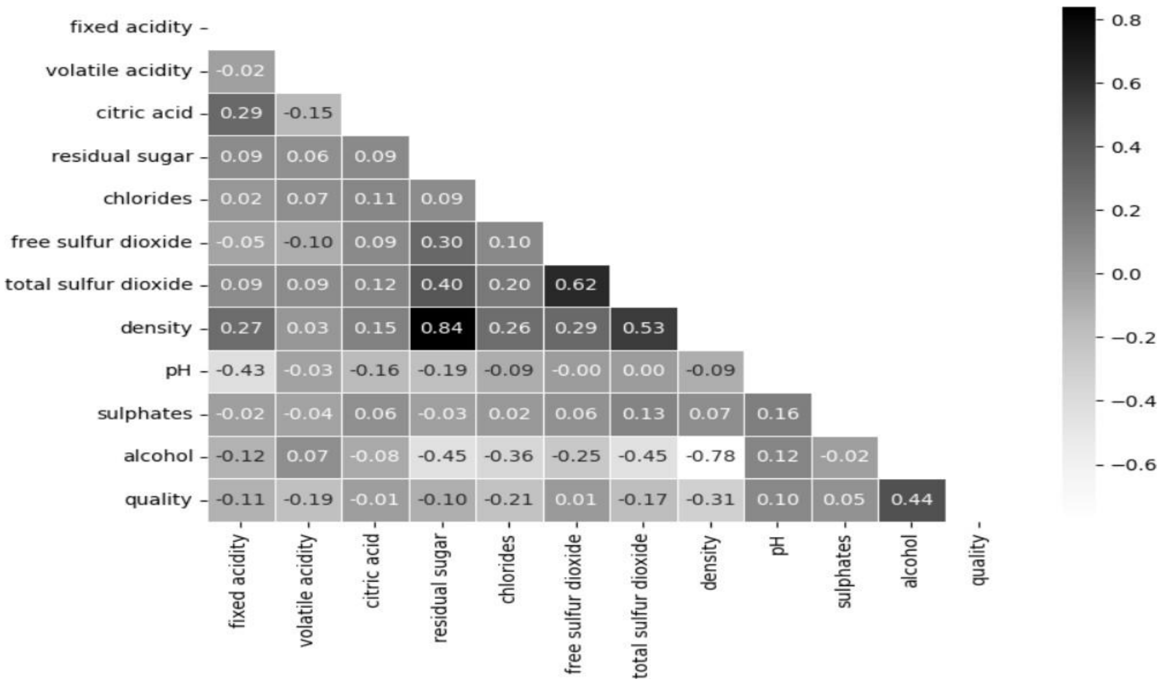


Figure 3. Correlation plot on initial wine quality data

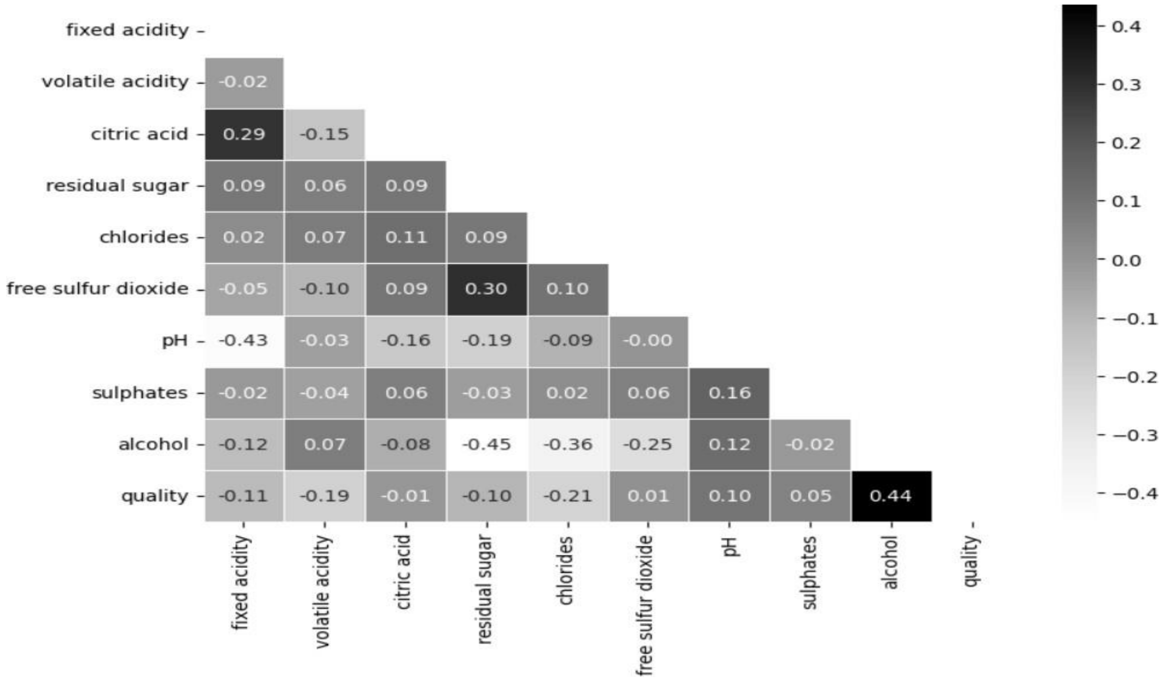


Figure 4. Correlation plot after removing correlated features in wine quality data

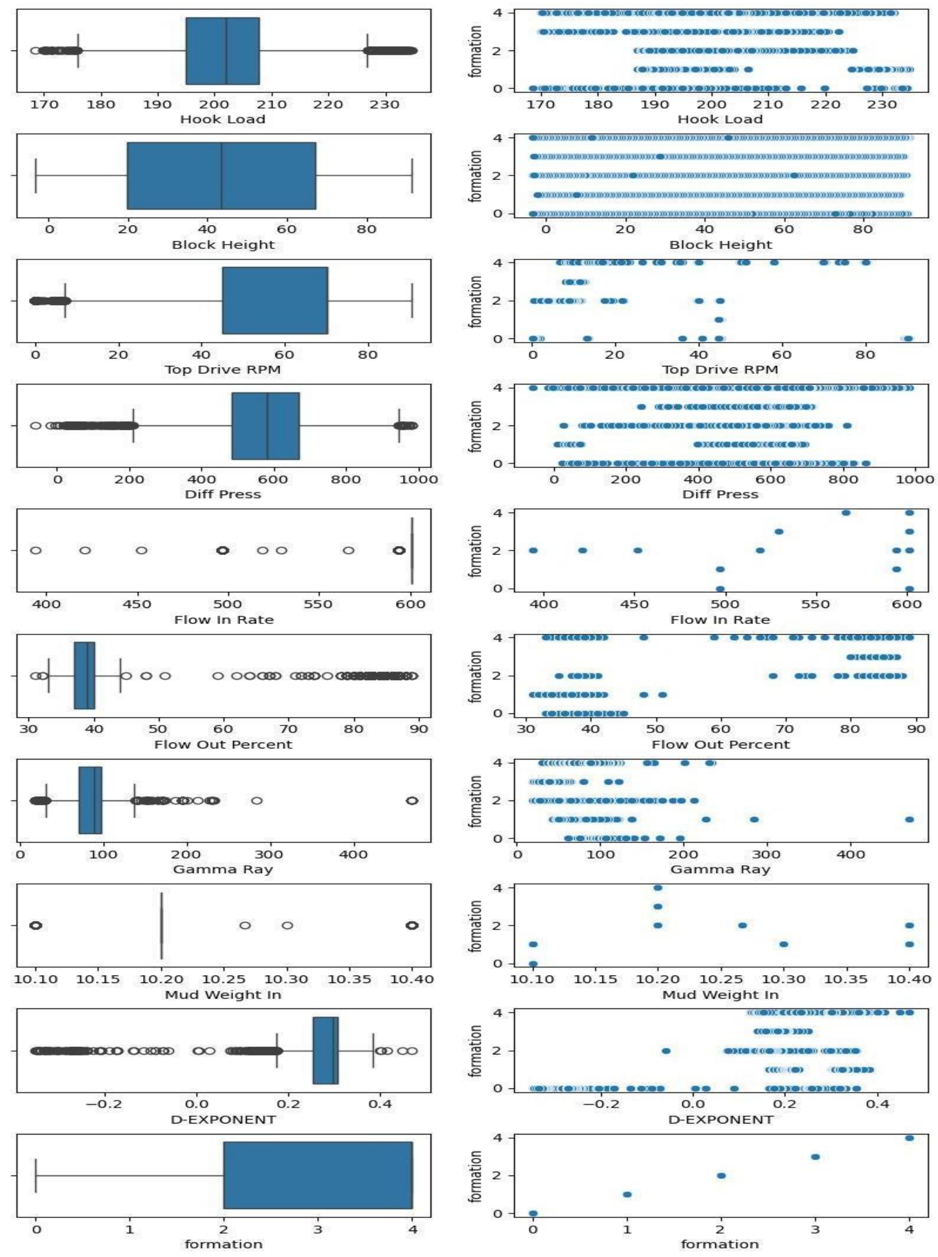
To determine the statistical characteristics of the non-correlated features, we now utilized the describe command. This provides us with a broad understanding of the range of values found within every feature. We generated a box plot for each feature because, as the feature values show, there are some outliers in the data. To examine how the features vary throughout the various wine quality, we have also plotted the quality against the non-correlated feature. The appendix contains the box plot and scatter plot. To exclude the outliers, we have now employed the z-score with a threshold of ± 3 standard deviations. The appendix contains the updated graphs with the outliers removed. We are using MinMax scalar to normalize the data because, as the describe() command indicates, the data's range is fluctuating, and we therefore need to normalize the data. After the data has been normalized, we made our prediction for classification of the quality based on input data using the random forest model. We considered it as a multi-class classification and began importing the required modules like Random Forest classifier, Cross validation score and accuracy score from the sklearn library. After this we divided the data into input variables and target variable which is "quality" in our case. Now we have fitted the random forest model in the input and target variables. We have now performed K-fold cross-validation where K=5 for our dataset. The general working of K-fold cross validation can be explained as a single parameter called k that refers to the number of groups that a given data sample is to be split into. When a specific value for k is chosen, it may be used in place of k in reference to the model, such as K=5 becoming 5-fold cross-validation. If K=5 the dataset will be divided into 5 equal parts and the process mentioned next will run 5 times, each time with a different holdout set. Firstly, take the group as a holdout or test data set. Secondly, take the remaining groups as a training data set. Thirdly, fit a model on the training set and evaluate it on test set and finally retain the evaluation score and discard the model. At the end of the process summarize the skill of the model using the sample of model evaluation scores which is "accuracy score" in our case. As can be inferred from the process of cross validation we get 5 accuracy score values for each fold or each iteration and finally by taking a mean of these accuracy scores we get a mean accuracy value of 0.52.

References:

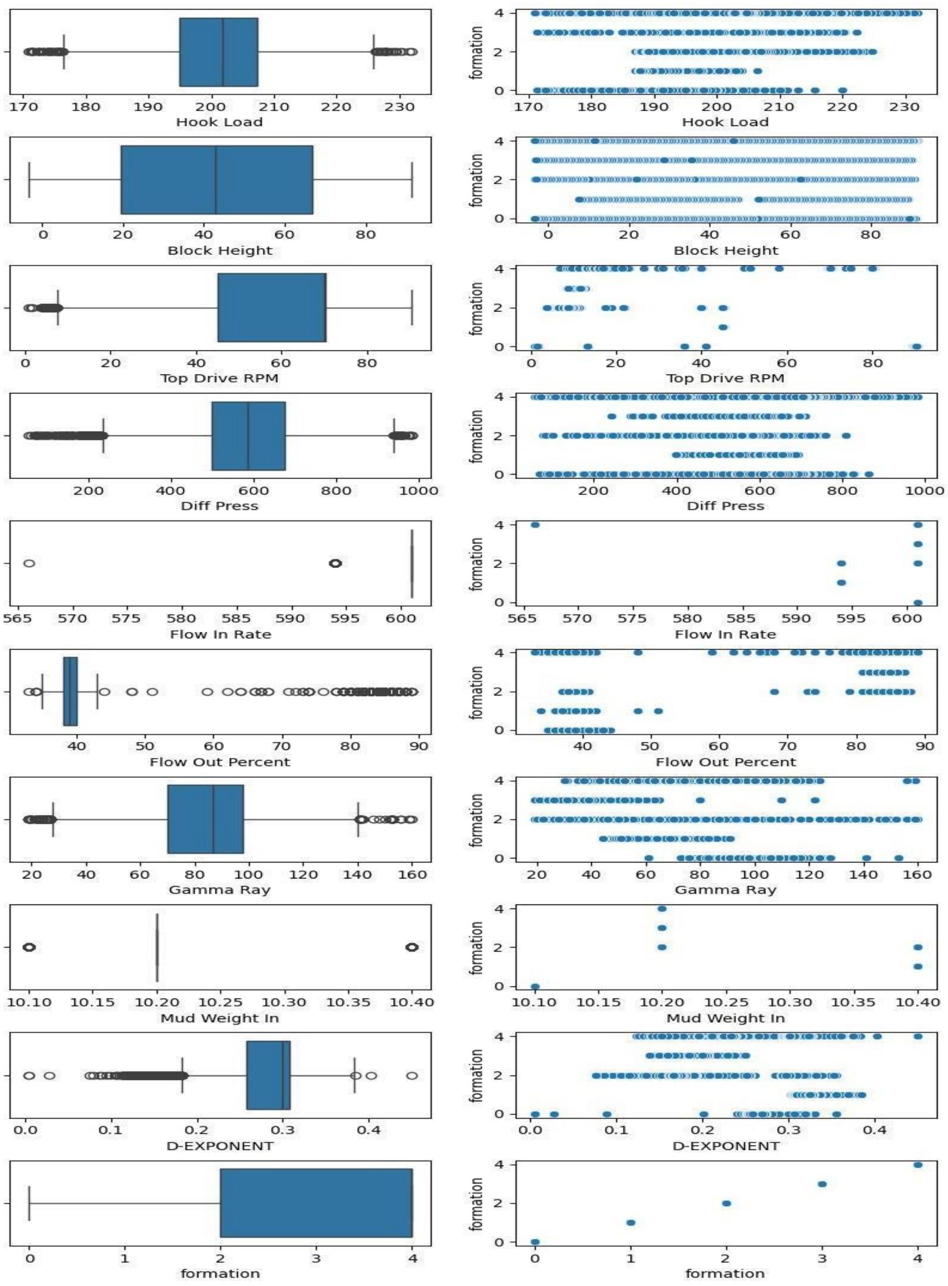
1) Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.
2) Agrawal, R., Malik, A., Samuel, R., and Saxena, A. (July 19, 2021). "Real-Time Prediction of Litho-Facies from Drilling Data Using an Artificial Neural Network: A Comparative Field Data Study with Optimizing Algorithms." ASME. *J. Energy Resour. Technol*. April 2022; 144(4): 043003. https://doi.org/10.1115/1.4051573
3) “Detect and Remove the Outliers using Python”, https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/
4) “How to use seaborn plotting”, https://www.geeksforgeeks.org/python-seaborn-tutorial/
5) “Normalizing the data by use of scaler”, https://www.geeksforgeeks.org/data-pre-processing-wit-sklearn-using-standard-and-minmax-scaler/
6) Z-score for normalization, https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55
7) Cross Validation Explained: Evaluating estimator performance. Improve your ML model using cross validation. https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

Appendix

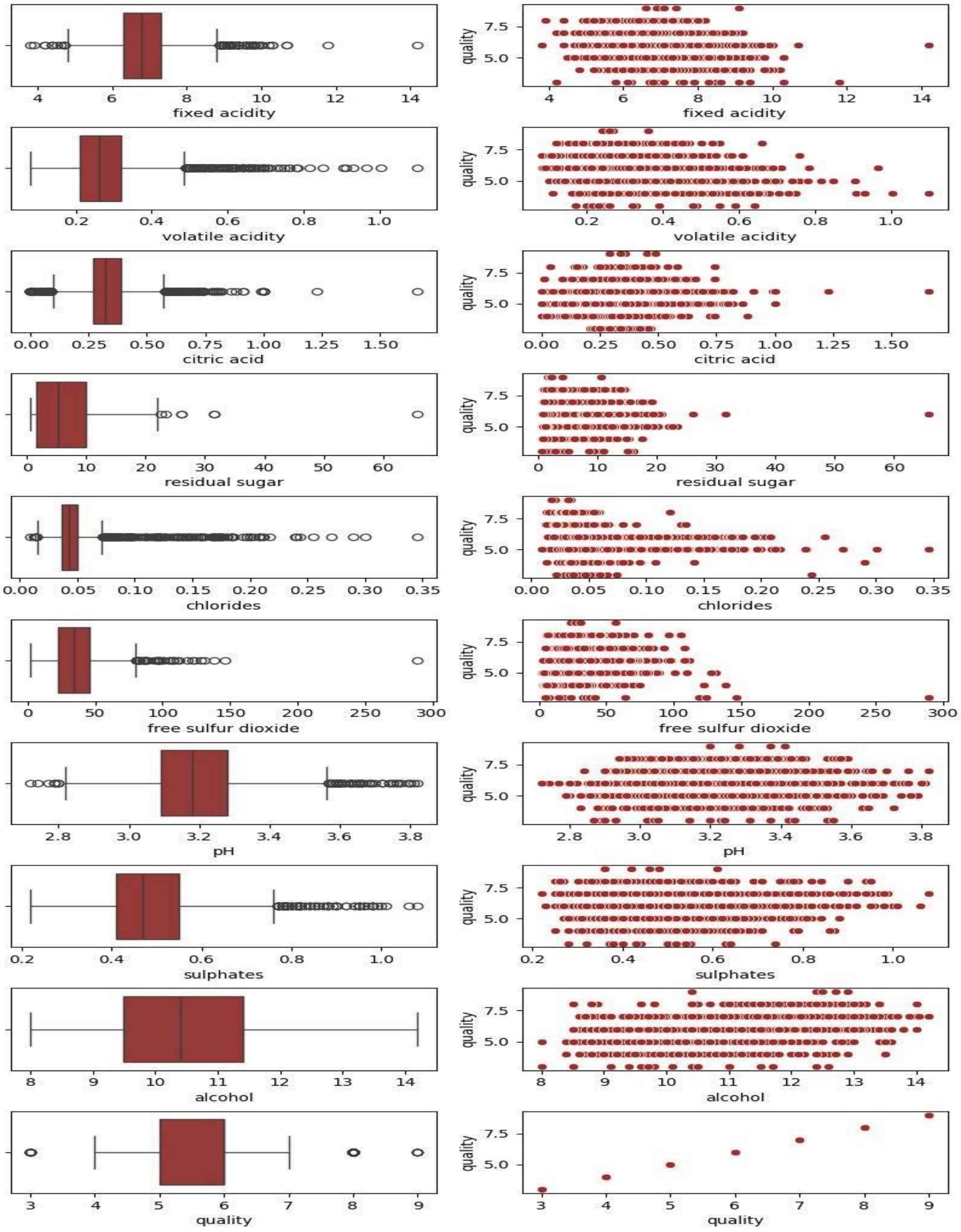
Box plot and scatter plot for the drilling data with outliers is given below.



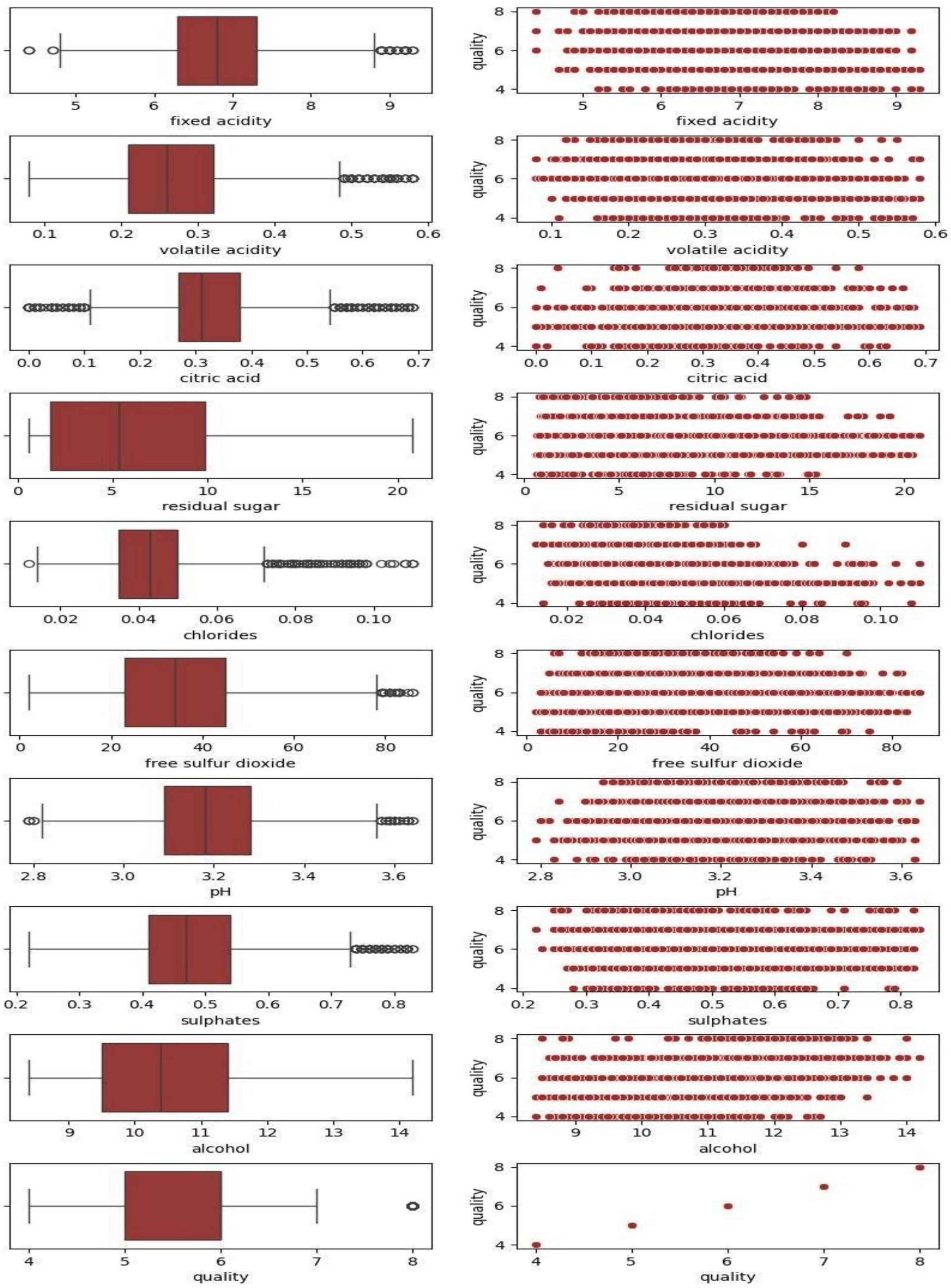
Box plot and scatter plot of the drilling data without outliers is given below.



Box plot and scatter plot for wine quality dataset with outliers.



Box plot and scatter plot for wine quality dataset without outliers.



5 6 7 8 9
fixed acidity
quality
8 6 4
fixed acidity
0.1 0.2 0.3 0.4 0.5 0.6
volatile acidity
volatile acidity
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
citric acid
citric acid
0 5 10 15 20
residual sugar
residual sugar
0.02 0.04 0.06 0.08 0.10
chlorides
chlorides
0 20 40 60 80
free sulfur dioxide
free sulfur dioxide
2.8 3.0 3.2 3.4 3.6
pH
pH
0.2 0.3 0.4 0.5 0.6 0.7 0.8
sulphates
sulphates
9 10 11 12 13 14
alcohol
alcohol
4 5 6 7 8
quality
quality