Hole Deviation Prediction in Oil Drilling

Onkar Dubey

210107060

Submission Date: April 25, 2024

Final Project submission

Course Name : Applications of Al and ML in chemical engineering

Course Code: CL653

Contents

1.Executive Summary

This project aims to analyze a dataset related to geological formations, specifically focusing on various properties such as depth, gamma-ray, shale volume, resistivity, delta T, Vp, Vs, density, calculated density, neuron porosity, density porosity, Poisson's ratio, and classification. These properties are crucial in understanding the characteristics of the geological formations and can provide valuable insights for various applications such as oil and gas exploration, groundwater management, and geotechnical engineering.

2.Introduction

The problem this project seeks to solve is to understand the relationships between these properties and how they contribute to the overall classification of the geological formations. This could potentially lead to more efficient exploration strategies and better management of geological resources

3.Methodology

Data Source: https://raw.githubusercontent.com/dubey-0nkar/Cl653/main/well_log.csv

Data Preprocessing:

1. Data Cleaning: Handling of missing data points, which may involve imputation techniques such as mean substitution or deletion of records with missing values.

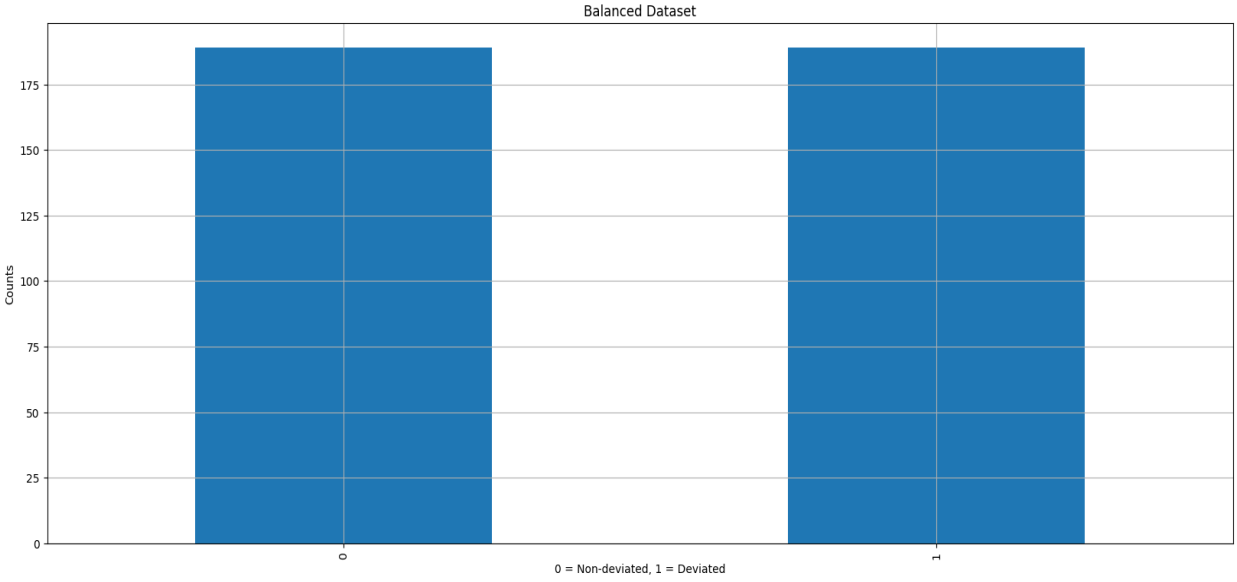
There were no missing data points in my dataset.

```
Depth                   0
Gamma-ray               0
Shale_Volume            0
Restivity               0
Delta T                 0
Vp                      0
Vs                      0
Density                 0
Density_Calculated      0
Neuron_Porosity         0
Density_Porosity        0
Possions_Ratio          0
Classification          0
dtype: int64
```

2. Normalization (numerical features): Scaling numerical features to a consistent range to prevent biases in model training, using techniques like min-max scaling or z-score normalization.

3. Text Preprocess: One-hot encoding or Label encoding for converting categorical features to numerical features**.**

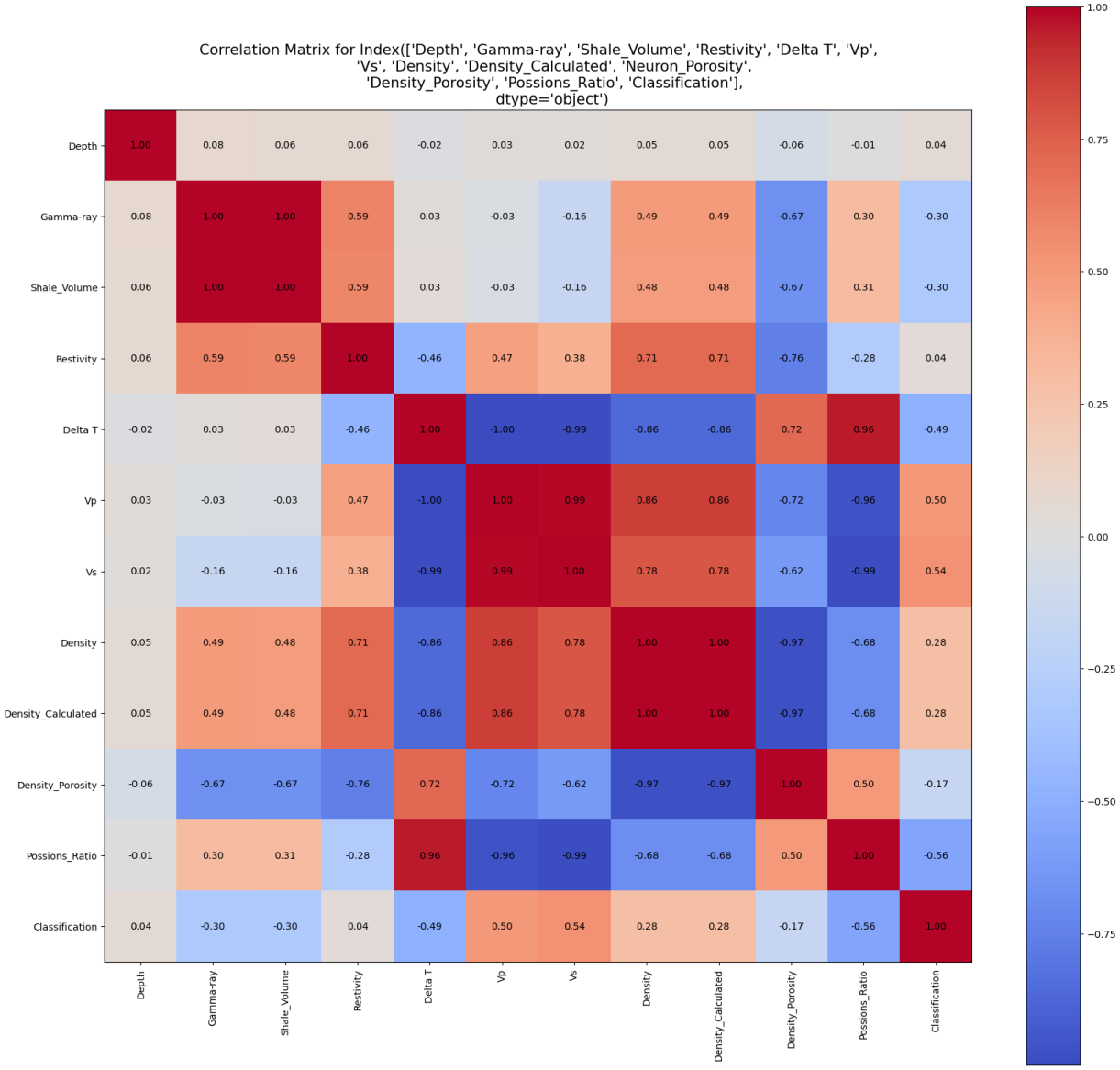
4. Handling Imbalanced Data: Apply techniques like oversampling or undersampling to balance class distributions if significant imbalances are observed.



5. Feature Engineering: Explore potential interactions or combinations of features that may provide additional predictive power.

6. Exploratory Data Analysis (EDA): Conduct exploratory data analysis to gain insights into the relationships between features and the target variable. Visualize distributions, correlations, and patterns in the data using techniques like

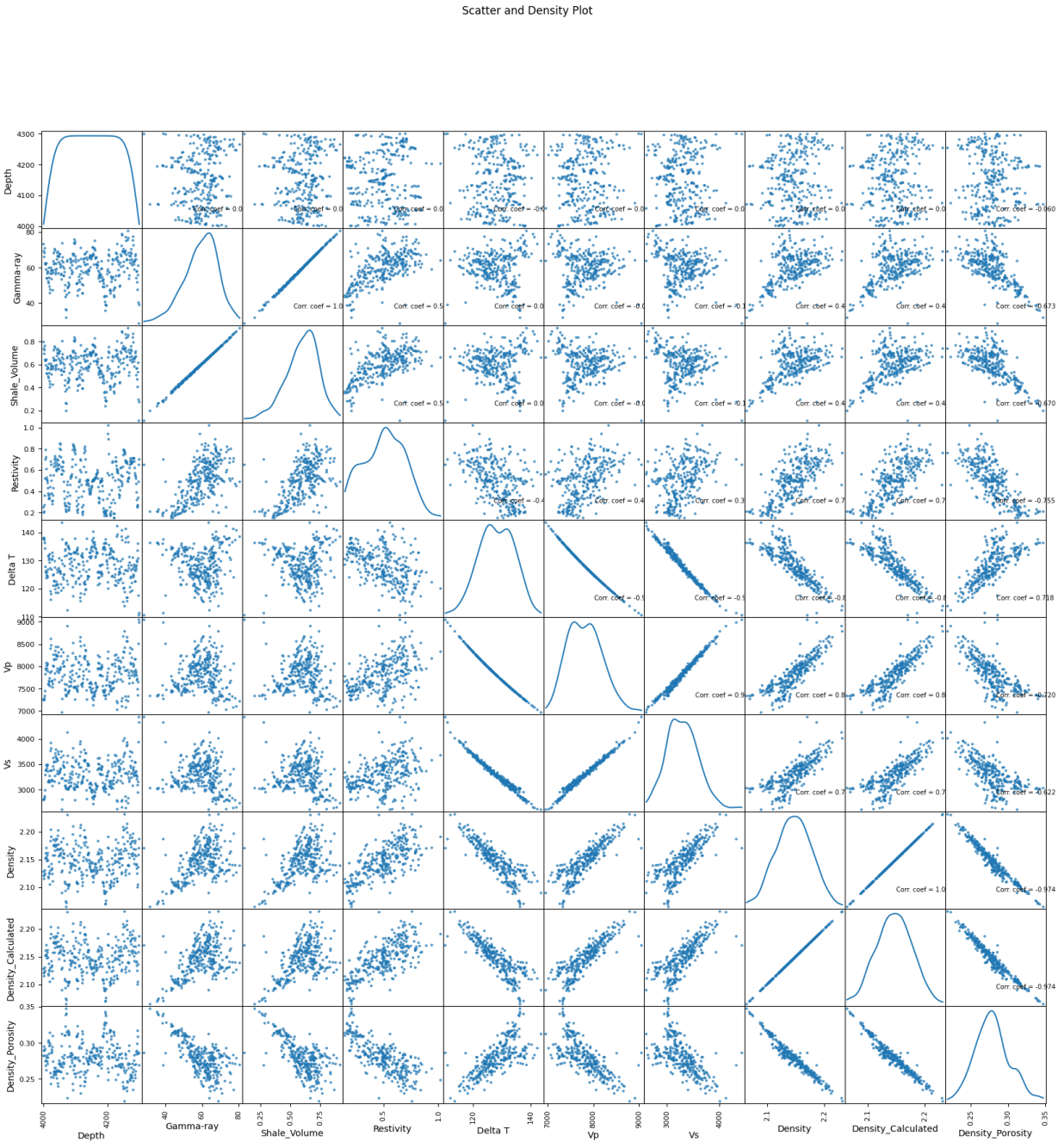
histograms, scatter plots, or correlation matrices. Identify any potential trends or outliers that may impact model training and interpretation.



Model Architecture: Four different models are considered for this project: Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest. These models are chosen because they are suitable for classification tasks.

Tools and Technologies: Google Colab , Python,Visual Studio Code.

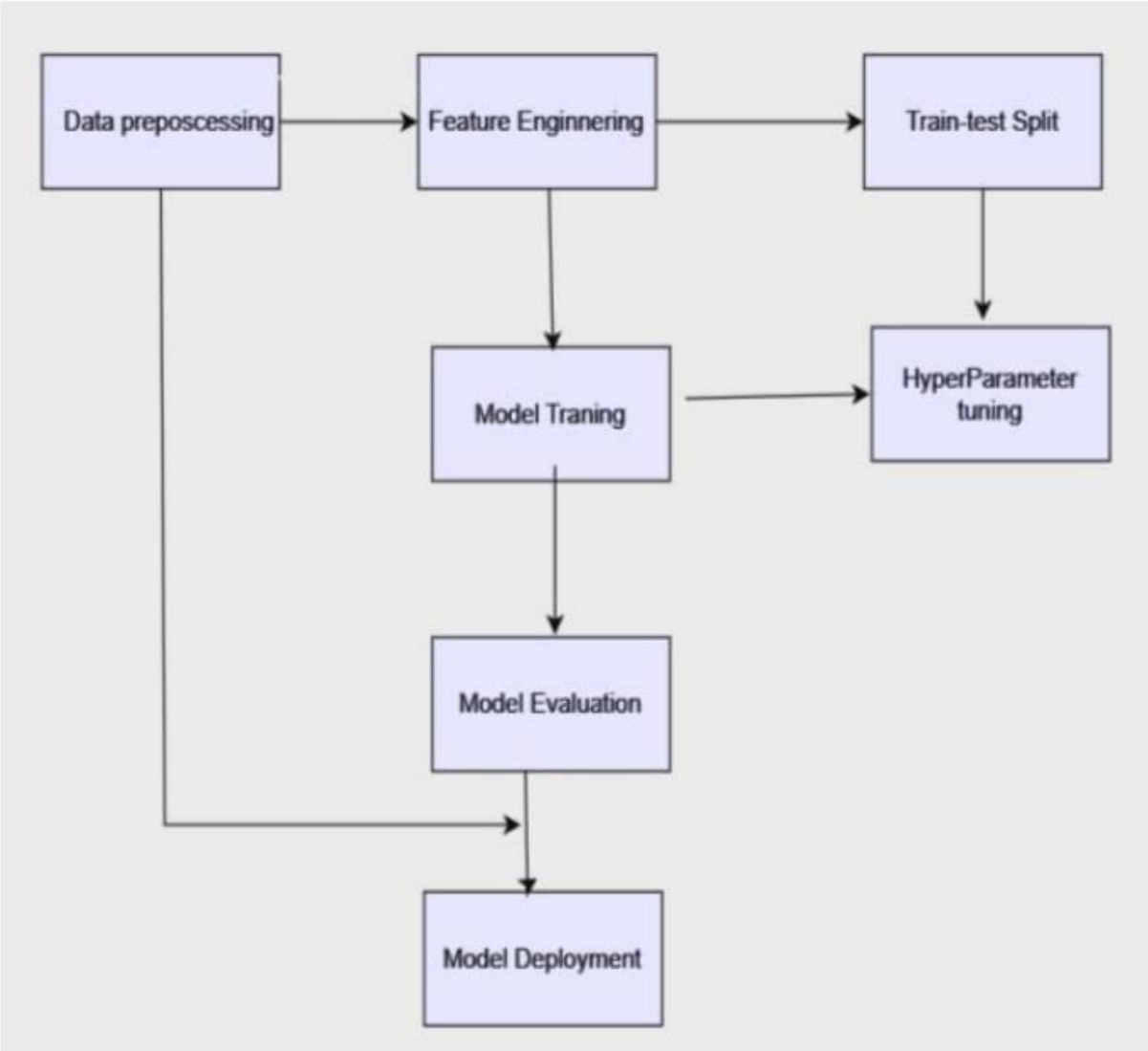
Different Plots :



4.Implementation Plan

Development Phases: The development phases for this dataset include data preprocessing, where the data is split into features and target variables and

standardized. This is followed by model selection and hyperparameter tuning using GridSearchCV for Logistic Regression, K-Nearest Neighbors, Decision Tree, and Random Forest models. The models are then evaluated based on their accuracy scores, and the best model is selected. Finally, the performance of the models is compared visually using a bar plot.



Model Training: We will train the selected models (Decision Trees, Random Forest, SVM, Neural Networks) using the training data. This involves feeding the data to the model and adjusting the model's parameters to minimize the prediction error.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=22)
```

```python
def apply_model(model,x_train,x_test,y_train,y_test):
    print(' '*20+re.search(r'(.*)\(',f'''{model}''',re.M).group(1))
    model.fit(x_train,y_train)
    y_pred = model.predict(x_test)
    print('')
    print('Train Score:  ',model.score(x_train,y_train))
    print('Test Score:   ',model.score(x_test,y_test))
    print('')
    print(classification_report(y_test,y_pred))
```

The above code snippet is a function to apply different models.

Model Evaluation: We will evaluate the performance of the models using the test data. We will use appropriate metrics such as accuracy, precision, recall, F1 score, and ROC AUC score.

The below picture represents accuracy scores for different models.

```
{'Logistic Regression': 0.8205128205128205,
 'KNN': 0.8192307692307693,
 'Random Forest': 0.9205128205128205,
 'Decision Tree': 0.885897435897436}
```

	precision	recall	f1-score	support
0	0.92	1.00	0.96	33
1	1.00	0.93	0.96	43
accuracy			0.96	76
macro avg	0.96	0.97	0.96	76
weighted avg	0.96	0.96	0.96	76

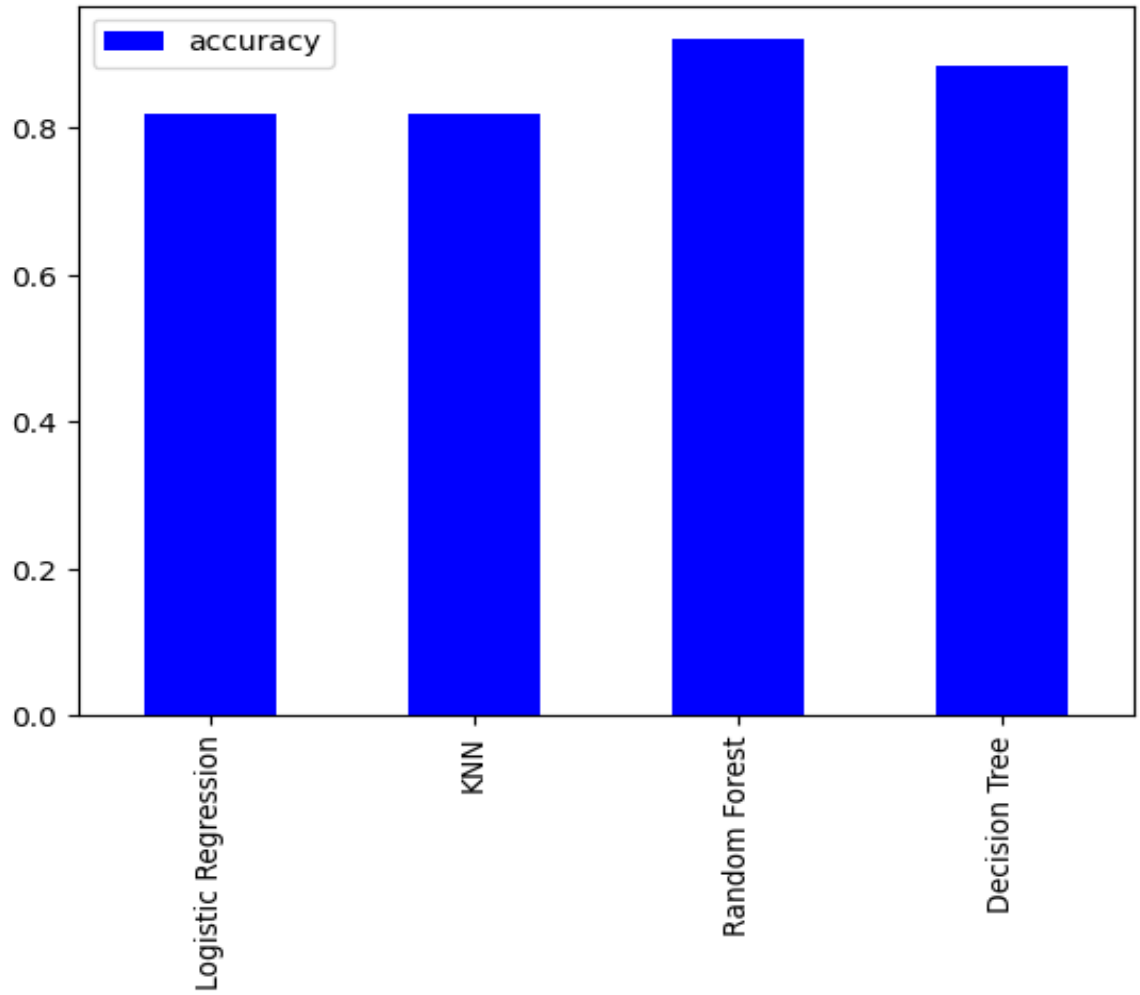
The above picture represents various scores(precision,recalls,etc)

5.Testing and Deployment

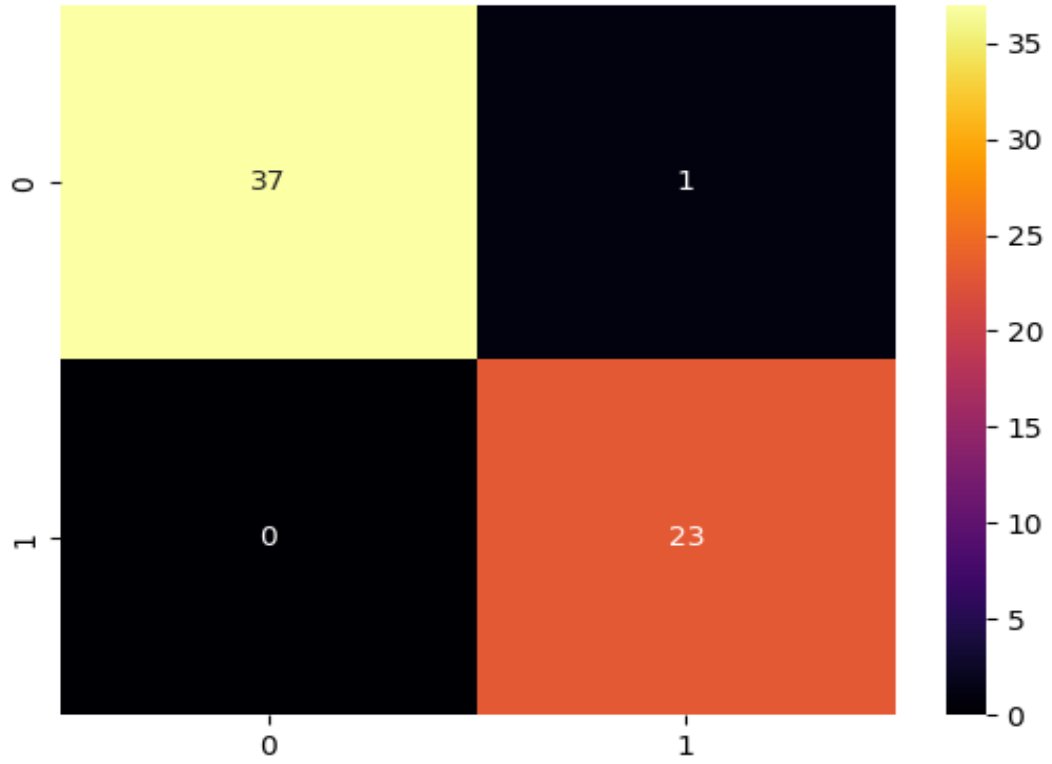
Testing Strategy: The testing strategy for this dataset involves using the test set that was held out during the train-test split. After the models are trained and their hyperparameters are tuned, they are used to make predictions on the test set. The performance of the models is then evaluated by comparing these predictions to

the actual values in the test set, using appropriate metrics such as accuracy. This helps in understanding how well the models generalize to unseen data.

Deployment Strategy:. The deployment strategy for this project would involve selecting the best performing model based on the testing results and then integrating it into a production environment.



The above plot represents which model have highest accuracy scores(without hyperparameter tuning).Here, random forest classifier have highest accuracy.



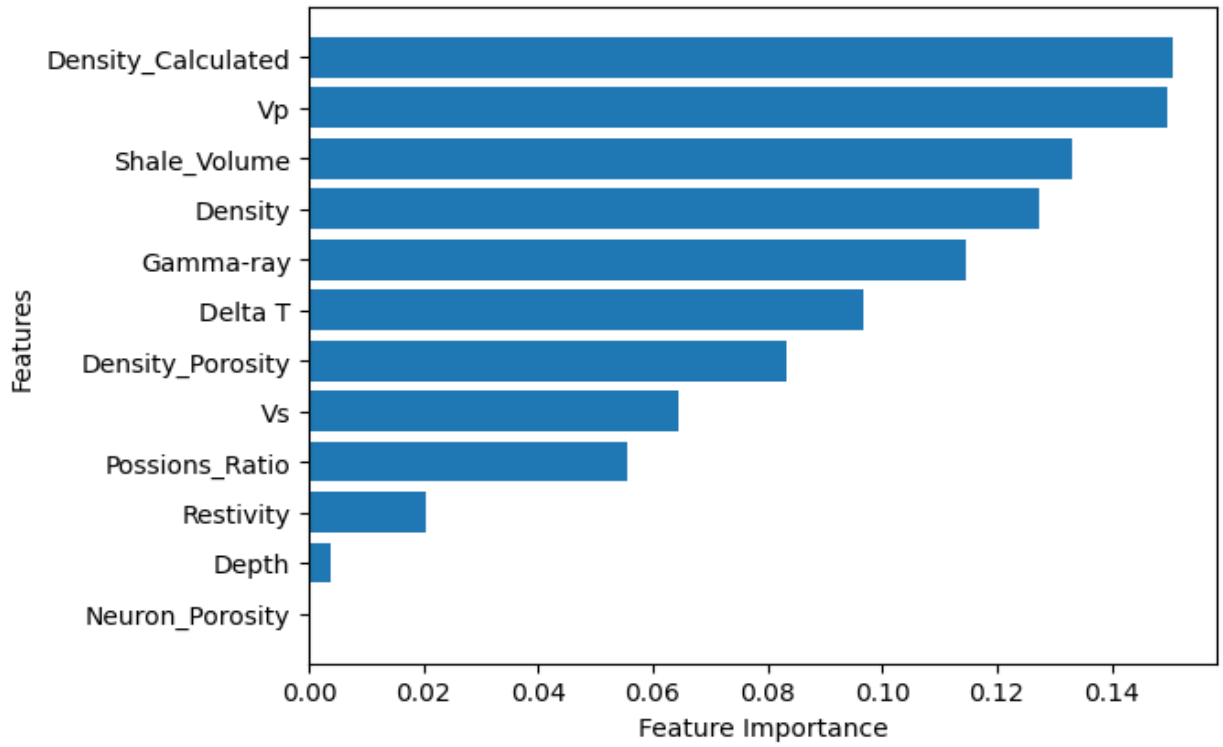
(Confusion Matrix) Here, the number of True Positive and true negative values are much higher than others.

Ethical Considerations: Deploying a model from this dataset raises ethical considerations. Data privacy is paramount, ensuring sensitive information is anonymized and securely stored. The model must be free from bias, ensuring fair and representative predictions. Transparency is crucial, users should understand how predictions are made. Accountability mechanisms must be in place to monitor performance and address incorrect predictions. Safeguards should be established to prevent misuse of the model's predictions.

6.Results and Discussion

Findings : The feature importance plot provides insights into the key drivers of the model's predictions. It ranks the top 20 features based on their contribution to the model's accuracy. The higher the feature on the plot, the more significant its impact on the model's output. This information is crucial for understanding the

model's decision-making process, identifying the most influential factors, and potentially simplifying the model by removing less important features.



Challenges and Limitations : The quality and representativeness of the data can significantly impact the model's performance. Any biases in the data can lead to biased predictions. Also, selecting model which model to use was quite challenging.

7.Conclusion and Future Work

Our project is focused on developing a machine learning model to analyze well log data, a critical component in oil and gas exploration. The innovative aspect of our project lies in the application of advanced machine learning techniques to predict subsurface properties such as shale volume, resistivity, and porosity. This approach not only automates the analysis process but also enhances the accuracy of predictions, reducing the risk of human error.

The potential impact of our project is substantial. It can contribute to making oil and gas exploration more efficient and environmentally friendly, and help address

challenges related to water conservation and soil stability. In essence, our project stands at the intersection of technology and environmental sustainability, making a significant contribution to the fields of oil and gas exploration, environmental science, and financial analysis.

8.References

Definition source: Hole deviation - PetroWiki

https://github.com/tannisthamaiti/ML_well_log

This well log dataset was forked from the github link provided, and was modified by Mbonu Chinedu. The classification column was added to the well log dataset for Hole Deviation Class.

9.Appendices

None

10.Auxiliaries

DataSource: https://raw.githubusercontent.com/dubey-0nkar/Cl653/main/well_log.csv

Pythonfile:

https://colab.research.google.com/drive/1qwOew5h5iQGX6teuzjKRYuNacHldBnlc?authuser=1#scrollTo=eIgErvWqnmm3