PHASE 3 PROJECT

NON-TECHNICAL PRESENTATION

Overview

- Business understanding
- Data understanding
- Correlation analysis
- Modelling
- Evaluation
- Recommendations
- Next steps

Business Understanding

The goal of this project is to predict the condition of water wells in Tanzania to help stakeholders—such as NGOs, government bodies, and water management organizations—optimize resource allocation, maintenance efforts, and future well construction planning. This analysis presents the approach to addressing water access challenges in Tanzania through predictive modeling.

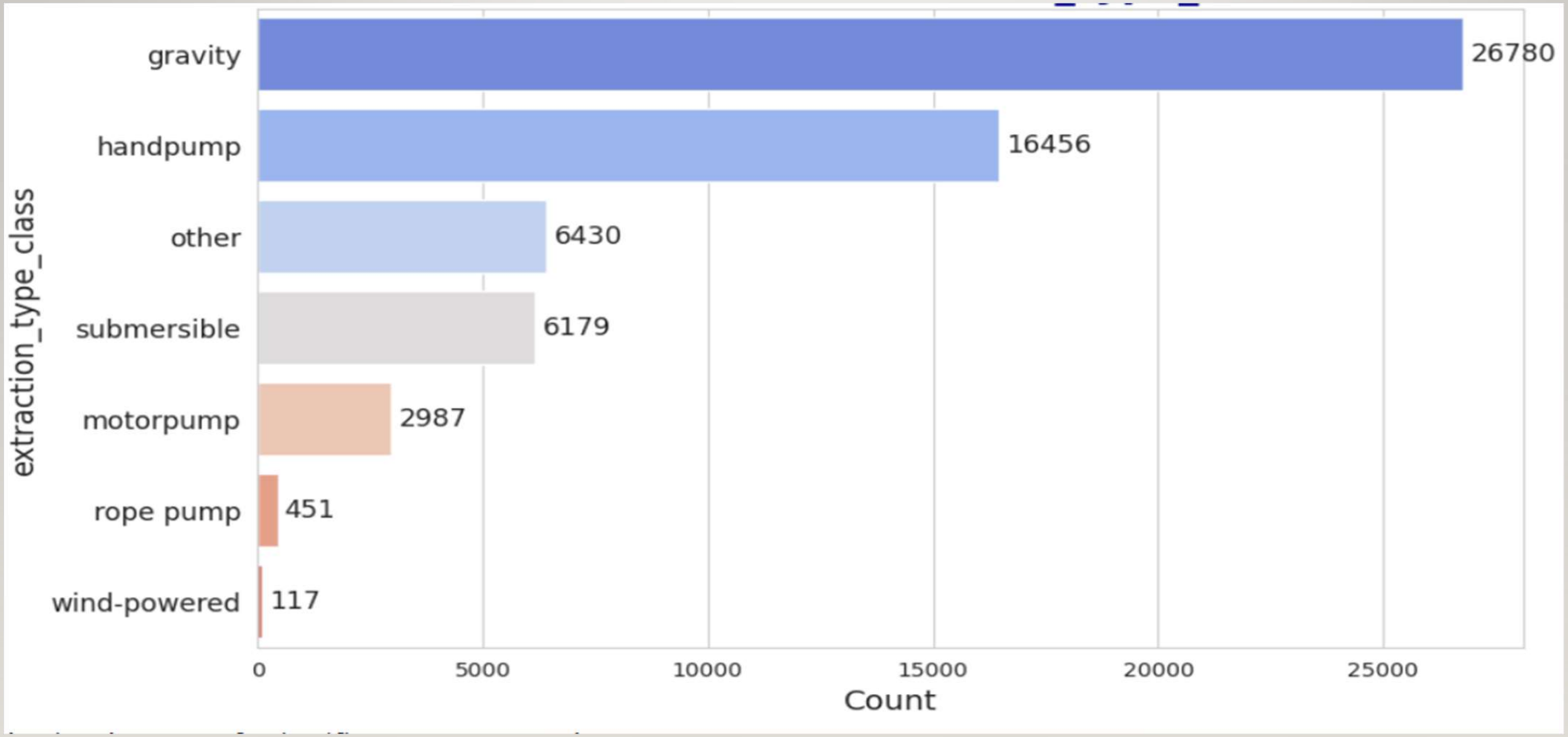
Data Understanding

The target variable categorizes water points into three groups:

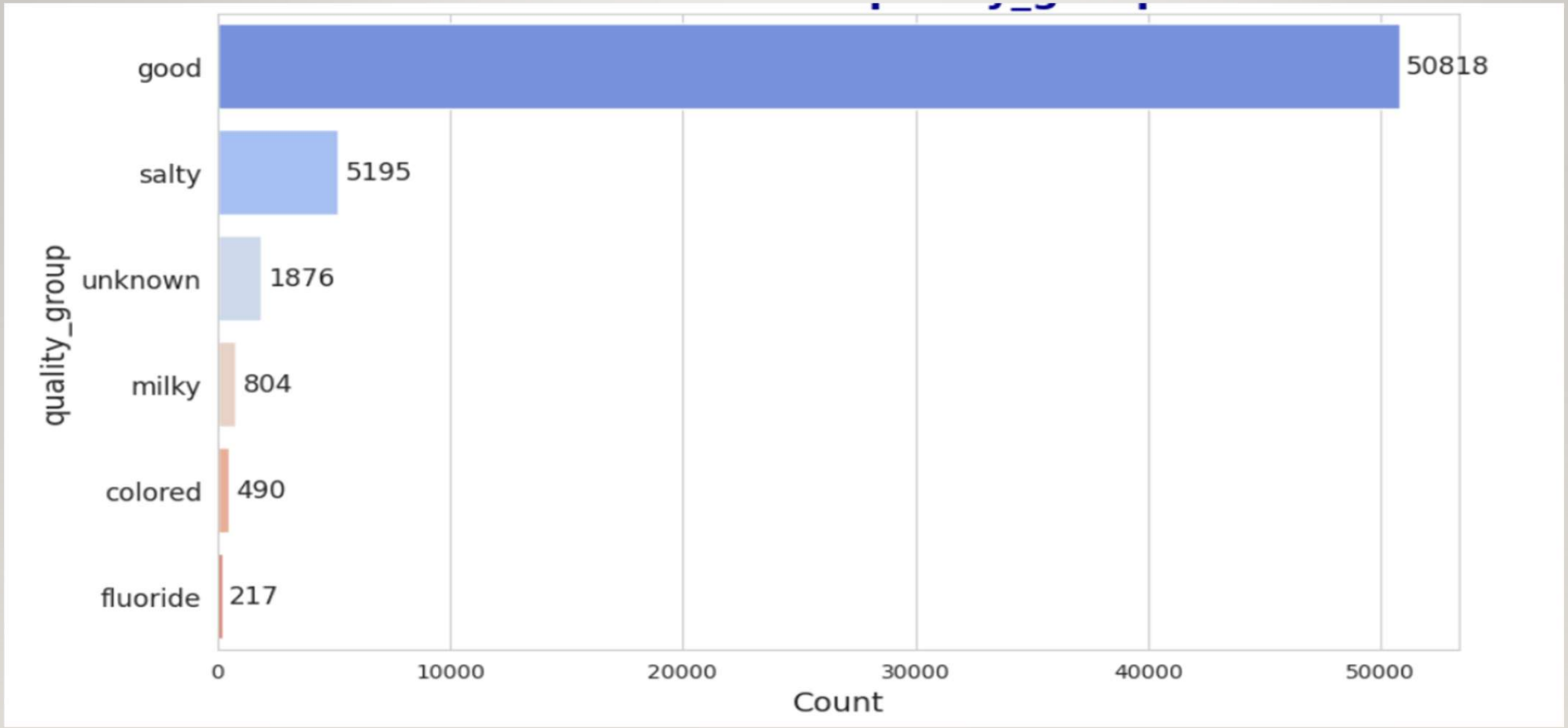
- **Functional** – The water point is fully operational with no repairs needed.
- **Functional but needs repair** – The water point is working but requires maintenance.
- **Non-functional** – The water point is not operational.

We are aiming to understand the data distribution among the target variables and get model performance insights from the data.

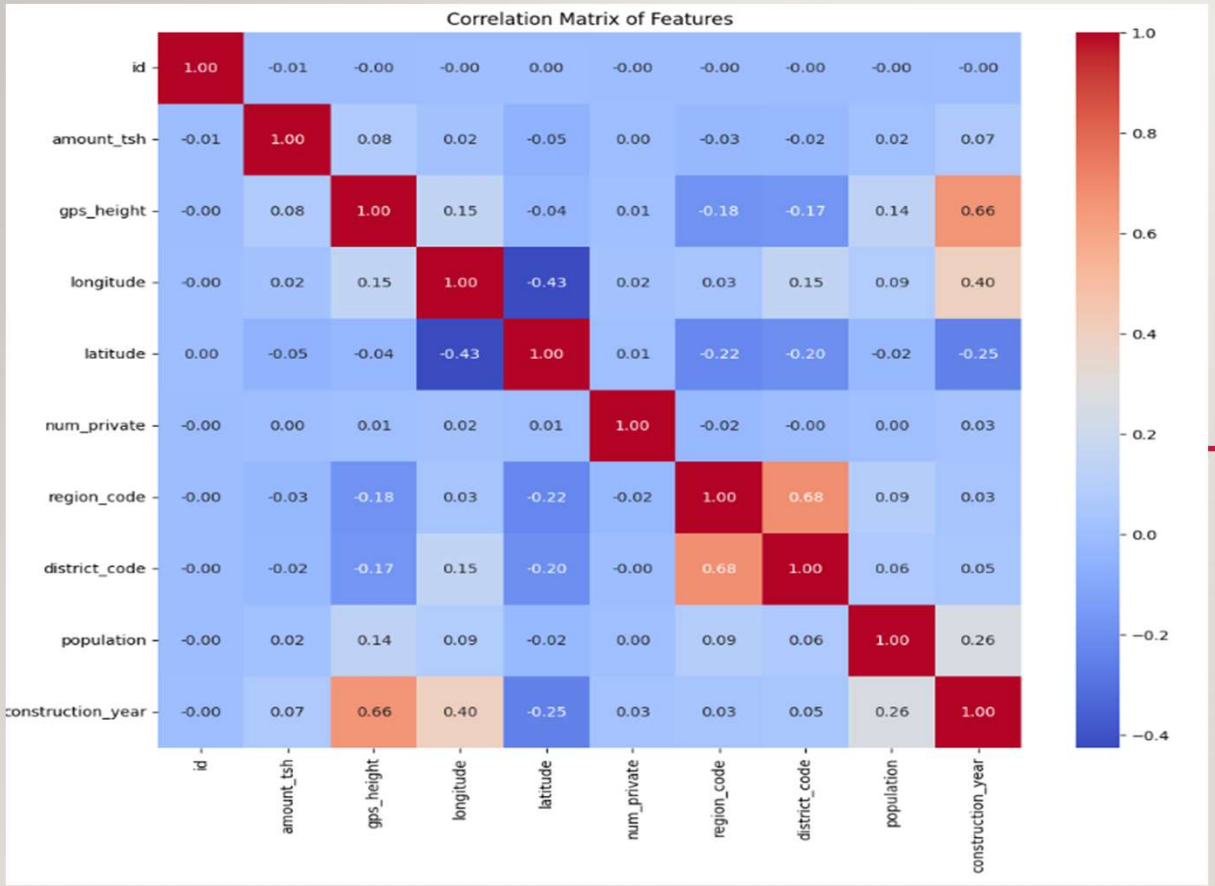
Distribution of the water Pumps Types



Distribution of the water Quality



Correlation Analysis



Conducted a correlation analysis to identify relation between the different variables

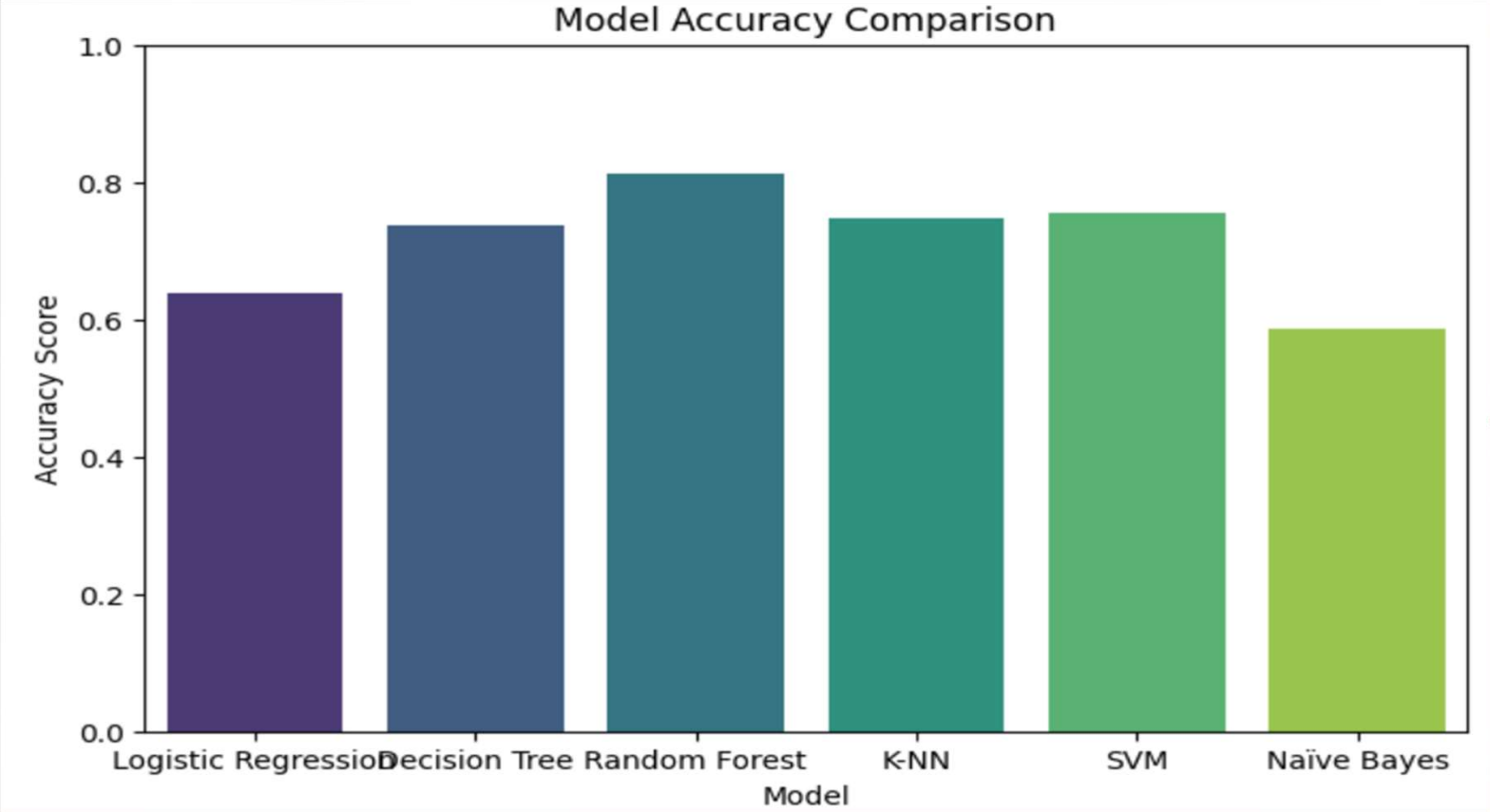
Modelling

Model	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)
Logistic Regression	0.6382	0.47	0.44	0.43
Decision Tree	0.7385	0.62	0.63	0.62
Random Forest	**0.8114**	**0.74**	**0.67**	**0.69**
K-NN	0.7468	0.66	0.60	0.62
SVM	0.7562	0.71	0.56	0.57
Naïve Bayes	0.5859	0.48	0.50	0.49

Six models were used to do the classification and give scores based on:

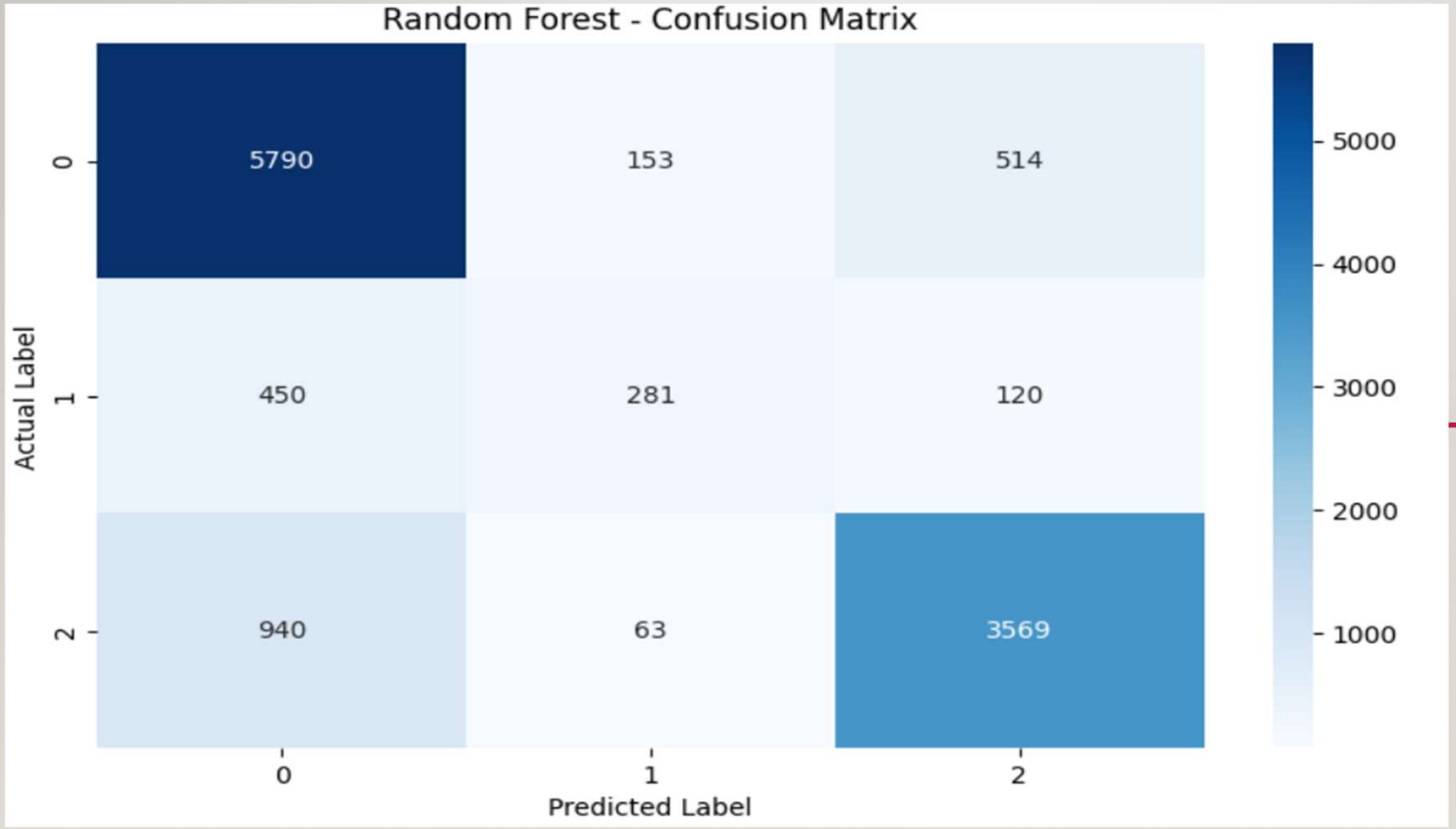
- Accuracy
- Precision
- Recall
- F1-Score

Evaluation- Accuracy Score



- Random Forest performs the best overall with an accuracy of 81%
- Decision tree and KNN are also decent alternatives with an accuracy of over 73%

Random Forest Confusion Matrix



- Accuracy: 81%
- Precision, Recall, F1-Score: Strong performance for Class 0 and Class 2, but weaker for Class 1

Recommendations

1) Best Model:

- Use Random Forest as it consistently outperforms the other models across

2) Improving Class 1 Performance:

- Investigate class imbalance. If class 1 is underrepresented, consider techniques like oversampling (e.g., SMOTE) or class weighting.
- Experiment with hyperparameter tuning for better performance on this class.

3) Alternative Models:

- If interpretability is important, consider Decision Tree or Logistic Regression (with improvements).
- If computational efficiency is a concern, K-NN or SVM are reasonable alternatives.

4) Avoid Naïve Bayes:

- Its performance is significantly worse than the other models, likely due to its assumptions

Next Steps

- **Data Preprocessing**:
  - Handle missing values, encode categorical variables, and normalize/scale numerical features.
- **Model Improvement**:
  - Experiment with advanced techniques like ensemble methods (e.g., Gradient Boosting, XGBoost) or deep learning models.
- **Deployment**:
  - Once the best model is identified, deploy it for real-world predictions and monitor its performance over time.

QUICKMART FRESH & EASY

THANK YOU