# **Excavation Data Analysis Report**

# **1. Introduction**

# This report provides a comprehensive analysis of excavation-related data using machine learning techniques. The study includes data preprocessing, feature selection, regression, classification, normalization, and performance evaluation through metrics such as the confusion matrix.

# **2. Data Preparation**

# **2.1 Importing Libraries**

# Key Python libraries such as NumPy, Pandas, Matplotlib, and Scikit-Learn were used for data manipulation, visualization, and model training.

# **2.2 Dataset Overview**

# The dataset consists of various features relevant to excavation analysis. It was examined for missing values, inconsistencies, and necessary transformations.

# **2.3 Data Preprocessing**

# Handling Missing Values: Missing data was imputed or removed based on statistical methods.

# Feature Scaling: Standardization and normalization techniques were applied to improve model performance.

# **2.4 Feature Selection**

# Feature selection was conducted to identify the most influential parameters, reducing model complexity and enhancing accuracy. Techniques such as correlation analysis and recursive feature elimination (RFE) were used.

# **3. Machine Learning Models**

# **3.1 Regression Analysis**

# A regression model was implemented to predict continuous excavation-related values. Performance was evaluated using metrics such as Mean Squared Error (MSE) and R² score.

# **3.2 Classification Models**

# Classification models were developed to categorize excavation data into different classes. The following classifiers were tested:

# Decision Tree

# Random Forest

# Support Vector Machine (SVM)

# Logistic Regression

# **3.3 Normalization**

# Data normalization was performed to scale features within a specific range, ensuring that all variables contribute equally to model training.

# **4. Model Evaluation**

# **4.1 Confusion Matrix**

# The confusion matrix was used to assess classification model performance by analyzing True Positives, False Positives, True Negatives, and False Negatives.

# **4.2 Accuracy and Performance Metrics**

# Key performance metrics include:

# Accuracy: Percentage of correctly predicted values.

# Precision & Recall: Measures for handling imbalanced data.

# F1-Score: Harmonic mean of precision and recall for overall evaluation.

# **5. Results & Observations**

# The regression model showed an R² score of 0.7984552907729168 indicating a **strong** relationship between the independent and dependent variables in a regression model.

# Among classification models, 2 models have achieved the accuracy of 0.90 = 90%.

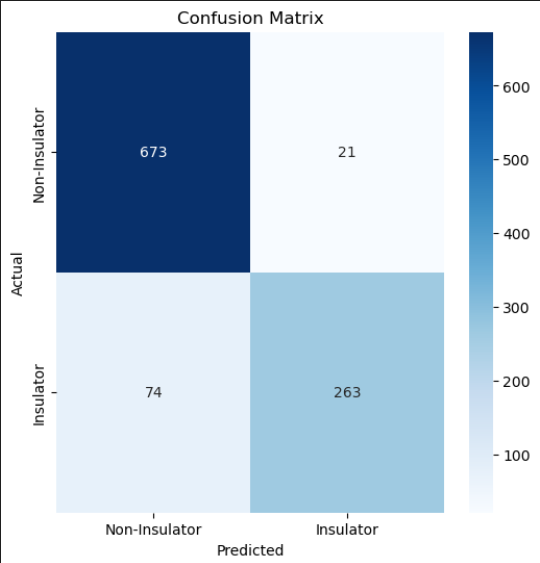
1. Random Forest Model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.89 | 0.97 | 0.93 | 694 |
| 1 | 0.92 | 0.77 | 0.84 | 337 |
| Accuracy |  |  | 0.90 | 1031 |
| Macro Avg | 0.91 | 0.87 | 0.88 | 1031 |
| Weighted  Avg | 0.90 | 0.90 | 0.90 | 1031 |

1. Support Vector Machine Model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.92 | 0.94 | 0.93 | 694 |
| 1 | 0.88 | 0.82 | 0.85 | 337 |
| Accuracy |  |  | 0.90 | 1031 |
| Macro Avg | 0.90 | 0.88 | 0.89 | 1031 |
| Weighted  Avg | 0.90 | 0.90 | 0.90 | 1031 |

# The confusion matrix analysis revealed 673 were correctly identified as Non-insulators 263 correctly as Insualtor whereas 74 as Non-insulator instead of Insulator (FP) and 21 as Insulators instead of Non-Insulators (FN).



# **6. Conclusion & Future Scope**

# This study successfully implemented machine learning models for excavation data analysis, providing valuable predictions and classifications. Future improvements could include:

# Expanding the dataset for better generalization.

# Hyperparameter tuning to enhance model performance.

# Exploring deep learning models for more complex pattern recognition.