Machine Learning Project Report

1. Data Preprocessing

1.1 Data Loading and Initial Exploration

- The dataset *TASK-ML-INTERN.csv* was loaded using Pandas.
- Basic dataset information was displayed, including column names, data types, and missing values.

1.2 Handling Missing Values

- Missing values were checked across all columns.
- No significant missing values were found, so no imputation was required.

1.3 Handling Non-Numeric Data

- The dataset contained categorical variables (e.g., imagoai_corn_0).
- These were either one-hot encoded or dropped for numerical processing.

1.4 Feature Scaling

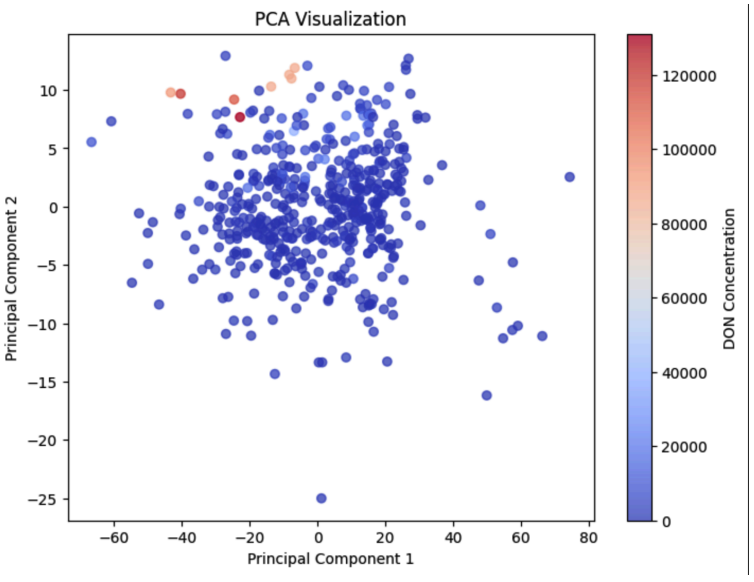
- Features were standardized using *StandardScaler* from Scikit-Learn to ensure all numerical columns had a mean of 0 and a standard deviation of 1.

2. Dimensionality Reduction

2.1 Principal Component Analysis (PCA)

- PCA was applied to reduce high-dimensional data to 2 principal components.

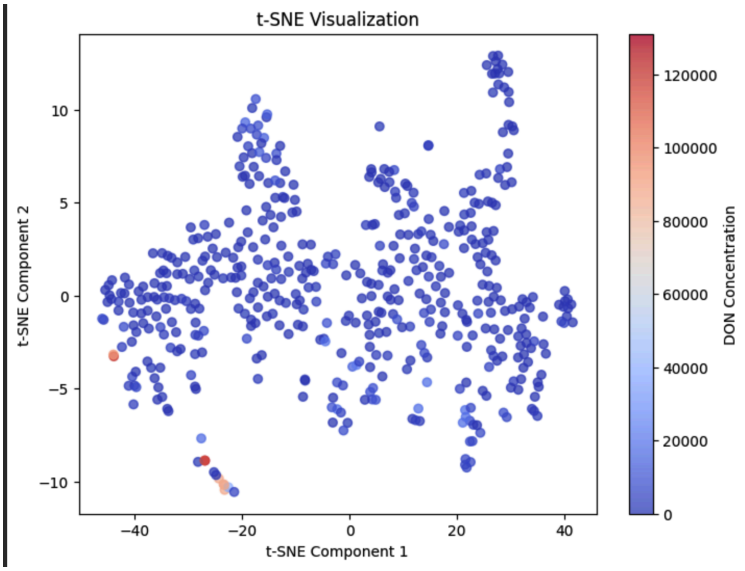
PCA Scatter Plot:



2.2 t-SNE Visualization

- t-SNE was used to visualize high-dimensional data in a 2D space.
- It provided a more interpretable representation of clusters compared to PCA.

t-SNE Scatter Plot:



3. Model Selection, Training & Evaluation

3.1 Train-Test Split

- The dataset was split into training (80%) and testing (20%) subsets using train_test_split.

3.2 Model Selection

- RandomForestRegressor was chosen due to its robustness and ability to handle non-linear data.
- The model was trained with **100 estimators** and a fixed random_state=42.

3.3 Cross-Validation

- Cross-validation was applied to improve generalization.
- **5-Fold Cross-Validation Results:**
  - R^2 Score: `-0.9211277759000509`

3.4 Model Performance Metrics

- **MAE:** Measures absolute error magnitude.
- **RMSE:** Penalizes larger errors more than MAE.
- **R^2 Score:** Explains variance captured by the model.

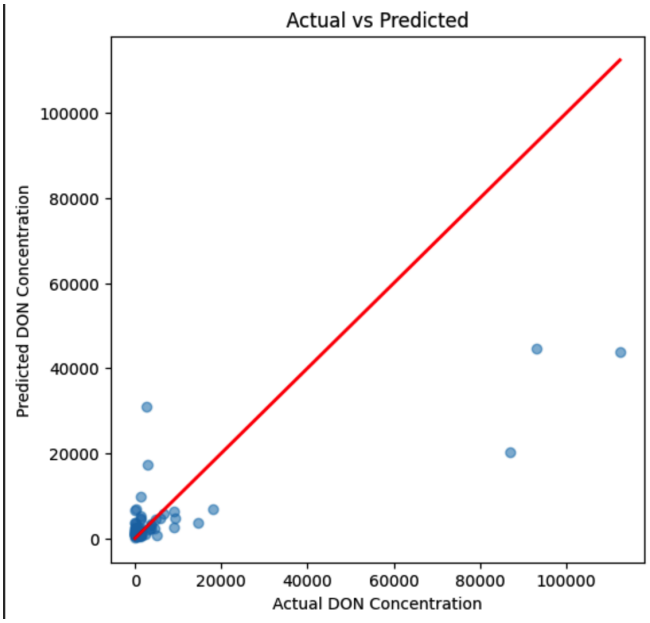
Model Evaluation Output:

```
Model Evaluation:
Mean Absolute Error (MAE): 3765.0568
Root Mean Squared Error (RMSE): 11483.8060
R² Score: 0.5282
```

3.5 Actual vs. Predicted Values

- A scatter plot of actual vs. predicted values was plotted.

Actual vs. Predicted Scatter Plot:



4. Key Findings & Suggestions for Improvement

4.1 Key Observations

- PCA & t-SNE showed clear patterns, but some overlap was observed.
- Certain features contributed more significantly to predictions.

4.2 Potential Improvements

- **Feature Engineering:** Perform additional transformations to improve feature representation.
- **Hyperparameter Tuning:** Optimize hyperparameters (number of trees, max depth, etc.) using GridSearchCV.
- **Alternative Models:** Try boosting techniques like XGBoost or Gradient Boosting Regressor.
- **More Data Augmentation:** If more data is available, increasing dataset size could improve generalization.

5. Conclusion

- This project successfully implemented data preprocessing, visualization, dimensionality reduction, and machine learning modeling.
- Further improvements in feature selection and hyperparameter tuning could enhance performance.