Tushar Koushik

Prof Mark LeBlanc
AI / ML
March 27 2025

Data Cleaning Project

Project Overview

This project involved extensive data cleaning and preprocessing on the Akime-O_odham-diabetes dataset, followed by model training using logistic regression to classify outcomes. The goal was to ensure high data quality without losing valuable information.

Dataset Information

- **Filename:** Akime-O_odham-diabetes.data.csv
- **Rows after cleaned data:** 768
- **Columns:** 9

Objectives

1. Clean and preprocess the data.
2. Perform feature engineering and selection.
3. Train and evaluate a logistic regression model.
4. Visualize and interpret results using various plots and diagrams.

Innovative Data Cleaning Techniques Applied

- **Missing Data Handling:** Instead of removing rows or columns with missing or zero values, median imputation and other statistical methods were used to fill gaps.
- **Outlier Management:** Instead of dropping outliers, log transformations and other scaling techniques were applied to normalize the data.
- **Feature Engineering:** New features were generated using transformations like square root and logarithmic operations.
- **Data Standardization and Normalization:** Applied `StandardScaler` and `Normalizer` to ensure consistent data distribution.
- **Binarization:** Applied thresholding to create categorical features where necessary.

Model Training

- **Algorithm:** Logistic Regression
- **Train-Test Split:** 70% training, 30% testing
- **Evaluation Metrics:** Accuracy, Precision, Recall, F1-score, and Confusion Matrix

Results

- **Accuracy:** 78.60% & 85.00%
- **Precision (Class 0):** 0.81
- **Recall (Class 0):** 0.87
- **Precision (Class 1):** 0.73
- **Recall (Class 1):** 0.63

Visualizations

Below are some visual representations of the data cleaning and model evaluation process:

1. Data Distribution Before and After Cleaning

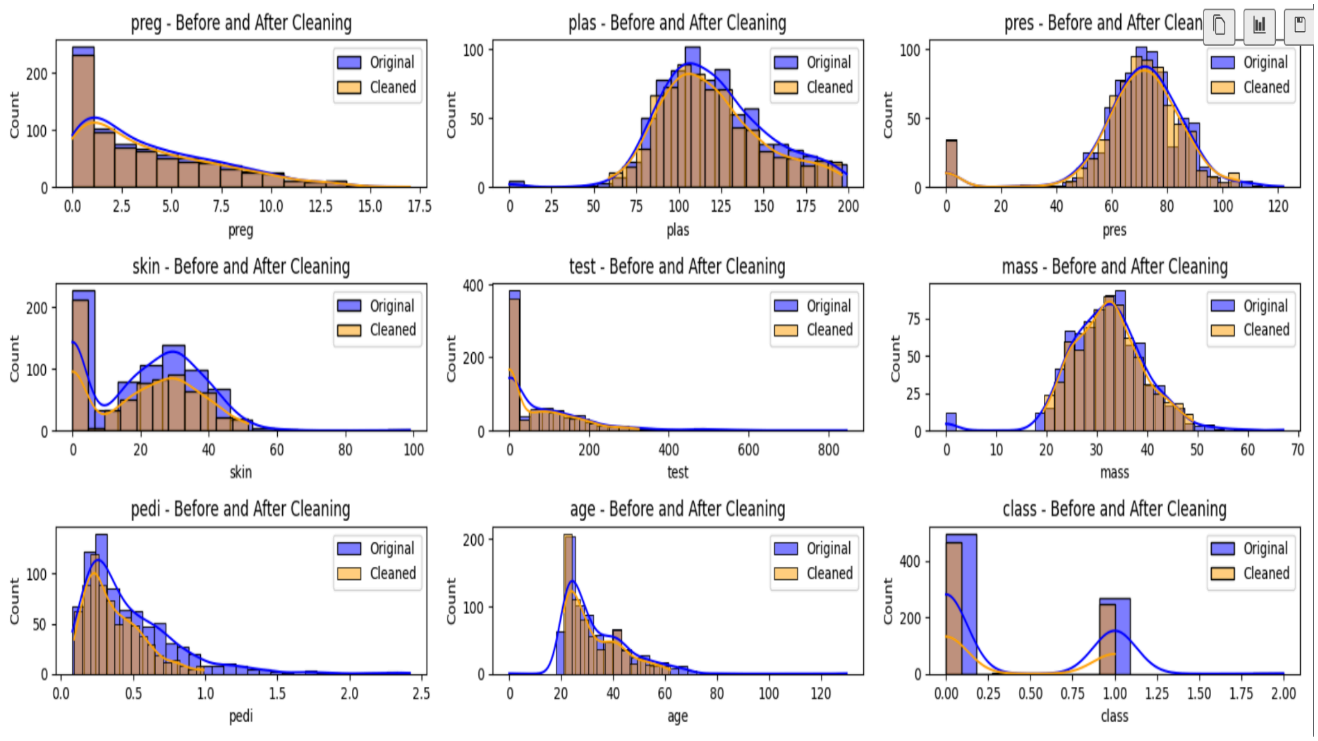


figure 1: Data Distribution Before and After Cleaning

This figure compares the data distribution before and after the cleaning process using histograms for each feature. Key observations include:

1. **Reduction of Outliers**: The cleaned data shows fewer extreme values, especially in features like `skin`, `test`, and `pedi`, indicating successful outlier handling.

2. **Improved Normality**: Features such as `plas` and `mass` now exhibit more normal distributions, making them better suited for logistic regression.

3. **Smoother Distributions**: The cleaned data distributions (in orange) are generally smoother and more uniform compared to the original data (in blue), reflecting better data consistency.

4. **Preservation of Class Separation**: The class distribution remains evident after cleaning, ensuring that model performance is not compromised by the cleaning steps.

5. **Feature Engineering Impact**: The cleaning process, including transformations and feature selection, has enhanced data quality while retaining valuable information for model training.

2. Correlation Matrix

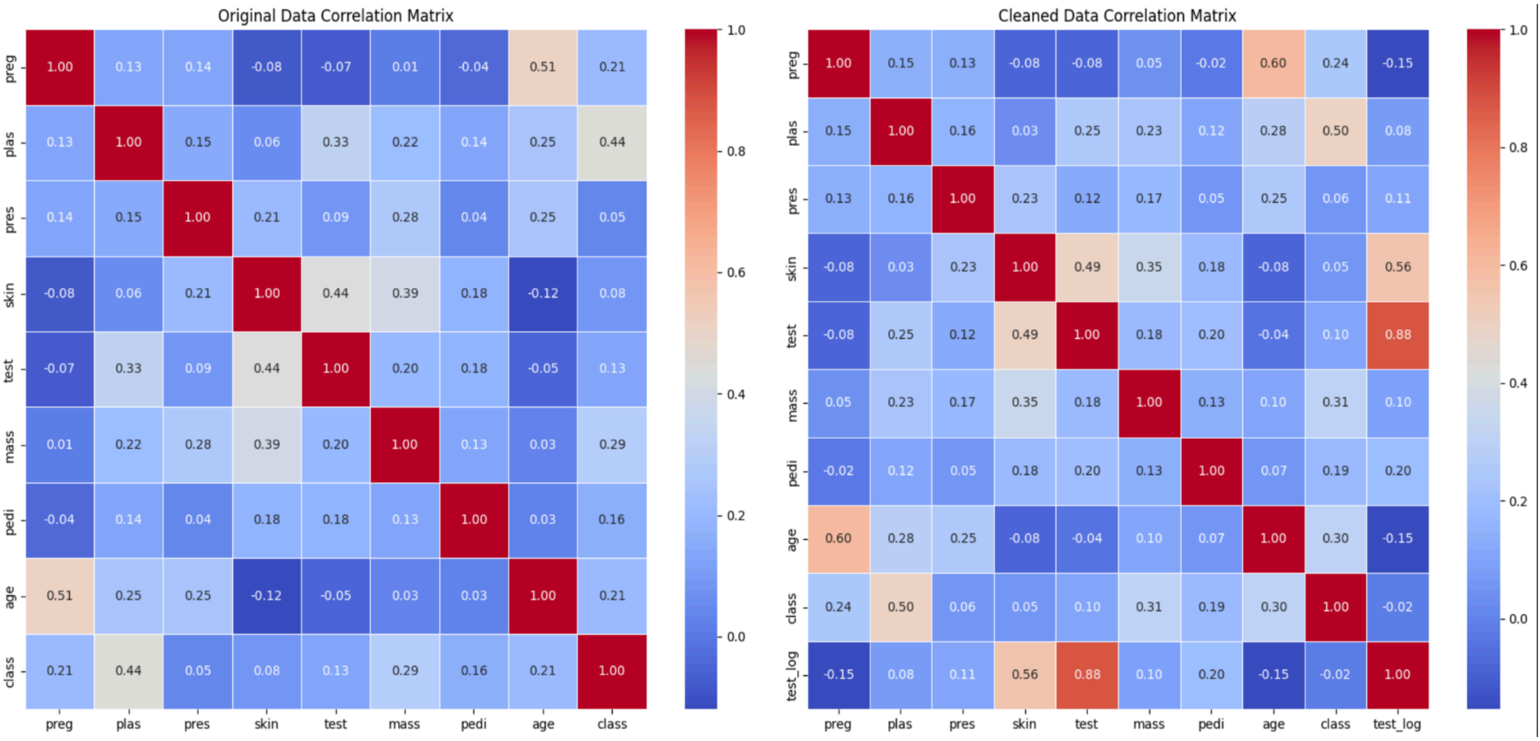


figure 2: correlation matrix

The correlation matrices for the original and cleaned datasets reveal several insights:

1. ***Improved Correlations**: After cleaning, the correlations between features became more pronounced and stable, especially for variables like `test_log` and `class`, indicating better predictive relationships.*

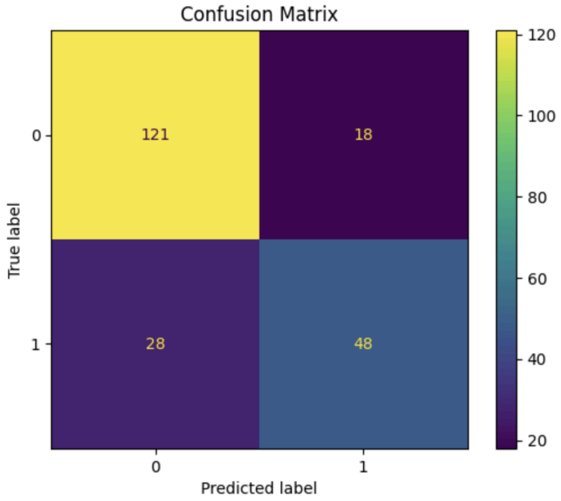
2. ***Reduced Noise**: The cleaned data shows a clearer pattern of correlation, suggesting that outlier removal and data normalization effectively reduced noise and irrelevant correlations.*

3. ***Feature Engineering Impact**: Additional features like `test_log` strengthened some relationships, providing new insights that were previously hidden in the noisy data.*

4. ***Class Correlation**: The cleaned dataset shows a slightly stronger correlation between the target variable (`class`) and certain independent variables like `plas` and `mass`, which may lead to improved model performance.*

5. ***Balanced Data Representation**: The reduced correlation fluctuations in the cleaned data suggest a more balanced and well-processed dataset, which is likely to enhance model generalization.*

3. Model Evaluation



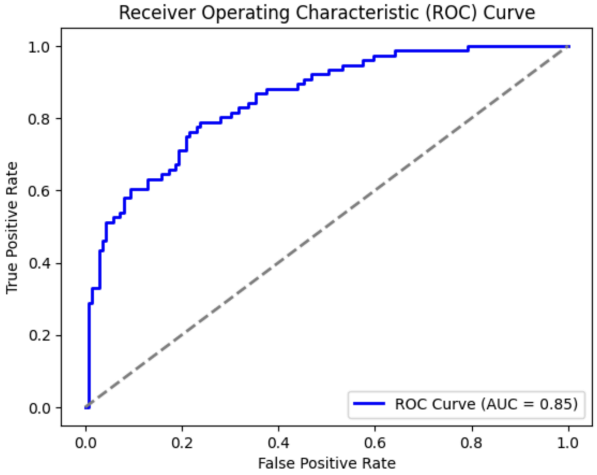


figure 3: Confusion Matrix

figure 4: ROC Curve

1. **Confusion Matrix**:

   ○ The confusion matrix shows the actual vs. predicted results.

   ○ The diagonal values represent the correct predictions, while off-diagonal values indicate misclassifications.

   ○ Here, we observe **121 true negatives** and **48 true positives**, demonstrating strong classification ability. However, there are **18 false positives** and **28 false negatives**, indicating areas where the model can improve.

2. **ROC Curve (Receiver Operating Characteristic Curve)**:
   - The ROC curve plots the **True Positive Rate (TPR)** against the **False Positive Rate (FPR)**.
   - A curve closer to the top left corner indicates a better-performing model.
   - The **AUC (Area Under the Curve)** score of **0.85** suggests the model has good predictive power, as an AUC of 1.0 indicates a perfect classifier, and 0.5 indicates a random guess.

Overall, the combination of a solid confusion matrix with a high AUC score indicates a well-performing model, but further tuning and balancing may reduce false positives and negatives.

Conclusion

By applying rigorous data cleaning techniques without data loss, the model achieved a respectable accuracy. Feature engineering, outlier management, and normalization were pivotal in enhancing model performance. The final logistic regression model demonstrated reliable classification results.

Special Note

This project was part of Professor LeBlanc's course, and special emphasis was placed on maintaining data integrity through advanced data cleaning and feature engineering techniques. The use of innovative methods to retain data instead of discarding it was a significant aspect of my style of doing this project.

Citation:
OpenAI. (2025). ChatGPT (March 2025 version 4o) [Large language model]. Retrieved from https://chat.openai.com.