Pima Diabetes Prediction Report

Project Overview

This report details the implementation and evaluation of a machine learning model for predicting diabetes presence using the **"Pima_Diabetes.csv"** dataset. The dataset consists of diagnostic measurements for Pima Indian women, with the goal of predicting whether or not a patient has diabetes (binary classification). The model employed is a **Decision Tree Classifier**, and its performance is assessed using accuracy and AUC (Area Under the Curve) across both simple train/test splits and k-fold cross-validation.

Methodology

The project was implemented in Python using **NumPy**, **pandas**, and **scikit-learn**. The workflow involved three key experiments:

1. Data Loading & Preprocessing

- The dataset was read from **"Pima_Diabetes.csv"**, where the final column represented the binary outcome (0 = no diabetes, 1 = diabetes).
- No explicit feature selection or normalization was performed during this run.
- Features included glucose level, BMI, insulin, blood pressure, and others.

2. Model Training & Evaluation Approaches

A. Simple Train/Test Split:

- The data was split into training (65%) and testing (35%) sets.
- Two variants of the Decision Tree Classifier were tested:
    - criterion='gini'
    - criterion='entropy'

B. 5-Fold Cross-Validation:

- Cross-validation was conducted with cv=5 to better estimate model generalizability.
- The model was trained and validated across multiple folds, with performance averaged across them.

3. Evaluation Metrics

- **Accuracy:** Measures the percentage of correct predictions.
- **AUC (Area Under ROC Curve):** Evaluates the classifier's ability to distinguish between classes.

- **Runtime:** Execution time for each experiment was recorded for performance comparison.

Results

Simple Train/Test Split with criterion='gini'

Run	Accuracy	AUC	Runtime (s)
1–5	0.714	0.689	~0.002

Simple Train/Test Split with criterion='entropy'

Run	Accuracy	AUC	Runtime (s)
1–5	0.714	0.689	~0.002

5-Fold Cross-Validation with criterion='gini'

Run	Accuracy (Mean ± SD)	AUC (Mean ± SD)	Runtime (s)
1	0.72 ± 0.06	0.77 ± 0.06	~0.017
2	0.73 ± 0.05	0.77 ± 0.04	~0.018
3	0.73 ± 0.05	0.77 ± 0.05	~0.018
4	0.72 ± 0.06	0.78 ± 0.04	~0.017
5	0.72 ± 0.05	0.76 ± 0.04	~0.017

Analysis

Accuracy (~71–73%)

- The model consistently achieved moderate accuracy across all evaluation strategies.
- Indicates reliable performance for binary classification but suggests room for improvement.

AUC (~0.69–0.78)

- AUC values imply the model has fair ability to discriminate between positive and negative diabetes cases.
- Performance was slightly better in cross-validation than in train/test splits, which aligns with expectations for improved generalization.

Stability

- Train/test results were extremely stable.
- Cross-validation introduced some variance, especially in accuracy, due to data splitting, but the model still remained within reasonable bounds.

Discussion

Strengths:

- **Stable and repeatable results** in simple splits.
- **Improved generalization** through cross-validation.
- Reasonable discrimination between classes as reflected by the AUC.

Limitations:

- Performance plateaued at ~71–73% accuracy and ~0.77 AUC.
- The Decision Tree model alone may not capture complex interactions among features.
- Lack of preprocessing or feature engineering may limit model capacity.

Recommendations

To improve the model's performance, the following steps are recommended:

1. **Feature Engineering:** Normalize inputs or create new derived features (e.g., age × BMI).
2. **Feature Selection:** Use correlation analysis or tree-based feature importance to reduce noise.
3. **Hyperparameter Tuning:** Experiment with max_depth, min_samples_leaf, and pruning.
4. **Try Ensemble Models:** Explore **Random Forest**, **Gradient Boosting**, or **XGBoost** for better generalization.
5. **Stratified Cross-Validation:** Maintain class proportions to improve stability in CV folds.

Conclusion

The Decision Tree Classifier demonstrated consistent and moderately accurate performance in predicting diabetes outcomes using the Pima Indian dataset. While the model was stable and reasonably discriminative, further improvements could be achieved through enhanced preprocessing, feature engineering, and advanced ensemble techniques. This project offers a solid foundation in model evaluation practices and highlights the importance of cross-validation for trustworthy results.