Diabetes Prediction using Classical Machine Learning Approaches

IIT Guwahati - Data Science & Machine Learning Project

By - Samiksha V. Wanjari

Abstract

This project develops and evaluates machine learning models for diabetes risk prediction using the Pima Indians Diabetes dataset. We employ two classical classification algorithms—Logistic Regression and Linear Discriminant Analysis (LDA)—to predict diabetes based on patient health metrics. The models achieve 81.3% AUC-ROC, demonstrating good discriminative ability. Feature importance analysis reveals Glucose and BMI as the most significant predictors, aligning with established medical knowledge. The project emphasizes proper data preprocessing, particularly handling invalid zero values in medical measurements, and comprehensive model evaluation using multiple metrics.

Keywords: Diabetes Prediction, Logistic Regression, Linear Discriminant Analysis, Medical Data Analysis, Classification

1. Introduction

1.1 Background

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels, affecting millions of people worldwide. Early detection and risk assessment are crucial for effective prevention and management. Machine learning techniques offer promising approaches for developing predictive models that can assist healthcare providers in identifying high-risk individuals.

1.2 Problem Statement

Traditional diabetes diagnosis relies on clinical assessment and laboratory tests, which may not always be readily available or cost-effective for screening large populations. There is a need for automated risk prediction systems that can identify individuals at high risk of developing diabetes based on readily available health metrics.

1.3 Objectives

The primary objectives of this project are:

1. To perform comprehensive exploratory data analysis on the Pima Indians Diabetes dataset
2. To preprocess medical data appropriately, handling missing values and ensuring data quality
3. To train and evaluate classification models using Logistic Regression and Linear Discriminant Analysis
4. To interpret model coefficients to understand feature importance and clinical significance
5. To assess model performance using multiple evaluation metrics

1.4 Scope

This project focuses on binary classification (diabetes vs. no diabetes) using classical machine learning approaches. The scope includes data preprocessing, model training, evaluation, and interpretation, but excludes advanced techniques such as deep learning or ensemble methods.

2. Dataset Description

2.1 Dataset Overview

The Pima Indians Diabetes dataset is a well-known benchmark dataset in machine learning, containing medical measurements from 768 female patients of Pima Indian heritage. The dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases.

Dataset Characteristics: - **Total Samples:** 768 - **Features:** 8 input features + 1 target variable - **Target Variable:** Outcome (0 = No Diabetes, 1 = Diabetes) - **Class Distribution:** 500 (65.1%) No Diabetes, 268 (34.9%) Diabetes

2.2 Feature Description

The dataset includes the following features:

1. **Pregnancies:** Number of times pregnant
2. **Glucose:** Plasma glucose concentration (mg/dL)
3. **BloodPressure:** Diastolic blood pressure (mmHg)
4. **SkinThickness:** Triceps skin fold thickness (mm)
5. **Insulin:** 2-Hour serum insulin (μU/mL)
6. **BMI:** Body Mass Index (kg/m^2)
7. **DiabetesPedigreeFunction:** Function representing genetic predisposition to diabetes
8. **Age:** Age in years

2.3 Data Quality Issues

Initial examination revealed that several features contain zero values, which are biologically impossible for medical measurements: - Glucose: 5 zeros (cannot be 0 mg/dL in living patients) - BloodPressure: 35 zeros (cannot be 0 mmHg) - SkinThickness: 227 zeros (cannot be 0 mm) - Insulin: 374 zeros (extremely rare, likely missing data) - BMI: 11 zeros (cannot be 0)

These zero values represent missing data rather than actual zero measurements, requiring appropriate preprocessing.

3. Exploratory Data Analysis

3.1 Data Overview

The dataset contains 768 records with 8 features and 1 target variable. All features are numerical, with no explicit missing values (NaN) in the original dataset. However, zero values in medical features indicate missing data.

3.2 Class Distribution

The target variable shows moderate class imbalance: - **No Diabetes (0):** 500 cases (65.1%) - **Diabetes (1):** 268 cases (34.9%)

This 2:1 ratio suggests the need for appropriate handling during model training, such as using class weights or stratified sampling.

3.3 Feature Distributions

Analysis of feature distributions reveals:

- **Right-skewed distributions:** Glucose, Age, Pregnancies
- **Approximately normal:** BMI, BloodPressure
- **Many zeros:** SkinThickness, Insulin (indicating missing data)

Box plots comparing feature distributions by outcome show clear differences for several features, particularly Glucose and BMI, which exhibit better class separation.

3.4 Correlation Analysis

Correlation analysis with the target variable (Outcome) reveals the following relationships:

1. **Glucose:** 0.467 (strongest positive correlation)
2. **BMI:** 0.293 (strong positive correlation)
3. **Age:** 0.238 (moderate positive correlation)
4. **Pregnancies:** 0.222 (moderate positive correlation)

5. **DiabetesPedigreeFunction:** 0.174 (weak positive correlation)
6. **Insulin:** 0.131 (weak positive correlation)
7. **SkinThickness:** 0.075 (very weak positive correlation)
8. **BloodPressure:** 0.065 (very weak positive correlation)

The correlation heatmap shows moderate correlation between Age and Pregnancies (0.544), which is expected as older women tend to have more pregnancies.

3.5 Key EDA Insights

Key Findings: - Glucose is the strongest predictor, with a correlation of 0.467 - BMI shows strong association with diabetes risk (0.293) - Features have vastly different scales (e.g., Insulin: 0-846 vs. DiabetesPedigreeFunction: 0.08-2.42), indicating the need for feature scaling - The dataset is moderately imbalanced, requiring appropriate handling during model training

4. Data Preprocessing

4.1 Handling Invalid Zero Values

Medical datasets often contain zero values that represent missing data rather than actual zero measurements. In this dataset, zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI are biologically impossible and must be treated as missing data.

Preprocessing Steps:

1. **Identification:** Identified 652 invalid zero values across 5 features
2. **Replacement:** Replaced zeros with NaN to properly mark missing values
3. **Imputation:** Imputed missing values using median (robust to outliers)
4. **Verification:** Confirmed all zeros and NaN values were handled

Imputation Values: - Glucose: 117.00 mg/dL - BloodPressure: 72.00 mmHg - SkinThickness: 29.00 mm - Insulin: 125.00 μU/mL - BMI: 32.30 kg/m²

4.2 Why Median Imputation?

Median imputation was chosen over mean imputation for several reasons:

1. **Robustness to outliers:** Median is not affected by extreme values
2. **Distribution preservation:** Better maintains the shape of the data distribution
3. **Skewed data:** Medical data often has skewed distributions, where median is more appropriate
4. **Clinical relevance:** Median represents a more typical value in medical contexts

4.3 Feature Scaling

Feature scaling is critical for Logistic Regression and LDA because:

1. **Algorithm requirements:** Both algorithms assume features are on similar scales
2. **Gradient descent:** Logistic Regression uses gradient descent, which converges faster with scaled features
3. **Distance calculations:** LDA uses distance-based calculations sensitive to feature scales
4. **Fair comparison:** Scaling ensures all features contribute equally to the model

StandardScaler was used to transform features to have mean=0 and standard deviation=1, ensuring fair treatment of all features.

5. Model Training

5.1 Data Splitting

The preprocessed dataset was split into training and test sets: - **Training set:** 614 samples (80%) - **Test set:** 154 samples (20%) - **Stratification:** Maintained class distribution in both sets using stratified splitting - **Random state:** 42 (for reproducibility)

5.2 Feature Scaling

Features were standardized using StandardScaler: - Fitted on training data only (prevents data leakage) - Applied to both training and test sets - Result: All features have mean≈0 and std≈1

5.3 Logistic Regression

Model Configuration: - Algorithm: Logistic Regression - Solver: LBFGS (Limited-memory BFGS) - Max iterations: 1000 - Class weights: Balanced (handles class imbalance) - Random state: 42

Training Process: The model was trained on scaled training data. The balanced class weights ensure that the minority class (diabetes) receives appropriate attention during training.

Key Features: - Interpretable coefficients - Probabilistic outputs - Handles class imbalance with class_weight='balanced'

5.4 Linear Discriminant Analysis (LDA)

Model Configuration: - Algorithm: Linear Discriminant Analysis - Solver: SVD (Singular Value Decomposition) - Shrinkage: None - Priors: Estimated from data

Training Process: LDA was trained on the same scaled training data. The SVD solver provides numerical stability and works well with scaled features.

Key Features: - Assumes normal distribution of features - Maximizes class separation - Provides discriminant function coefficients

6. Model Evaluation

6.1 Evaluation Metrics

Both models were evaluated using multiple metrics to provide comprehensive assessment:

1. **Accuracy:** Overall correctness of predictions
2. **Precision:** Proportion of positive predictions that are correct
3. **Recall (Sensitivity):** Proportion of actual positives correctly identified
4. **F1-Score:** Harmonic mean of precision and recall
5. **ROC-AUC:** Area under the ROC curve (threshold-independent)

6.2 Logistic Regression Performance

Test Set Results: - **Accuracy:** 73.38% - **Precision:** 60.32% - **Recall:** 70.37% - **F1-Score:** 64.96% - **ROC-AUC:** 81.26%

Confusion Matrix: - True Negatives (TN): 75 - False Positives (FP): 25 - False Negatives (FN): 16 - True Positives (TP): 38

Interpretation: Logistic Regression demonstrates good performance with balanced precision and recall. The model correctly identifies 70.37% of diabetes cases (recall) and is correct 60.32% of the time when predicting diabetes (precision).

6.3 LDA Performance

Test Set Results: - **Accuracy:** 70.13% - **Precision:** 59.09% - **Recall:** 48.15% - **F1-Score:** 53.06% - **ROC-AUC:** 81.26%

Confusion Matrix: - True Negatives (TN): 82 - False Positives (FP): 18 - False Negatives (FN): 28 - True Positives (TP): 26

Interpretation: LDA achieves similar AUC-ROC but shows lower recall (48.15%), meaning it misses more diabetes cases. However, it has better precision (59.09%) and fewer false positives.

6.4 Model Comparison

Side-by-Side Comparison:

Metric	Logistic Regression	LDA
Accuracy	73.38%	70.13%
Precision	60.32%	59.09%
Recall	70.37%	48.15%
F1-Score	64.96%	53.06%
ROC-AUC	81.26%	81.26%

Key Observations:

1. **Similar AUC-ROC:** Both models achieve 81.26% AUC-ROC, indicating comparable discriminative ability
2. **Recall difference:** Logistic Regression has significantly better recall (70.37% vs. 48.15%), making it better at catching diabetes cases
3. **Precision similarity:** Both models have similar precision (~60%), indicating similar false positive rates
4. **Overall performance:** Logistic Regression performs better overall, with higher accuracy and F1-score

6.5 ROC Curves

ROC curve analysis confirms that both models perform well above random chance (AUC = 0.5). The curves show good separation between classes, with both models achieving AUC-ROC of 0.813, indicating strong discriminative ability.

7. Model Interpretation

7.1 Logistic Regression Coefficients

The Logistic Regression coefficients, sorted by absolute value, provide insights into feature importance:

Rank	Feature	Coefficient	Absolute Value	Effect
1	Glucose	1.1834	1.1834	↑ Increases risk
2	BMI	0.7097	0.7097	↑ Increases risk
3	Pregnancies	0.3730	0.3730	↑ Increases risk
4	DiabetesPedigreeFunction	0.2877	0.2877	↑ Increases risk
5	Age	0.1864	0.1864	↑ Increases risk
6	Insulin	-0.0447	0.0447	↓ Decreases risk
7	BloodPressure	-0.0145	0.0145	↓ Decreases risk
8	SkinThickness	0.0139	0.0139	↑ Increases risk

Intercept: -0.2570

7.2 Feature Importance Analysis

Top 4 Most Important Features:

1. **Glucose (1.18)** - Most Important Predictor
2. **Clinical Significance:** Blood glucose is the primary diagnostic marker for diabetes
3. **Correlation:** 0.467 (strongest correlation with outcome)
4. **Interpretation:** A one standard deviation increase in glucose increases the log-odds of diabetes by 1.18
5. **Medical Relevance:** Elevated glucose directly reflects impaired insulin function, the core pathology of diabetes
6. **BMI (0.71)** - Second Most Important
7. **Clinical Significance:** Obesity is a major risk factor for Type 2 diabetes
8. **Correlation:** 0.293 (strong correlation)
9. **Interpretation:** Higher BMI increases diabetes risk
10. **Medical Relevance:** Excess adipose tissue promotes insulin resistance through inflammatory mechanisms
11. **Pregnancies (0.37)** - Third Most Important
12. **Clinical Significance:** Gestational diabetes history increases future diabetes risk
13. **Correlation:** 0.222 (moderate correlation)
14. **Interpretation:** More pregnancies associated with higher diabetes risk
15. **Medical Relevance:** Pregnancy-induced insulin resistance can unmask underlying β-cell dysfunction

16. **DiabetesPedigreeFunction (0.29)** - Fourth Most Important
17. **Clinical Significance:** Genetic predisposition is a non-modifiable risk factor
18. **Correlation:** 0.174 (weak but significant)
19. **Interpretation:** Stronger family history increases diabetes risk
20. **Medical Relevance:** Type 2 diabetes has strong genetic component (heritability ~30-70%)

7.3 Coefficient Interpretation

Positive Coefficients: - Increase the probability of diabetes - Larger values indicate stronger influence - Six features have positive coefficients

Negative Coefficients: - Decrease the probability of diabetes - Insulin and BloodPressure show weak negative effects - May indicate protective factors or measurement artifacts

Standardized Features: Since features were standardized, coefficients are directly comparable. The coefficient values represent the change in log-odds for a one standard deviation increase in the feature.

7.4 Clinical Validation

The coefficient ranking aligns perfectly with established medical knowledge: - Glucose is the primary diagnostic criterion (validated) - BMI reflects obesity-diabetes link (validated) - Pregnancy history is recognized risk factor (validated) - Genetic predisposition is well-documented (validated)

This alignment provides confidence in the model's clinical relevance and predictive validity.

8. Conclusion

8.1 Project Summary

This project successfully developed and evaluated machine learning models for diabetes risk prediction using the Pima Indians Diabetes dataset. The work demonstrates the application of classical machine learning approaches to medical data analysis, with emphasis on proper data preprocessing and comprehensive model evaluation.

8.2 Key Achievements

1. **Data Quality:** Identified and handled 652 invalid zero values representing missing data in medical measurements
2. **Model Performance:** Achieved 81.3% AUC-ROC with both Logistic Regression and LDA
3. **Feature Insights:** Identified Glucose and BMI as the most important predictors, aligning with medical knowledge
4. **Clinical Relevance:** Model coefficients validate established clinical risk factors

8.3 Main Findings

- **Best Performing Model:** Logistic Regression (73.4% accuracy, 81.3% AUC-ROC, 70.4% recall)
- **Most Important Feature:** Glucose (coefficient: 1.18, correlation: 0.467)
- **Data Quality Issue:** Zero values in medical features represent missing data, requiring appropriate preprocessing
- **Class Imbalance:** Moderate imbalance (65% no diabetes, 35% diabetes) handled using balanced class weights

8.4 Model Selection

Logistic Regression is recommended for this application because: - Higher overall accuracy (73.4% vs. 70.1%) - Better recall (70.4% vs. 48.1%) - catches more diabetes cases - Higher F1-score (65.0% vs. 53.1%) - better balance of precision and recall - Interpretable coefficients for clinical understanding

8.5 Practical Implications

The models can assist healthcare providers in: - Early diabetes risk assessment - Prioritizing patients for screening - Identifying high-risk individuals for preventive intervention - Supporting clinical decision-making with data-driven insights

9. Limitations and Future Scope

9.1 Limitations

1. **Dataset Size:** 768 samples may limit model generalization to larger populations
2. **Population Specificity:** Pima Indian population may not generalize to other ethnic groups
3. **Feature Engineering:** Limited to original features; could benefit from interaction terms (e.g., Glucose × BMI)
4. **Model Complexity:** Linear models may miss non-linear relationships between features
5. **Temporal Aspects:** Dataset lacks temporal information; cannot capture disease progression
6. **External Validation:** Models tested only on internal test set; external validation needed

9.2 Future Work

1. **Advanced Models:** Explore ensemble methods (Random Forest, XGBoost) or neural networks for potentially better performance
2. **Feature Engineering:** Create interaction features (Glucose × BMI, Age × Pregnancies) to capture non-linear relationships
3. **Hyperparameter Tuning:** Optimize model parameters using grid search or Bayesian optimization
4. **External Validation:** Test models on independent datasets from different populations

5. **Clinical Integration:** Develop user-friendly interfaces (e.g., web applications) for healthcare providers
6. **Cost-Sensitive Learning:** Incorporate misclassification costs (false negatives more costly in medical context)
7. **Explainable AI:** Implement SHAP values or LIME for enhanced model interpretability
8. **Longitudinal Studies:** Incorporate temporal data to track disease progression

9.3 Recommendations

For practical deployment: - Validate models on diverse populations before clinical use - Incorporate domain expert knowledge in feature selection - Develop risk stratification guidelines based on model outputs - Ensure compliance with medical device regulations - Maintain model performance monitoring and periodic retraining

References

1. Smith, J. W., et al. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261-265.
2. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
3. American Diabetes Association. (2023). Standards of Medical Care in Diabetes. *Diabetes Care*, 46(Supplement 1).
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
5. James, G., et al. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Appendix

A. Dataset Source

- **Dataset:** Pima Indians Diabetes Dataset
- **Source:** UCI Machine Learning Repository / National Institute of Diabetes and Digestive and Kidney Diseases
- **License:** Public domain

B. Software and Tools

- **Programming Language:** Python 3.x

- **Libraries:** pandas, numpy, scikit-learn, matplotlib, seaborn
- **Development Environment:** Jupyter Notebook

C. Code Availability

All code, including preprocessing scripts, model training scripts, and evaluation modules, is available in the project repository.

Project completed for IIT Guwahati Data Science Course

This report demonstrates the application of classical machine learning approaches to medical data analysis, emphasizing proper methodology, comprehensive evaluation, and clinical interpretation.

Project completed for IIT Guwahati Data Science Course