DIABETES PREDICTION USING CLASSICAL MACHINE LEARNING APPROACHES

A Comparative Analysis of Logistic Regression and Linear Discriminant Analysis

Author: Abhishek Kumar

Roll/ID: IITG_DS_25011354

Date: 13 January, 2026

Institution: Masai DS Program Capstone Project

ABSTRACT

Abstract: This research presents a comparative analysis of classical machine learning approaches for the early detection of Type 2 Diabetes Mellitus. Using the Pima Indians Diabetes Database (PIDD), we evaluated Logistic Regression and Linear Discriminant Analysis (LDA) models to predict diabetes onset based on diagnostic measurements. The study methodology involved rigorous data preprocessing, including medical range filtering and median imputation, reducing the dataset to 382 high-quality samples. Key results demonstrate that both models achieved an identical accuracy of **79%** and an Area Under the Curve (AUC) of **0.85**. However, Logistic Regression demonstrated superior clinical utility with a higher recall of **76.92%** compared to LDA's 61.54%, making it the preferred model for screening purposes. The final model was deployed as an interactive Streamlit web application, providing real-time risk assessment and lifestyle guidelines. This work highlights the efficacy of interpretable, classical machine learning in medical diagnostics where transparency is paramount.

1. INTRODUCTION & CLINICAL CONTEXT

1.1 Diabetes Mellitus: Global Health Challenge 🩺

Diabetes Mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels, which over time leads to serious damage to the heart, blood vessels, eyes, kidneys, and nerves. Type 2 diabetes is the most common form, largely driven by excess body weight and physical inactivity.

- **Complications:** Uncontrolled diabetes can result in nephropathy (kidney failure), neuropathy (nerve damage), cardiovascular disease, and retinopathy (blindness).
- **Early Detection:** Timely intervention is critical. Prediabetes and early-stage Type 2 diabetes can often be managed or reversed through lifestyle changes if detected early.
- **ML Role:** Machine learning algorithms offer a non-invasive, cost-effective screening tool to identify high-risk individuals using routine clinical data, reducing the burden on healthcare systems.

1.2 Research Motivation

While complex "black-box" models like Deep Neural Networks often achieve high accuracy, they lack interpretability, which is crucial in clinical settings. Physicians need to understand *why* a model predicts high risk.

- **Interpretability:** Classical models like Logistic Regression provide clear coefficients that align with medical knowledge.

- **Transparency:** Essential for building trust with healthcare practitioners and patients.
- **Deployment:** Lightweight classical models are easier to deploy in resource-constrained environments.

2. PROBLEM STATEMENT & OBJECTIVES

2.1 Problem Formulation

The task is defined as a **binary classification problem** where the objective is to predict the binary class variable `Outcome`:

- **1 (Positive):** Patient tests positive for diabetes.
- **0 (Negative):** Patient tests negative for diabetes.

The prediction relies on routinely collected clinical variables such as Glucose, BMI, Age, and Blood Pressure. The focus is strictly on transparency and ethical deployment, ensuring the model serves as a decision-support tool rather than a diagnostic replacement.

2.2 Research Objectives

- ✅ **EDA:** Perform comprehensive Exploratory Data Analysis with strict medical validation.
- ✅ **Comparison:** Systematically compare Logistic Regression vs. Linear Discriminant Analysis (LDA).
- ✅ **Data Quality:** Handle medical implausibilities (zero values, outliers) through domain-informed preprocessing.
- ✅ **Recall Priority:** Prioritize sensitivity (recall) to minimize missed diagnoses (false negatives).
- ✅ **Deployment:** Develop an interactive Streamlit application for real-time inference.

3. DATASET DESCRIPTION

3.1 Pima Indians Diabetes Database (PIDD) 📊

Source: National Institute of Diabetes and Digestive and Kidney Diseases.
Population: 768 female patients of Pima Indian heritage, age ≥21 years.

Features (8 Predictors + 1 Target):

1. **Pregnancies:** Number of times pregnant (0-17)
2. **Glucose:** Plasma glucose concentration (mg/dL) after 2-hour oral glucose tolerance test
3. **BloodPressure:** Diastolic blood pressure (mm Hg)
4. **SkinThickness:** Triceps skin fold thickness (mm)
5. **Insulin:** 2-Hour serum insulin (μU/mL)
6. **BMI:** Body mass index (weight in kg/(height in $m)^2$)
7. **DiabetesPedigreeFunction:** Genetic predisposition indicator (0.078-2.42)
8. **Age:** Age in years (21-81)

9. **Outcome:** Class variable (0 or 1)

3.2 Dataset Characteristics BEFORE Preprocessing

- **Total entries:** 768
- **Class distribution:** 500 non-diabetic (65.1%), 268 diabetic (34.9%)
- **Data Issues:** Missing values are represented as zeros in biologically impossible columns (Glucose, BloodPressure, SkinThickness, Insulin, BMI).

3.3 Dataset Characteristics AFTER Filtering

After applying rigorous medical range filtering, the dataset consists of **382 high-quality samples**. The statistical summary is as follows:

Feature	Mean	Std	Min	25%	50% (Median)	75%	Max
Pregnancies	3.34	3.22	0	1	2	5	17
Glucose	123.33	30.42	71	99.25	120	143.75	198
BloodPressure	70.93	11.61	40	62.5	70	78	110
SkinThickness	29.07	10.38	7	21	29	36	60
Insulin	158.45	119.28	16	78.25	127.5	191.75	846
BMI	32.96	6.57	18.2	28.43	33.2	36.98	57.3
DPF	0.522	0.334	0.085	0.270	0.452	0.687	2.329
Age	30.96	10.25	21	23	27	36	81

4. METHODOLOGY

4.1 Data Preprocessing Pipeline

4.1.1 Medical Range Filtering (Outlier Removal)

To ensure data integrity, physiological plausibility was enforced using the following ranges:

- **Pregnancies:** 0-20
- **Glucose:** 70-200 mg/dL
- **BloodPressure:** 40-140 mm Hg
- **SkinThickness:** 5-100 mm
- **Insulin:** 15-900 µU/mL
- **BMI:** 10-60 kg/m^2
- **DiabetesPedigreeFunction:** 0.05-3.0
- **Age:** 18-100 years

Result: Dataset reduced from 768 → 382 samples (50.3% retained).
Rationale: Removing physiologically impossible measurements improves model robustness.

4.1.2 Zero Value Imputation

Strategy: Median Imputation.
Rationale: Zero values in Glucose, BP, Skin, Insulin, and BMI were treated as measurement errors. Median imputation was chosen over mean due to the skewed nature of these distributions.

4.1.3 Feature Scaling

Method: StandardScaler (z-score normalization).
Formula: `z = (x - μ) / σ`
Reason: Standardizing features to zero mean and unit variance is essential for the convergence of Logistic Regression and LDA.

4.1.4 Train-Test Split

- **Split Ratio:** 80% Training, 20% Testing
- **Total Samples:** 382
- **Training Set:** 305 samples
- **Test Set:** 77 samples
- **Stratification:** Applied to maintain class distribution in splits.

4.2 Exploratory Data Analysis (EDA)

Visual analysis revealed several key insights:

- **Feature Distributions:** Glucose showed a bimodal distribution, suggesting two underlying subpopulations (healthy vs. diabetic). Age and Pregnancies were right-skewed.
- **Feature Correlation Heatmap:**
  - Glucose vs Outcome: **0.51** (Strongest Predictor)
  - Age vs Outcome: 0.36
  - BMI vs Outcome: 0.24
  - Inter-feature correlation: Glucose-Insulin (0.58), SkinThickness-BMI (0.66).
- **Pairplot Analysis:** Scatter plots of Glucose vs BMI color-coded by outcome showed a clear separation, with higher glucose levels strongly associated with positive outcomes.
- **Box Plots:** Outcome 1 (Diabetic) showed a significantly higher median Glucose (~145 mg/dL) compared to Outcome 0 (~110 mg/dL), confirming Glucose as a primary discriminator.

4.3 Model Selection & Training

4.3.1 Logistic Regression

Algorithm: Binary classification using the sigmoid function to model probability.
Hyperparameters: `max_iter=1000`, `class_weight='balanced'`, `solver='lbfgs'`, `random_state=42`.

Feature Coefficients (Exact):

1. **Glucose:** 1.162929 (Dominant predictor)
2. **BMI:** 0.439254
3. **DiabetesPedigreeFunction:** 0.429931
4. **Age:** 0.275627
5. **Pregnancies:** 0.173074
6. **SkinThickness:** 0.163036
7. **BloodPressure:** 0.078242
8. **Insulin:** 0.061324

Interpretation: Glucose has >2.6x stronger effect than BMI. The top 3 features (Glucose, BMI, DPF) account for the majority of the predictive power.

4.3.2 Linear Discriminant Analysis (LDA)

Algorithm: Finds a linear combination of features that characterizes or separates two or more classes of objects or events. It assumes Gaussian distributions and equal covariance matrices for classes.
Advantages: Computationally efficient and less prone to overfitting on small datasets.

5. PERFORMANCE EVALUATION 📈

5.1 Metrics Selection

- **Recall/Sensitivity:** The critical metric for this medical screening task to avoid missing positive cases (False Negatives).
- **Precision:** To minimize unnecessary stress from False Alarms.
- **F1-Score:** Harmonic mean for a balanced view.
- **AUC-ROC:** Threshold-independent measure of discriminative ability.

5.2 Logistic Regression Results

Confusion Matrix:

```
Predicted | 0 | 1 | Actual 0 | 41 | 10 | 1 | 6 | 20 |
```

Detailed Classification Report:

```
precision recall f1-score support Class 0 0.87 0.80 0.84 51 Class 1 0.6
0.77 0.79 0.78 77 weighted avg 0.80 0.79 0.80 77
```

Key Logistic Regression Metrics:
Accuracy: **79.22%** | Recall (Class 1): **76.92%** | Precision (Class 1): **66.67%**

5.3 Linear Discriminant Analysis (LDA) Results

Confusion Matrix:

```
Predicted | 0 | 1 | Actual 0 | 45 | 6 | 1 | 10 | 16 |
```

Detailed Classification Report:

```
precision recall f1-score support Class 0 0.82 0.88 0.85 51 Class 1 0.7
0.77 0.75 0.76 77 weighted avg 0.79 0.79 0.79 77
```

Key LDA Metrics:
Accuracy: **79.22%** | Recall (Class 1): **61.54%** | Precision (Class 1): **72.73%**

5.4 Comparative Analysis

Both models achieved an identical AUC of **0.85**, indicating excellent discriminative ability. However, their operating points differ.

Metric	Logistic Regression	LDA	Winner
Accuracy	79.22%	79.22%	Tie
Precision (Class 1)	66.67%	72.73%	LDA
Recall (Class 1)	76.92%	61.54%	Logistic Regression
F1-Score (Class 1)	71.43%	66.67%	Logistic Regression
False Negatives	6	10	Logistic Regression

Clinical Recommendation

✅ **Logistic Regression is PREFERRED**.

In medical screening, missing a diabetic patient (False Negative) is significantly more dangerous than raising a false alarm. Logistic Regression detects **76.92%** of positive cases compared to LDA's 61.54%, missing only 6 cases versus LDA's 10. The interpretability of coefficients further supports its adoption.

6. MODEL INTERPRETABILITY

6.1 Feature Importance Analysis

The Logistic Regression coefficients reveal the model's decision logic:

- **Glucose (1.163):** By far the most critical factor. A 1 standard deviation increase in Glucose roughly triples the odds of a diabetes prediction.
- **BMI (0.439):** Obesity remains a major risk factor, holding the second highest weight.
- **DiabetesPedigreeFunction (0.430):** Genetic history is nearly as important as BMI.
- **Age (0.276):** Risk increases modestly with age.

6.2 Clinical Validation

The model learned patterns strictly from data that align with established medical science (American Diabetes Association guidelines). The strong weighting of Glucose and BMI confirms the model is not relying on spurious correlations.

7. DEPLOYMENT: STREAMLIT APPLICATION 🚀

7.1 Application Architecture

The final model was deployed using **Streamlit**, a Python web framework.

- **Backend:** Scikit-learn (Model Inference), Pandas (Data Handling).
- **Frontend:** Streamlit (User Interface).
- **Pipeline:** Real-time median imputation and scaling for user inputs.

7.2 Application Features

- **Interactive Inputs:** Sliders for all 8 clinical features (e.g., Glucose 50-200 mg/dL).
- **Real-time Prediction:** Instantly calculates probability.
- **Risk Categorization:**
  - 🟢 **Low Risk (<30%):** Advice: Maintain healthy lifestyle (Walk, Yoga).
  - 🟡 **Moderate Risk (30-60%):** Advice: Monitor glucose, reduce sugar.
  - 🔴 **High Risk (>60%):** Advice: Consult doctor immediately.
- **Cultural Context:** Diet guidelines tailored to Indian context (e.g., advising against samosas, suggesting millets/roti).

7.3 Ethical Considerations

Disclaimers Included:

- "⚠️ This model estimates Type 2 diabetes risk for adults (18+)."
- "These suggestions are for general awareness only 📘. They do NOT replace professional medical advice 🩺."
- "This application is for educational purposes only and is not a diagnostic tool."

8. LIMITATIONS & FUTURE WORK

8.1 Current Limitations

- **Data Size:** Rigorous filtering reduced the dataset to 382 samples, potentially limiting generalizability.
- **Demographics:** Trained solely on Pima Indian females; validation on males and other ethnicities is required.
- **Features:** Lacks critical clinical markers like HbA1c and lipid profiles.
- **Trade-off:** Logistic Regression accepts more False Positives (10) to maximize Recall.

8.2 Future Enhancements

- **Ensemble Methods:** Test Random Forest or XGBoost for potentially higher accuracy (at the cost of interpretability).
- **Data Augmentation:** Incorporate larger, multi-ethnic datasets.
- **SHAP Values:** Integrate SHAP plots into the app for personalized explanations of *why* a specific user was flagged as high risk.

9. CONCLUSION

9.1 Key Findings

- ✅ Successfully developed interpretable ML models for diabetes prediction.
- ✅ Logistic Regression achieved **79.22% accuracy** and **76.92% recall**.
- ✅ Glucose was identified as the strongest predictor (Coefficient: 1.163).
- ✅ Deployed a user-friendly Streamlit application with risk-stratified guidance.

9.2 Clinical Impact

The developed system serves as a cost-effective, non-invasive early screening tool. By prioritizing recall, it minimizes the risk of missing diabetic patients, while its interpretability fosters trust among medical professionals.

9.4 Final Recommendation

Logistic Regression is recommended for deployment. Its superior ability to detect positive cases (higher recall) and transparent decision-making process make it the ideal candidate for a clinical decision support system in this context.

REFERENCES

1. National Institute of Diabetes and Digestive and Kidney Diseases. Pima Indians Diabetes Database.
2. American Diabetes Association. Standards of Medical Care in Diabetes—2024. Diabetes Care.
3. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. JMLR.
4. Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics.
5. Hosmer, D.W., Lemeshow, S. (2000). Applied Logistic Regression. Wiley.

6. World Health Organization. Global Report on Diabetes (2016).