PREDICTING DIABETES OUTCOMES: A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS ON THE PIMA INDIANS' DATASET

MAGALY ANTUANET CAROLINA DIAZ GUARNIZ

ISOM 835 – PREDICTIVE ANALYTICS AND MACHINE LEARNING:

TERM PROJECT

BOSTON, MASSACHUSETTS, 2025

1. INTRODUCTION & DATASET DESCRIPTION

Diabetes is a chronic metabolic disease that occurs when the body either does not produce enough insulin or cannot effectively use the insulin it produces. This leads to elevated levels of blood glucose, which, over time, can result in serious damage to the heart, blood vessels, eyes, kidneys, and nerves (World Health Organization, 2023). Early detection and intervention are essential for preventing complications and improving health outcomes.

This project applies predictive analytics and machine learning techniques to identify individuals at risk of developing diabetes using health-related features. The goal is to develop accurate and interpretable predictive models that can help healthcare professionals make informed decisions and prioritize preventive care efforts.

For this analysis, the Pima Indians Diabetes Dataset is used. This dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases and focuses on female patients of Pima Indian heritage, aged 21 years and older, residing in Arizona. It contains 768 records with 8 independent variables representing various health metrics and 1 binary target variable indicating diabetes diagnosis (0 = no diabetes, 1 = diabetes).

The dataset includes the following variables:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration (2-hour oral glucose tolerance test)
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skinfold thickness (mm)
- Insulin: 2-hour serum insulin (mu U/ml)
- BMI: Body mass index
- DiabetesPedigreeFunction: Score that estimates diabetes risk based on family history

- Age: Age in years
- Outcome: Class variable (0 = non-diabetic, 1 = diabetic)

This dataset is widely used in machine learning education and research because it represents a real-world medical prediction challenge with measurable health indicators. The data was accessed from Kaggle: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

This dataset is suitable for this project because it provides real clinical measurements from a high-risk population, allowing for the development and evaluation of predictive models in a healthcare context.

PROJECT OBJECTIVES

1. Identify the most important health-related predictors of diabetes among Pima Indian women.
2. Develop and compare the performance of at least two predictive models for diabetes diagnosis.
3. Generate insights that could support healthcare professionals in screening and preventive strategies.

2. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis was conducted to understand the structure, distribution, and relationships within the Pima Indians Diabetes dataset. The analysis focused on detecting patterns, identifying potential data issues, and gaining insights that could inform the modeling process.

2.1. DATA OVERVIEW:

The dataset's structure was explored using .head() and .describe(). No missing values were detected using .isnull().sum(), but several features had zero values where biologically impossible (e.g., Glucose = 0), indicating potential missing data that needs handling.

2.2. FEATURE DISTRIBUTIONS (HISTOGRAMS):

The histograms display the frequency distribution of each variable in the dataset, offering insights into their shapes and central tendencies:

- Glucose: The distribution is slightly right-skewed, with most values between 80 and 150, but a long tail toward higher glucose levels.
- BloodPressure: The distribution is approximately normal but shows a slight skew to lower values, with most readings around 60–80 mm Hg.
- BMI: The BMI variable is right-skewed, with a concentration around 30, and some individuals exhibiting higher BMI values.
- Insulin and SkinThickness: Both show strong right skewness, with many low or zero values and a few very high readings.
- Pregnancies: This variable is highly skewed toward lower numbers; most women had between 0 and 4 pregnancies.
- DiabetesPedigreeFunction: The distribution is right-skewed, with most values clustering below 0.5.
- Age: The age distribution is skewed left, with most participants between 20 and 40 years old.

2.3. BOXPLOTS:

The boxplot visualization provides an overview of the distribution and presence of outliers across all variables in the dataset. It shows that several variables contain extreme values beyond the typical range (outliers), especially in:

- Insulin: Significant number of high outliers, suggesting unusually large insulin levels in some individuals.
- SkinThickness: Outliers are present above the upper whisker, indicating higher-than-typical skinfold thickness measurements.
- Pregnancies: Some high outliers reflect women with unusually high numbers of pregnancies.
- Glucose and BMI: A few mild outliers, but the distributions are more concentrated.

2.4. CORRELATION MATRIX (HEATMAP):

The correlation matrix shows the strength and direction of linear relationships between the variables in the dataset:

- Glucose has the strongest positive correlation with Outcome (diabetes diagnosis), with a correlation coefficient around 0.47, suggesting higher glucose levels are associated with higher diabetes risk.
- BMI and DiabetesPedigreeFunction also show moderate positive correlations with Outcome, indicating that higher BMI and a stronger family history increase the likelihood of diabetes.
- Age shows a weak positive correlation with Outcome, suggesting older individuals may have a slightly higher risk.

- Most other variables (e.g., BloodPressure, SkinThickness, Insulin) exhibit very weak or negligible correlations with Outcome.

No strong negative correlations were observed between any variables and the diabetes outcome.

2.5. OUTCOME DISTRIBUTION

The outcome variable shows the distribution of diabetes diagnosis in the dataset:

- Approximately 65% of the individuals are non-diabetic (Outcome = 0)
- Around 35% are diabetic (Outcome = 1)

This indicates that the dataset is slightly imbalanced, with more non-diabetic cases than diabetic cases.

2.6. SCATTER PLOT: GLUCOSE VS. BMI, COLORED BY OUTCOME

- Individuals diagnosed with diabetes (red points) tend to have higher glucose levels and moderate to high BMI compared to non-diabetic individuals.
- There's a visible separation at higher glucose levels where most individuals are diabetic, suggesting glucose is a strong predictor.
- However, BMI alone does not create a clear separation; high BMI is present in both diabetic and non-diabetic groups, meaning BMI may be less discriminative on its own.

This visualization reinforces the idea that glucose is a key factor in predicting diabetes and that BMI might act as a supporting factor rather than a standalone indicator.

2.7. KEY INSIGHTS:

- Glucose, Insulin, BMI, and Age appear to be the most important predictors based on their distribution and correlation with the outcome.
- Several features contain biologically implausible zero values, suggesting they should be treated as missing values.
- The dataset shows a moderate imbalance between diabetic and non-diabetic cases, which may require adjustments during modeling.
- Visual patterns indicate meaningful differences in certain features (Glucose, BMI) between diabetic and non-diabetic individuals.

3. PREPROCESSING

3.1. MISSING VALUES:

The dataset contained zero values in features like Glucose, BloodPressure, SkinThickness, Insulin, and BMI, which are not physiologically possible. These zeros were treated as missing values and replaced with the median of each feature.

3.2. OUTLIER DETECTION AND HANDLING:

Outliers were visualized using boxplots. For extreme outliers, values were capped at the 1st and 99th percentiles to reduce their impact without removing data.

3.3. FEATURE SCALING:

Since many machine learning models are sensitive to feature scale, all numeric features were standardized using StandardScaler to have mean = 0 and standard deviation = 1.

3.4. TRAIN-TEST SPLIT:

The dataset was split into training (80%) and testing (20%) sets to evaluate model performance on unseen data.

4. BUSINESS QUESTIONS

- **What key patient characteristics are associated with a higher likelihood of having diabetes?**

  Identifying these factors can help healthcare providers target interventions and prioritize screening efforts.

- **How accurately can we predict the likelihood of diabetes using patient health metrics?**

  Understanding the model's predictive performance helps assess its potential use in clinical decision support systems.

- **Can predictive modeling assist healthcare providers in identifying high-risk individuals for early intervention?**

  By flagging patients at higher risk, preventive care measures can be implemented to reduce future complications and costs.

5. MODELING

Two predictive models were applied — Logistic Regression and Random Forest — to predict whether a patient is diabetic based on medical and demographic data.

Table 1: *Performance Comparison:*

Model	Accuracy	F1 Score	AUC
Logistic Regression	0.75	0.64	0.82
Random Forest	0.74	0.64	0.83

- Both models performed similarly in accuracy and F1 score, correctly classifying diabetes status approximately 74–75% of the time while maintaining a balanced precision-recall tradeoff.
- Random Forest achieved a slightly higher AUC (0.83) compared to Logistic Regression (0.82), indicating a marginally stronger ability to distinguish diabetic from non-diabetic patients.
- Logistic Regression offers greater interpretability through its coefficients, which allows stakeholders to understand how features like glucose or BMI influence predictions.
- Random Forest, as a non-linear model, is better suited for capturing more complex relationships among variables.

6. INSIGHTS & RECOMMENDATIONS

6.1. KEY FINDINGS

- Glucose and BMI as Primary Predictors: Both models consistently identified glucose levels and BMI as the strongest predictors of diabetes risk. This supports the widely accepted medical understanding of how elevated glucose and BMI contribute to increased diabetes susceptibility.

- Model Insights: Logistic Regression demonstrated that each unit increase in glucose or BMI raised the likelihood of diabetes, offering straightforward insights for healthcare professionals. Meanwhile, the Random Forest model uncovered more intricate, non-linear relationships, suggesting it may be better suited for capturing complex patterns in the data.

6.2. RECOMMENDATIONS

- Targeted Screening: Focus on individuals with high glucose and BMI levels for early diagnosis and intervention.
- Early Intervention Programs: Prioritize patients with higher risks for lifestyle changes or medical management programs.
- Clinical Decision Support: Implement Logistic Regression in electronic health records to provide healthcare professionals with transparent, actionable insights.
- Operational Use: While these models can assist in clinical decision-making, they should be viewed as complementary tools that support but do not replace human expertise.

6.3. LIMITATIONS

- The absence of crucial variables, such as family history, diet, or physical activity, limits the models' accuracy and generalizability.
- Features like insulin and skin thickness had missing or zero values, which were imputed, potentially reducing the precision of the models.
- The models were evaluated on the same dataset; external validation on new, diverse patient data is necessary to ensure real-world applicability.

- While Random Forest had slightly better accuracy, it is less interpretable than Logistic Regression, which might pose challenges for clinicians requiring clear, understandable model outputs.

6.4. CONCLUSION

The predictive models demonstrate that it is feasible to classify diabetes risk using the available clinical data with about 75% accuracy. Glucose and BMI were consistently the strongest predictors, providing actionable insights for targeted screening and early intervention. However, further data collection and validation are recommended to enhance predictive performance and ensure applicability across broader patient populations.

7. ETHICS & INTERPRETABILITY

When using predictive models in healthcare, ethical considerations are critical. The dataset used in this project includes only female patients of Pima Indian heritage, which limits its generalizability. Applying the model to other populations without proper validation could lead to biased or unfair outcomes. Additionally, features like insulin or skin thickness had missing or imputed values, which may affect fairness if not handled carefully. Relying on biased data can unintentionally reinforce health disparities.

Interpretability also plays a key role in ethical use. Logistic Regression offers transparency, allowing stakeholders to see how each feature impacts the prediction. This supports accountability and informed decision-making in clinical environments. In contrast, while Random Forest performs slightly better, its complexity makes it harder to explain. For ethical and practical reasons, models deployed in healthcare should balance accuracy with

transparency, ensuring that patients and providers can understand and trust the system's decisions.

REFERENCES

Dua, D., & Graff, C. (2017). *Pima Indians Diabetes Database*. UCI Machine Learning Repository. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

World Health Organization. (2024). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes

APPENDIX (CODE, VISUALS)

Appendix A: Code for Data Preprocessing

```
# Replace zeros with NaN
cols_to_replace_zeros = ['Glucose', 'BMI', 'BloodPressure',
'SkinThickness', 'Insulin']
df[cols_to_replace_zeros] = df[cols_to_replace_zeros].replace(0, pd.NA)

# Fill missing values
df.fillna(df.median(), inplace=True)

# Scaling
scaler = StandardScaler()
columns_to_scale = ['Pregnancies', 'Glucose', 'BloodPressure',
'SkinThickness', 'Insulin', 'BMI', 'Age']
df[columns_to_scale] = scaler.fit_transform(df[columns_to_scale])

# Split
X = df.drop('Outcome', axis=1)
y = df['Outcome']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Preview data
df.head()
```

Appendix B: Code for Model Building

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, f1_score, roc_auc_score,
confusion_matrix, classification_report

# Initialize and train
lr_model = LogisticRegression()
lr_model.fit(X_train, y_train)

# Predict
y_pred_lr = lr_model.predict(X_test)
y_prob_lr = lr_model.predict_proba(X_test)[:,1]

# Evaluate
accuracy_lr = accuracy_score(y_test, y_pred_lr)
f1_lr = f1_score(y_test, y_pred_lr)
auc_lr = roc_auc_score(y_test, y_prob_lr)
print("Logistic Regression Metrics:")
print("Accuracy:", accuracy_lr)
print("F1 Score:", f1_lr)
print("AUC:", auc_lr)
```

Appendix C: Code for Model Evaluation

```
from sklearn.ensemble import RandomForestClassifier

# Initialize and train
rf_model = RandomForestClassifier(random_state=42)
rf_model.fit(X_train, y_train)

# Predict
y_pred_rf = rf_model.predict(X_test)
y_prob_rf = rf_model.predict_proba(X_test)[:,1]

# Evaluate
accuracy_rf = accuracy_score(y_test, y_pred_rf)
f1_rf = f1_score(y_test, y_pred_rf)
auc_rf = roc_auc_score(y_test, y_prob_rf)
print("Random Forest Metrics:")
print("Accuracy:", accuracy_rf)
print("F1 Score:", f1_rf)
print("AUC:", auc_rf)
```