Analysis and Evaluation of Machine Learning Models for Diabetes Risk Prediction

Venkat Karthik Poreddy Sabin Adhikari Siddhartha Phuyal Sonam Gurung

University of South Dakota

Introduction

- Chronic Condition: Diabetes is a long-term health condition that impacts millions globally.
- Significance of Early Diagnosis: Early detection and prediction of diabetes are crucial for effective disease management and reducing healthcare expenses.
- Role of Classification Models: Classification models are widely used for predicting diabetes. They leverage health data to make predictions.
- Advantages Over Traditional Methods: These models can identify complex patterns in datasets. They are often more effective than traditional statistical techniques.

Project Goals

- Develop machine learning models to predict diabetes risk.
- Understand the impact of key features on predictions.
- Balance accuracy and computational efficiency.
- Ensure model interpretability for clinical use.

Literature Review

1. Role of Machine Learning in Diabetes Prediction
   - Leverages structured datasets (e.g., glucose levels, age, BMI, family history).
   - Identifies complex patterns beyond human capability, enhancing diagnostic accuracy.
2. Classification Models Overview
   - Logistic Regression: Simple and interpretable. Effective on datasets like PIMA Indian Diabetes Dataset (PIDD).
   - Decision Trees: Highly interpretable, using branching logic. Perform well on small datasets but prone to overfitting.
   - Ensemble Models (Random Forest Gradient Boosting): Combine multiple models for robust predictions. Random Forest reduces overfitting. Gradient Boosting (e.g., XGBoost) captures nonlinear interactions.
3. Data and Challenges
   - Class Imbalance: Diabetes datasets often have fewer positive cases. Techniques like SMOTE address imbalance.
   - Feature Selection: Key predictors: glucose levels, BMI, family history. Use techniques like PCA to remove redundant features.
   - Evaluation Metrics: Accuracy, precision, recall, F1-score, AUC-ROC. Prioritize recall and F1-score in imbalanced datasets.

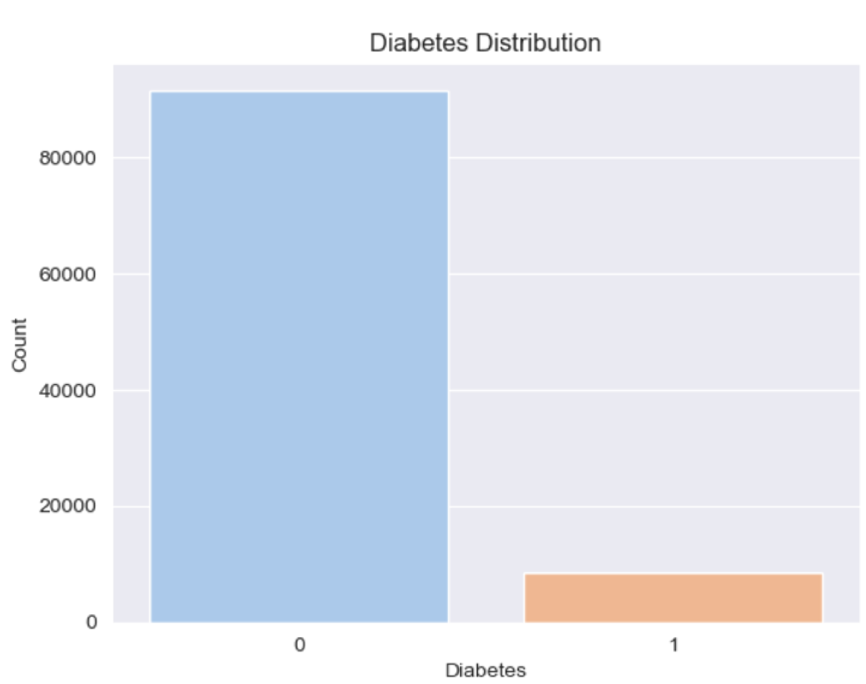

Figure 1. Class imbalance between people with diabetes (1) and people without (0)

Methodology

- Data Pre-processing:
  - Data Cleaning: Removed entries with missing values and "Other" gender.
  - Categorical Encoding: One-hot encoded "gen der" and "smoking history".
  - Feature Scaling: Standardized continuous variables (age, BMI, HbA1c, blood glucose level).
- Model Implementation

  The implemented models include Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost. The dataset was split into (80-20) percent testing sets using stratified sampling to ensure the class distribution remained consistent across both subsets.
- Model Evaluation
  - Precision
  - Recall
  - F1-score
  - ROC-AUC
  - Confusion matrix
  - Cross validation accuracy and cross validation standard deviation

Models	Precision	Recall	F1 Score	ROC AUC	CV Accuracy
Decision Tree	0.7030	0.7435	0.7227	0.8570	0.9519
Random Forest	0.9430	0.6906	0.7973	0.9632	0.9699
XGBoost	0.9562	0.6929	0.8035	0.9783	0.9710
LightGBM	0.9655	0.6912	0.8056	0.9792	0.9717
CatBoost	1.0	0.6729	0.8045	0.9580	0.9718

Figure 2. Confusion Matrix

- Proposed Framework
  - The framework consists of the following main steps:
    - Data Source: The dataset containing various health features (age, gender, BMI, HbA1c level, etc.).
    - Data Preprocessing: Includes cleaning the data, encoding categorical variables, handling missing values, and scaling continuous features.
    - Model Training: Training multiple machine learning models such as Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost.
    - Evaluation Metrics: Metrics like precision, recall, F1-score, ROC-AUC, and cross-validation are used to evaluate the models.

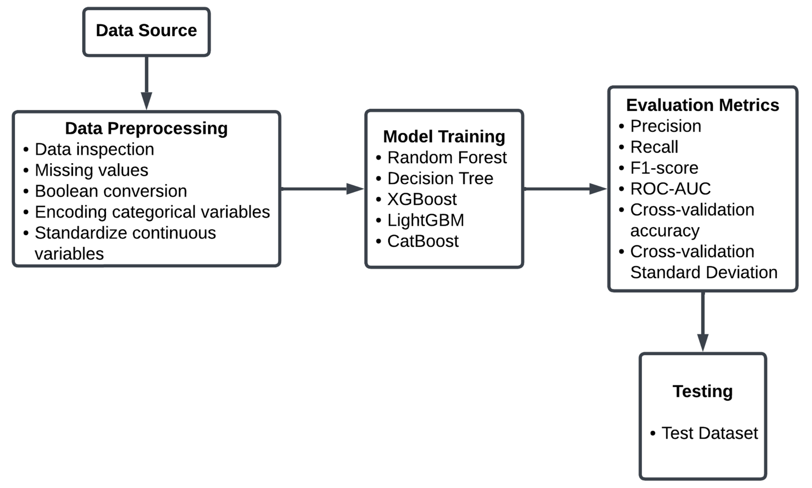

Figure 3. Proposed Framework

Evaluation Results

1. Evaluation Metrics Used:
   Models were assessed using precision, recall, F1-score, ROC-AUC, confusion matrix, cross-validation accuracy, and standard deviation.
2. Performance of Models:
   - Decision Tree: Worst overall but had the highest recall.
   - Random Forest XGBoost: Performed consistently well across all metrics.
   - CatBoost: Excelled in precision and CV-accuracy but lower recall affected its F1-score.
   - LightGBM: Best overall with the highest F1-score and ROC-AUC, crucial for classification evaluation.
3. Comparison of Top Models:
   LightGBM outperformed others but CatBoost was competitive with strong recall and similar F1-score.
4. Feature Importance Across Models: HbA1c level and blood glucose level were most important for Decision Tree, Random Forest, XGBoost, and CatBoost.

Models	Precision	Recall	F1 Score	ROC AUC	CV Accuracy
Decision Tree	0.7030	0.7435	0.7227	0.8570	0.9519
Random Forest	0.9430	0.6906	0.7973	0.9632	0.9699
XGBoost	0.9562	0.6929	0.8035	0.9783	0.9710
LightGBM	0.9655	0.6912	0.8056	0.9792	0.9717
CatBoost	1.0	0.6729	0.8045	0.9580	0.9718

Figure 4. Evaluation Result Table

Conclusion

1. LightGBM's Strength: LightGBM was the most effective model, showcasing the highest F1-score and ROC AUC, demonstrating its ability to balance precision and recall effectively for diabetes prediction.
2. CatBoost's Competitiveness: CatBoost closely followed LightGBM, excelling in precision and cross-validation accuracy, though its lower recall slightly impacted its overall performance.
3. Consistency of Random Forest and XGBoost: Both models provided reliable and consistent results across all evaluation metrics, making them strong alternatives in predictive modeling.
4. Decision Tree's Limitation: The Decision Tree model underperformed in most metrics, though its high recall highlighted its potential in prioritizing positive cases.
5. Feature Insights: HbA1c levels and blood glucose were identified as critical predictors across most models, emphasizing their importance in diabetes risk assessment.

Recommendations

1. Feature Engineering: Explore techniques like Principal Component Analysis (PCA) to uncover new predictive variables or derive more informative features from existing ones.
2. Address Class Imbalance: Use methods like SMOTE (Synthetic Minority Oversampling Technique) to improve model performance, particularly in recall and F1-score.
3. Hyperparameter Optimization: Apply advanced tuning techniques such as grid search or Bayesian optimization to achieve significant performance enhancements.
4. Enhance Generalization: Focus on strategies that ensure the models perform effectively across varied patient populations and healthcare settings.