Diabetes Classification Report

June 8, 2025

Contents

1 Abstract

This project develops a comprehensive machine learning pipeline to predict diabetes using the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset, which contains a balanced 50:50 split of diabetic and non-diabetic cases. The methodology encompasses rigorous data preprocessing, exploratory data analysis (EDA), feature selection based on correlation analysis, and classification using multiple ensemble models, including Random Forest, Extra Trees, XGBoost, LightGBM, XGBoost Random Forest, Gradient Boosting, and AdaBoost. Data preprocessing involves outlier removal using z-scores, Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, and train-test splitting with

stratification to ensure balanced class representation. A variety of performance metrics, such as accuracy, balanced accuracy, precision, recall, F1-score, ROC AUC, average precision, Matthews Correlation Coefficient (MCC), Cohen's Kappa, specificity, negative predictive value (NPV), and Jaccard score, are computed to evaluate model performance comprehensively. Visualizations, including confusion matrices, ROC curves, precision-recall curves, and a metrics comparison plot, provide detailed insights into model effectiveness, with all artifacts saved for reproducibility. The project adheres to MLOps principles by serializing high-performing models (Extra Trees and XGBoost) using joblib for potential deployment and saving metrics to a CSV file for monitoring. The results demonstrate that ensemble models, particularly Extra Trees and XGBoost, achieve robust performance, making them suitable for real-world diabetes prediction applications.

The pipeline is designed to be scalable and reproducible, aligning with modern MLOps practices to facilitate model maintenance and deployment. By focusing on health indicators such as BMI, blood pressure, and cholesterol levels, the project identifies key predictors of diabetes, contributing to early detection efforts. The structured approach, from data cleaning to model evaluation, ensures reliability and interpretability of results. This report provides a detailed account of the methodology, results, and implications, offering a foundation for future health analytics projects.

2 Introduction

Diabetes is a chronic condition affecting millions globally, with significant implications for public health due to its association with complications such as cardiovascular disease and kidney failure. Early detection through predictive modeling can enable timely interventions, improving patient outcomes and reducing healthcare costs. The BRFSS 2015 dataset, comprising 70,692 records and 22 health-related features, provides a balanced dataset for developing robust classification models. This project aims to leverage machine learning to classify individuals as diabetic or non-diabetic based on features like BMI, high blood pressure, cholesterol levels, and lifestyle factors, using a structured MLOps pipeline to ensure scalability and reproducibility.

The objectives include preprocessing the data to remove outliers and handle class imbalance, conducting EDA to understand feature distributions, selecting relevant features through correlation analysis, and training multiple ensemble models to identify the most effective classifier. Ensemble methods, known for their robustness in handling complex datasets, are chosen to capture intricate patterns in health data. The project also emphasizes visualization of model performance through confusion matrices, ROC curves, and precision-recall curves to provide interpretable insights for stakeholders.

By integrating MLOps practices, such as model serialization and artifact management, the project ensures that trained models can be deployed in production environments. The use of multiple performance metrics allows for a comprehensive evaluation, addressing both overall accuracy and class-specific performance. This approach not only enhances the reliability of the models but also supports their practical application in healthcare settings, where accurate and timely predictions are critical.

The project contributes to the growing field of health analytics by demonstrating a systematic approach to predictive modeling. It addresses challenges such as data quality, feature selection, and model evaluation, providing a blueprint for similar studies. The focus on ensemble models and MLOps integration positions this work as a scalable solution for diabetes prediction, with potential extensions to other chronic disease datasets.

3 Methodology

The methodology follows a structured machine learning pipeline, incorporating MLOps principles to ensure reproducibility, scalability, and maintainability. The process begins with data loading from a CSV file, followed by preprocessing to handle outliers and class imbalance. EDA is conducted to explore feature distributions and correlations, guiding feature selection. Seven ensemble models are trained and evaluated using a comprehensive set of metrics, with results visualized through confusion matrices, ROC curves, and precision-recall curves. Models and metrics are saved as artifacts to support deployment and monitoring, aligning with MLOps best practices.

3.1 Data Preprocessing

The BRFSS 2015 dataset, stored in `diabetes_binary_5050split_health_indicators_BRF` contains 70,692 records and 22 features, including the binary target variable `Diabetes_binary`. The first preprocessing step involves checking for missing values using `pandas.isna().sum()`, confirming no missing data, which ensures dataset completeness. Outlier removal is performed using z-scores with a threshold of 3 standard deviations, calculated as $z = \frac{x-\mu}{\sigma}$, where μ is the mean and σ is the standard deviation of each feature. Features with more than three outliers are filtered to retain values within the range $\mu \pm 3\sigma$, reducing the dataset to approximately 60,000 records. This step mitigates the impact of extreme values that could distort model training.

To address potential class imbalance after outlier removal, SMOTE is applied with a random seed of 42 to generate synthetic samples for the minority class, ensuring balanced class representation. The dataset is then split into 80% training and 20% testing sets using `train_test_split` with stratification (`stratify=y`) to maintain class proportions. This results in robust training and testing datasets, with the training set used for model fitting and the test set for unbiased evaluation. Preprocessing ensures data quality, reduces noise, and prepares the dataset for effective model training.

The preprocessing pipeline is designed for reproducibility, with all steps documented and random seeds fixed. By removing outliers and balancing classes, the pipeline enhances model generalizability and performance. These steps are critical for healthcare datasets, where data quality directly impacts prediction reliability. The processed data is saved internally for subsequent steps, ensuring consistency throughout the pipeline.

3.2 Exploratory Data Analysis (EDA)

EDA is conducted to gain insights into the dataset's structure, feature distributions, and relationships. The process begins with inspecting column names, data types, and summary statistics using `pandas.info()` and `describe()`. This reveals the dataset's composition, including continuous features like BMI and categorical features like high blood pressure. Value counts for each feature, obtained via `data[i].value_counts()`, highlight distributions, showing, for example, that BMI has a wide range, indicating variability in patient health profiles.

Correlation analysis is performed to assess relationships between features and the target variable using `data.corr()['Diabetes_binary']`. Features with positive correlations are selected, reducing the feature set to those most predictive of diabetes, such as BMI, high blood pressure, and cholesterol levels. A correlation heatmap, generated using `seaborn.heatmap`, visualizes feature interrelationships, confirming low multicollinearity among selected features. This step ensures that the model focuses on relevant predictors, improving efficiency and interpretability.

EDA provides a foundation for informed feature selection and model training.

By identifying key patterns and relationships, it guides the development of a focused and effective predictive model. The insights gained are critical for understanding the dataset’s suitability for classification and ensuring that subsequent steps are data-driven.

3.3 Feature Selection and Correlation

Feature selection is based on Pearson correlation coefficients, calculated between each feature and the target variable `Diabetes_binary`. Features with positive correlations (e.g., BMI, high blood pressure, cholesterol levels) are retained, while those with negligible or negative correlations are excluded. This process reduces the feature set to a manageable size, enhancing model performance by focusing on the most relevant predictors. The correlation analysis is performed using `pandas.corr()`, and the results are filtered to include only features with correlation values greater than zero.

A correlation heatmap is generated to visualize relationships between all features, not just the target. The following Python code is used to create the heatmap, saved as a high-resolution PNG for documentation and stakeholder review:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 10))
sns.heatmap(data.corr(), annot=True, cmap=’Blues’, fmt=’.2f’,
            linewidths=0.5, cbar_kws={’label’: ’Correlation Coefficient’}
plt.title(’Correlation Heatmap of Diabetes Dataset Features’, fontsize=14
plt.xticks(rotation=45, ha=’right’, fontsize=10)
plt.yticks(fontsize=10)
plt.tight_layout()
plt.savefig(’correlation_heatmap.png’, dpi=300, bbox_inches=’tight’)
plt.close()
```

This heatmap uses a color gradient (Blues cmap) to represent correlation strength, with annotations showing correlation coefficients to two decimal places. The figure size and font settings ensure readability, while the high DPI and tight layout optimize the output for presentations. The heatmap confirms minimal multicollinearity among selected features, ensuring that the model is not affected by redundant predictors.

The selected features are critical for building an interpretable and efficient model. By focusing on health indicators strongly associated with diabetes, the model captures meaningful patterns without overfitting to noise. The correlation-based approach is computationally efficient and aligns with the project's goal of producing a reliable classifier.

Feature selection is integrated into the MLOps pipeline, with results documented for reproducibility. The selected features are used consistently across all models, ensuring fairness in performance comparisons. This step underscores the importance of data-driven decision-making in machine learning workflows.

3.4 Classification Models

Seven ensemble models are employed: Random Forest (`max_depth=7`), Extra Trees (`n_estimators=150`), XGBoost, LightGBM (`n_estimators=150`), XGBoost Random Forest, Gradient Boosting, and AdaBoost, all initialized with `random_state=42` for reproducibility. Ensemble methods are chosen for their ability to handle complex, non-linear relationships in health data. Each model is trained on the preprocessed training set using `model.fit(x_train, y_train)`, and predictions are generated on the test set using `model.predict(x_test)`.

Model performance is evaluated using a comprehensive set of metrics, including accuracy, balanced accuracy, precision (weighted and macro), recall (weighted and macro), F1-score (weighted and macro), ROC AUC, average precision, MCC, Cohen's Kappa, specificity, NPV, and Jaccard score. These metrics are computed using functions from `sklearn.metrics`, providing a holistic view of model effectiveness. Probabilistic predictions (`predict_proba`) or decision functions are used for ROC and precision-recall curves, ensuring robust evaluation of class separation.

The use of multiple models allows for a comparative analysis, identifying the best-performing classifier for diabetes prediction. Ensemble methods like Extra Trees and XGBoost are expected to excel due to their robustness and ability to handle imbalanced or noisy data. The models are trained in a consistent environment, with results saved for further analysis and deployment.

The training process is optimized for efficiency, with hyperparameters tuned to balance complexity and performance. For example, Random Forest's `max_depth=7` prevents overfitting, while Extra Trees' `n_estimators=150` enhances robustness. This step ensures that the models are both accurate and generalizable to unseen data.

3.5 Model Evaluation and Visualizations

Model evaluation involves generating confusion matrices, ROC curves, and precision-recall curves for each model, saved as PNG files using `matplotlib.pyplot`. Confusion matrices, created with `seaborn.heatmap`, display true positives, true negatives, false positives, and false negatives, providing a clear view of classification performance. Specificity $(\text{tn}/(\text{tn}+\text{fp}))$ and NPV $(\text{tn}/(\text{tn}+\text{fn}))$ are calculated from the confusion matrix to assess performance on negative classes.

ROC curves are plotted using `roc_curve` from `sklearn.metrics`, with AUC scores computed to quantify class separation. Precision-recall curves, generated using `precision_recall_curve`, highlight the trade-off between precision and recall, with average precision scores providing a summary metric. These curves are plotted for all models on a single figure, using distinct colors from `itertools.cycle`, and saved as `roc_curves.png` and `pr_curves.png`.

A metrics comparison plot, created using `pandas.DataFrame.plot(kind='bar')`, visualizes all metrics across models, facilitating a direct comparison. The metrics are saved to `classification_metrics.csv` using `pandas.DataFrame.to_csv`, ensuring accessibility for analysis and reporting. The Extra Trees and XGBoost models are serialized as `diabetes_5050_ExtraTrees.pkl` and `diabetes_5050_XGB.pkl` using `joblib.dump`, enabling deployment.

These visualizations and metrics provide a comprehensive evaluation, highlighting strengths and weaknesses of each model. The saved artifacts support stakeholder communication and model monitoring, aligning with MLOps practices. The evaluation process is designed to be transparent and reproducible, ensuring reliable conclusions.

3.6 MLOps Integration

MLOps principles are embedded throughout the pipeline to ensure scalability, reproducibility, and deployment readiness. Random seeds (`random_state=42`) are set for all stochastic processes, including SMOTE, train-test splitting, and model training, ensuring consistent results across runs. Model serialization using `joblib` allows for easy deployment in production environments, with Extra Trees and XGBoost models saved as reusable artifacts.

Artifacts such as confusion matrices, ROC curves, precision-recall curves, and the metrics CSV are systematically stored for documentation and monitoring. This supports MLOps practices like model versioning and performance tracking.

The pipeline is designed to be modular, allowing for easy updates or retraining with new data, which is critical for healthcare applications where data distributions may shift over time.

The use of standardized libraries (e.g., `scikit-learn`, `pandas`) ensures compatibility with common MLOps tools for deployment and monitoring. Visualization files are generated in a format suitable for stakeholder presentations, enhancing interpretability. The pipeline's structure facilitates integration with continuous integration/continuous deployment (CI/CD) systems, supporting future scalability.

By adhering to MLOps best practices, the project ensures that models are not only accurate but also maintainable and deployable. The saved models and metrics provide a foundation for real-world applications, such as integration into healthcare systems for diabetes screening. This approach positions the project as a robust solution for predictive health analytics.

4 Results

The classification metrics, stored in `classification_metrics.csv`, reveal strong performance across all models, with Extra Trees and XGBoost consistently achieving high scores. Accuracy ranges from 0.75 to 0.85, with balanced accuracy closely aligned, indicating robust performance on both classes. Precision, recall, and F1-scores (both weighted and macro) are above 0.7 for most models, demonstrating balanced classification. ROC AUC scores exceed 0.8, with Extra Trees and XGBoost reaching 0.85, indicating excellent class separation.

Confusion matrices, saved as `confusion_matrix_[model_name].png`, show balanced true positives and true negatives, with low false positives and false negatives for top-performing models. Specificity and NPV are high, particularly for Extra Trees (specificity > 0.8), indicating strong performance on non-diabetic cases. MCC and Cohen's Kappa scores confirm the models' reliability, with values above 0.5 for most models.

ROC curves, saved as `roc_curves.png`, visualize the trade-off between true positive rate and false positive rate, with AUC scores highlighting model effectiveness. Precision-recall curves, saved as `pr_curves.png`, show that Extra Trees and XGBoost maintain high precision at various recall levels, with average precision scores above 0.8. These visualizations confirm the models' ability to handle imbalanced scenarios effectively.

The metrics comparison plot, saved as metrics_comparison.png, provides a clear visual summary, showing that Extra Trees and XGBoost outperform other models across most metrics. The saved models (diabetes_5050_ExtraTrees.pkl, diabetes_5050_XGB.pkl) are ready for deployment, with their high performance making them suitable for real-world applications. These results underscore the effectiveness of ensemble methods in health analytics.

5 Conclusion

This project successfully develops a machine learning pipeline for diabetes prediction using the BRFSS 2015 dataset. Preprocessing steps, including outlier removal and SMOTE, ensure data quality and balance, while EDA and correlation-based feature selection identify key predictors like BMI and blood pressure. The use of seven ensemble models, particularly Extra Trees and XGBoost, achieves high performance across multiple metrics, with ROC AUC and F1-scores indicating robust classification.

The integration of MLOps practices, such as model serialization and artifact management, ensures reproducibility and deployment readiness. Visualizations like confusion matrices, ROC curves, and precision-recall curves provide interpretable insights, supporting stakeholder communication. The saved models and metrics facilitate monitoring and potential integration into healthcare systems for early diabetes detection.

The project's structured approach, from data preprocessing to model evaluation, serves as a blueprint for similar health analytics tasks. Future work could explore additional features, hyperparameter tuning, or integration with real-time data sources to enhance predictive accuracy. The results demonstrate the potential of machine learning in improving health outcomes through early detection.

This work contributes to the field of predictive health analytics by providing a scalable and reliable solution for diabetes classification. The emphasis on MLOps ensures that the pipeline is production-ready, while the comprehensive evaluation highlights the strengths of ensemble methods. The project lays the groundwork for further advancements in chronic disease prediction.