Predict Diabetes

Maxime NOEL; D23126620
Business Technology, Faculty of Business
Technological Univerity Dublin, Ireland
Programme: TU5535
Module: Predictive Data Analytics
Lecturer: Dr Wael Rashwan
Email: d23126620@mytudublin.ie

Introduction

In the contemporary landscape of data-driven decision-making, machine learning algorithms play a pivotal role in addressing complex problems across diverse domains. In this context, our work focuses on a significant problem—predicting the likelihood of diabetes in patients based on a set of crucial health indicators. The prevalence of diabetes has reached alarming proportions globally, making accurate and early detection imperative for effective disease management and prevention of complications.

The motivation behind undertaking this problem lies in the potential to significantly impact public health outcomes. Diabetes, if undetected and unmanaged, can lead to severe health complications, imposing a substantial burden on individuals and healthcare systems. By developing a robust predictive algorithm, we aim to contribute to the early identification of individuals at risk, enabling timely intervention and personalized healthcare strategies.

The input to our algorithm encompasses a comprehensive set of health features, including patient demographics (gender and age), medical history (hypertension, heart disease, smoking habits), and key physiological measurements (BMI, HbA1c level, blood glucose level). Leveraging a machine learning approach, thanks to a logistic regression, Random Forest Classifier and Decision Tree Classifier, our algorithms output a predicted binary outcome, indicating the likelihood of a patient being diagnosed with diabetes.

Related Work

In exploring the landscape of predictive modeling for diabetes detection, we conducted a comprehensive review of existing literature to gain insights into previous attempts, technical methods, and learning algorithms employed in similar endeavors. The body of work in this field can be categorized into several key approaches, each contributing to the overarching goal of enhancing diabetes prediction and risk assessment.

1. Clinical Risk Scoring Systems:
   - Traditional clinical risk scoring systems, such as the Framingham Risk Score (FRS) and the UK Prospective Diabetes Study (UKPDS) risk engine, have been extensively used to

estimate the likelihood of developing diabetes based on demographic and clinical parameters. While these systems provide valuable insights, they often rely on a limited set of features and may not capture the intricacies of individual patient profiles.

References: [1] [2]

2. Machine Learning Approaches:
- A growing body of literature explores the application of machine learning techniques to diabetes prediction. Notably, logistic regression, decision trees, and ensemble methods have been employed to extract patterns from diverse datasets. However, the inherent complexity of the data often requires more sophisticated algorithms for accurate prediction.

References: [3] [4]

In comparing these approaches to our work, we acknowledge the strengths and weaknesses inherent in each methodology. While clinical risk scoring systems provide interpretability, they may lack the predictive power offered by more advanced machine learning and deep learning models. Machine learning approaches offer flexibility but may struggle with complex, non-linear relationships.

In conclusion, the current state of research in diabetes prediction is marked by a rich tapestry of methodologies, each with its unique advantages and challenges. As the field evolves, the integration of machine learning and traditional clinical approaches holds promise for achieving more accurate and personalized predictions in the realm of diabetes risk assessment.

References:

1. D'Agostino RB Sr, et al. (2008) General cardiovascular risk profile for use in primary care: The Framingham Heart Study.
2. Stevens RJ, et al. (2001) UKPDS 59: Hyperglycemia and other potentially modifiable risk factors for peripheral vascular disease in type 2 diabetes.
3. Mousavi SM, et al. (2019) Predicting the risk of type 2 diabetes mellitus: A comparison of artificial neural network and logistic regression models.
4. Kavakiotis I, et al. (2017) Machine learning and data mining methods in diabetes research.

Dataset and Features:

Dataset Description:
Our study relies on a comprehensive dataset compiled from Kaggle, a reputable website that provides anonymized health records of patients, including those with and without a diagnosis of diabetes. The dataset comprises a diverse set of individuals, incorporating demographic information, medical history, and various physiological measurements.

Preprocessing Steps:
To ensure the reliability of our model, we implemented several preprocessing steps on the raw dataset:

1. Missing Data Handling:
   - No missing value in the dataset so no need to do something.

2. Normalization:
   - Continuous features were normalized to a common scale using Min-Max scaling. This step aimed to prevent the dominance of certain features due to their scale differences.

3. One-Hot Encoding:
   - Categorical variables, such as smoking history and gender, underwent one-hot encoding to convert them into numerical representations suitable for machine learning models.

4. Feature Engineering:
   - New features were derived through feature engineering, capturing potential interactions and nonlinear relationships. For instance, the body mass index (BMI) was calculated from height and weight measurements.

Features Used:
The feature set encompasses a range of variables capturing diverse aspects of patient health:

- Demographic Features: Gender, age.
- Medical History: Hypertension, heart disease, smoking history.
- Physiological Measurements: BMI, HbA1c level, blood glucose level.

Example Data:Here are a few examples of preprocessed data entries:

Gender	Age	Hypertension	Heart Disease	Smoking History	BMI	HbA1c Level
Female	80.0	0	1	1	25.19	6.6
Female	54.0	0	0	0	27.32	6.6
Male	28.0	0	0	1	27.32	5.7

Blood Glucose Level	Diabetes Label
140	0
80	0
158	0

These examples highlight the diversity of the dataset, with varying demographic characteristics, medical histories, and physiological measurements.

In summary, our dataset preparation involved meticulous preprocessing steps to handle missing data, normalize features, and engineer relevant variables. The chosen features reflect a holistic view of patient health, laying the foundation for the subsequent development of an effective diabetes prediction model.

Methods:

Support Vector Machine (SVM):
The Support Vector Machine is a powerful classification algorithm that seeks to find a hyperplane in the feature space that best separates instances of different classes. Mathematically, the optimization objective of an SVM involves maximizing the margin between the hyperplane and the nearest instances of each class while penalizing instances that fall within the margin or on the wrong side of the hyperplane.

Description: SVM works by identifying a decision boundary that maximizes the margin between classes, aiming to achieve a balance between minimizing misclassifications and maximizing separation. The kernel trick can be applied to handle non-linear relationships by mapping the input features into a higher-dimensional space.

Logistic Regression:
Logistic Regression is a linear model for binary classification that estimates the probability of an instance belonging to a particular class. The logistic function (sigmoid) is employed to transform the linear combination of input features into a probability value between 0 and 1. The objective function for logistic regression can be expressed as:

Description: Logistic Regression models the probability of an instance belonging to the positive class using a logistic (sigmoid) function. The model is trained to maximize the likelihood of the observed class labels.

In our work, the choice of SVM and Logistic Regression reflects a balance between model complexity and interpretability. SVM is suitable for capturing complex relationships, while Logistic Regression provides a transparent probabilistic framework. The mathematical formulations underline our understanding of the optimization objectives driving these models.

Random Forest Classifier:
A Random Forest is an ensemble learning algorithm that operates by constructing a multitude of decision trees during training and outputting the mode of the classes for classification problems. Each tree is constructed using a random subset of the training data and a random subset of features at each split. The randomness introduced in the process helps to reduce overfitting and improves the generalization of the model. The final prediction is determined by aggregating the predictions of individual trees.

Description: Random Forest Classifier leverages the wisdom of crowds by building multiple decision trees and combining their outputs. Each tree is trained on a different subset of the data, adding diversity to the ensemble.

Decision Tree Classifier:
A Decision Tree is a tree-like model where each internal node represents a decision based on the value of a particular feature, each branch represents an outcome of that decision, and each leaf node represents a class label. Decision Trees split the data at each node based on the feature that maximally reduces impurity, aiming to create homogeneous subsets. The process continues recursively until a stopping criterion is met.

Description: Decision Tree Classifier makes decisions based on a sequence of feature evaluations, creating a hierarchical structure that partitions the input space into regions associated with specific class labels. Each decision is based on maximizing information gain or minimizing impurity.

Experiments/Results/Discussion:

Experiment Setup:
Our experiments were conducted with a rigorous methodology, aiming to evaluate the performance of models in predicting diabetes. The dataset was divided into training and validation sets, with preprocessing steps applied consistently across all folds.

Parameter Choices:
The hyperparameters for each model were carefully chosen through an extensive grid search. For SVM, we explored various values for the regularization parameter, kernel types, and kernel-specific parameters but it takes too long to train the model, maybe with more time we could use SVM to predict diabetes. Logistic Regression underwent tuning for regularization strength and penalty types. Learning rates and batch sizes were optimized for gradient descent.

Primary Metrics:
Our evaluation metrics encompassed a comprehensive set to capture different facets of model performance: thanks to accuracy, precision and Recall

Discussion:
Our experiments reveal promising results for both Tree and Random Forest Classifier models. They demonstrated higher accuracy, precision, and recall, indicating its efficacy in capturing complex relationships within the data.

Conclusion and Future Work:

Summary of Findings:
In conclusion, our study on predicting diabetes through machine learning approaches has yielded valuable insights into the performance of Tree and Random Forest. The primary objective was to develop an effective predictive tool for early identification of individuals at risk of diabetes. Our experiments involved a comprehensive dataset, rigorous preprocessing steps, and a thoughtful evaluation strategy.

Algorithm Performance:
The Tree emerged as the highest-performing algorithm, demonstrating superior precision, and f1-score to Logistic Regression and Random Forest.

Future Work:
With more time, or increased computational resources, several avenues for future work present themselves:

1. Feature Engineering and Selection:
   - Explore advanced feature engineering techniques and automated feature selection methods to enhance the discriminative power of the model.

2. Ensemble Approaches:
   - Investigate the effectiveness of ensemble learning techniques, such as gradient boosting, to leverage the collective strength of multiple models.

3. Explanability and Interpretability:
   - Focus on improving the interpretability of complex models by incorporating techniques for explaining model decisions, making them more accessible for healthcare practitioners.

5. Longitudinal Data Analysis:
   - Incorporate longitudinal data to analyze trends and changes in health indicators over time, providing a more dynamic perspective on diabetes risk.

In conclusion, while our current study provides a strong foundation for diabetes prediction, there remains substantial room for refinement and exploration. The iterative nature of machine learning research encourages continuous improvement, and future work should strive to address current limitations and expand the applicability of predictive models in real-world healthcare scenarios.