Term Project Milestone 5

Diabetic Analysis Through Machine Learning

Modeling Prediction

DSC 630

Prof. Andrew Hua

Carlos Cano

Carlos Cantu

Alexander Hamedaninia

Table of Contents

Abstract

The Diabetes Prediction Dataset sourced through Kaggle provides a valuable opportunity to leverage Machine Learning Techniques to predicting diabetes based on patients' medical and demographic information. This paper will propose an outline of the key aspects of our project and its proposed executability using Data Science to extract, modify as needed, process, apply models and finally interpret the results that is relevant to Healthcare Professionals as well as Data Scientist.

Introduction

Dataset Background

The Diabetes prediction dataset encompasses a diverse set of features, including age, gender, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. These features offer a comprehensive view of patients' health profiles, making it an ideal dataset for developing predictive models. The ultimate goal is to assist healthcare professionals in identifying individuals at risk of diabetes and to contribute to the development of personalized treatment plans.

Types of Models & Rationale

Logistic Regression:

Our plan is to start with Logistic Regression due to its simplicity and interpretability as the data is fairly complete in its nature. Logistic Regression is also well-suited for binary classification problems. Using the various features of a patient's health status, we can predict with a certain amount of accuracy a patient's diabetes status (positive or negative).

Random Forest:

To capture non-linear relationships and interactions among features, we intend to explore Random Forest models. The ensemble nature of Random Forest can enhance predictive accuracy with a robustness to outliers. Random Forest models also provide feature importance, allowing us to determine which factors in a patient's health contribute more significantly to diabetes.

Neural Networks:

For a more complex analysis, we will consider neural networks. Neural networks, particularly deep learning models, have shown success in handling intricate patterns within data. This will allow us to capture nuanced relationships within the dataset. We can utilize its feature learning capabilities to determine which features are the most relevant when predicting diabetes in patients.

Linear Regression:

For its simplicity and interpretability, linear regression provides coefficients for each predictor, allowing interpretation of the relationship between predictors and the target variable. Linear regression is also a computationally efficient modeling process, an advantageous aspect for our dataset containing 100,000 inputs.

K-Nearest Neighbors (KNN) Classifier:

KNN is a non-parametric and instance-based classification algorithm used for classification. Similar to linear regression, KNN is easy to understand and implement. It does not make strong assumptions about the underlying data distribution. There is also no training phase – it can store the data and make predictions based on similarity measures during inference.

Evaluation Strategy

To evaluate the performance of our models, we will employ the following metrics:

- *Accuracy:* Overall correctness of predictions.
- *Precision:* Proportion of true positives among predicted positives.

- *Recall:* Proportion of true positives identified among all actual positives.
- *F1 Score:* Harmonic mean of precision and recall, balancing false positives and false negatives.

Additionally, we will use a stratified cross-validation approach to ensure robustness in our model evaluation. Based on these model and evaluation and approach we will be able to explore the efficacy of our modeling techniques as it relates to our data.

Hopeful Takeaways

As we continue to work in unison, we will rely heavily on these takeaway principals to gain continued insight to our data driven model.

Our primary learning objectives include:

1. *Model Performance:* Assess the effectiveness of different models in predicting diabetes.
2. *Feature Importance:* Identify key factors contributing to diabetes prediction.
3. *Optimization Techniques:* Explore and implement Hyperparameter Tuning to enhance model overfitting & performance.
4. *Interpretability:* Understand the interpretability of models, especially important for healthcare applications.

Project Risk

- *Data Bias:* The dataset may contain biases that could lead to unfair predictions, especially regarding underrepresented groups.

- *Overfitting:* Models might perform well on training data but fail to generalize to new, unseen data. Hyperparameter tuning should be applied for optimal model configuration.

Ethical Implications

- *Privacy Concerns:* Ensure the responsible handling of sensitive medical data to protect patient privacy.
- *Equity*: Address potential bias to ensure fair and equitable predictions for all demographic groups.
- *Data Selection & Manipulation:* Ensure that data that is removed is notated and does not skew the meaningfulness of the original data.

Contingency Plan

In the event that our original project plan encounters challenges, our contingency plan involves:

1. *Model Iteration:* Iteratively refine models based on feedback and performance metrics.
2. *Additional Feature Engineering:* Explore additional features or transformations to enhance model interpretability and performance.
3. *Consultation with Expert:* Seek guidance Professor Hua on our rationale and decision-making processes.

Challenges

1. *Continued Teammate Communication:* Finding a regular schedule to discuss weekly topics.

2. *Group Unanimous Voting:* Allowing for group decision processes to occur naturally.
3. *Conflict Resolution:* Understanding rationale and through processes of other teammates.

Opportunities

1. *Learning Opportunities:* Observing other teammates coding and skills all the while gaining insight.
2. *Coding Refining:* Finding new ways from different teammate experiences how to refine coding.
3. *Teammate Insights:* Combining professional and personal background experiences in medical applications related to this field of study.

Will the Data Answer Our Questions?

With the comprehensive dataset we have obtained, encompassing a wide array of demographic and medical features, our project is well-equipped to address pivotal questions surrounding diabetes prediction and the formulation of personalized treatment plans. Our strategic choice of machine learning models, including Logistic Regression and Random Forest, demonstrates a nuanced approach to capturing both linear and non-linear relationships within the data.

The use of robust evaluation metrics, such as accuracy, precision, recall, and F1 Score, underscores a thorough strategy for assessing model performance. Ethical considerations, including awareness of potential biases, along with a commitment to addressing them, reflect a responsible approach to our data analysis. The project's contingency plan underscores a readiness to navigate challenges and refine the analysis iteratively. Our project appears well positioned to

extract meaningful insights from the data, contributing significantly to diabetes prediction and the development of tailored treatment plans for individuals.

Useful Explanatory Visualizations

To comprehensively convey the findings of our diabetes prediction project, a combination of key visualizations is essential. Histograms are pivotal for illustrating the distributions of continuous variables, such as age, BMI, HbA1c level, and blood glucose level, offering insights into their underlying patterns. A correlation heatmap can help visualize relationships among features, aiding in the identification of multicollinearity and guiding feature selection. Scatter plots may prove valuable for exploring relationships between variables, emphasizing their impact on the target variable, diabetes. ROC and Precision-Recall curves will provide a holistic evaluation of predictive model performance, showcasing model sensitivity and precision tradeoffs.

The incorporation of a confusion matrix breaks down model predictions into true positives, true negatives, false positives, and false negatives, offering a detailed understanding of our model errors. The possibility of including Feature importance plots (especially for Random Forest) reveals the significance of different factors in diabetes prediction. Demographic breakdowns using bar charts help identify disparities in diabetes prevalence across different groups.

Lastly, presenting a summary of overall model performance metrics will ensure a clear and concise overview of the models' effectiveness. A combination of these visualizations collectively provides a robust and insightful representation of the data, model outcomes, and key insights for stakeholders involved in diabetes prediction. As we progress within our analysis, the

team will evaluate the proper combination of visualizations to present and evaluate our models' performance.

Reflecting on the Data

In our preliminary analysis, it becomes evident that the dataset at our disposal forms a robust foundation for addressing the problem of diabetes prediction. The inclusion of a diverse set of features, spanning a wide range, offers a comprehensive view of patients' health profiles. The variation not only aligns with the intricacies of diabetes risk factors but also provides an opportunity to explore relationships within the data.

The dataset is thorough with a wide range of values in central variables like age and BMI, and it enhances our ability to discern patterns and potential correlations. Additionally, the inclusion of categorical variables, such as smoking history, allows for a more holistic understanding, acknowledging lifestyle factors that may contribute to diabetes risk. Our preliminary assessment suggests that the dataset is well-suited for our machine learning endeavors, promising meaningful insights into diabetes prediction.

Reflections on Data Modeling & Evaluation Options

At our current analysis stage, there is no immediate indication that a modification to the chosen models and evaluation strategies is necessary. Our continued assessment suggests that the selected models Logistic Regression, Random Forest, and Neural Networks align well with the characteristics of the dataset, which includes diverse features relevant to diabetes prediction.

The balance between interpretability and complexity of the models that is to be maintained, and the chosen evaluation metrics, such as accuracy, precision, recall, and F1 Score, are ultimately the best methods for determining the most accurate models. Additionally, there are

no apparent issues with model performance or ethical considerations that necessitate an adjustment at this point or the foreseeable future.

Finally, it is acknowledged that as the analysis progresses, there may be new discoveries or insights that could influence the model choices. Depending on future discoveries there remains a possibility of model refinement or complexity reduction of the applied models to ensure their optimal alignment with the emerging understanding of the data and the project's objectives.

Original Expectations Revisited

As we go further into the analysis, it is important to reflect on our original expectations, and if they are still reasonable. Thus far, there has been no compelling reason to believe otherwise. There has been set a wild field of expectations, with the use of the three different modeling techniques. Each one will be evaluated and compared against each other using the accuracy, precision, recall and F1 score, from which we may determine the most optimal model to use.

With the use of Linear Regression, Random Forest, and Neural Networks models, and the advantages that each one provides, including interpretability & linear relationships, each one will allow us to reach our original expectations. With this in mind, we may proceed further into the analysis stage.

Data Preparation

Before beginning to build the model, we must prepare the data first. First, we conducted a missing data review to ensure data integrity and readiness for subsequent analyses and model training, and we found no missing values. We also noted that the “gender” feature included a third category labeled as “other”, affecting only an inconsequential 0.018% of the data. For the purpose of building a prediction model based on biological features, the “other” category will be excluded from the gender feature, with remaining genders being represented by binary values (0 or 1), enabling the model to effectively interpret and utilize gender information in its predictions. The remaining categorical features will be converted to dummy indicators.

Next, we looked at smoking history. The smoking history columns represents a categorical variable encompassing six distinct features, ranging from “never” to “current”. To effectively utilize this information in our predictive models, it necessitates transformation into dummy indicator variables. In our final stage of data processing, we aim to standardize all non-numeric type variables to a consistent format of float64, ensuring compatibility across the dataset as the original dataset was few categorical data present the data was recoded for numeric instead of their previous categorical value with ones and zeros in their place. This standardized numerical representation simplifies data handling and enhances the interoperability of various modeling techniques as a result.

Building the Model

We explored two models: K-Nearest Neighbors (KNN) Classifier and Linear Regression. First, we explored the KNN model. This algorithm classifies a data point based on the majority class of its k nearest neighbors in a feature space, in this case being classifying individuals into diabetic and non-diabetic groups based on their features. First, we split the data into an X and Y variables, with X being the predictor variables, and Y the target variable "diabetes". We then split the data into a training and test set utilizing an 80/20 split. We scaled the data using a MinMax scaler, ensuring all features contribute equally to the distance computation used in KNN modeling. Scaling the data thus prevents features with larger scales from having an undue influence, improving the classifier's performance. After fitting the pipeline to the training data, we evaluate the accuracy of the classifier on the test data using the score method, resulting in a 96.03% accuracy.

We attempted to increase the accuracy of this model with the use of hyperparameter tuning. We conducted a Grid Search analysis, which searches through a specified grid of hyperparameters to find the combination that yields the best performance. These parameters include the pipeline, parameter grid, cross-validation strategy, and scoring method. We fit the training test sets to the grid search and ran the search to find the best parameters. This returned a k-fold cross-validation of 9. Adjusting our model to this optimal k-fold cross-validation, this increased our model accuracy from 96.03% to 96.11%. Overall, we found that our results can be considered highly accurate in prediction of diabetes status using the KNN modeling technique.

The next model we explore is Linear Regression, a classification modeling technique primarily used for predicting a continuous outcome variable based on one or more predictor

variables. We chose to utilize statsmodels' linear regression model, this specific model being more suited for statistical modeling and analysis. It implements ordinary least squares (OLS) linear regression, aiming to minimize the sum of the squared differences between the observed and predicted values. This allows for greater flexibility in specifying and examining the regression model, including options for hypothesis testing, parameter estimation, and inference. We then fit and predicted the results, and printed the summary of the results, obtaining the R-squared score, which was 0.16, the F-statistic, and other various measurements.

Results

Each model showed promising performance, indicating potential room for improvement. We can first look at the KNN model, which achieved a final accuracy of 96.11%, and a R-squared value of 0.51. This accuracy exceeded our expectations. While we already increased our accuracy with the use of hyperparameter tuning by conducting a grid search for the best parameters, there is still room for improvement to increase the accuracy further. Next, we can look at the Linear Regression model results. We obtained an R-squared (and adjusted R-squared) score of 0.16. A higher R-squared value indicates that the model explains a large proportion of the variance in the dependent variable and is generally desirable. The R-squared value sits quite low in accordance with this, indicating that the model and the predictor variables make up for 16% of the variance of the target variable. Although it presents quite low currently, there is still much more room for improvement to be made with this linear regression model.

When looking to compare these two models, because they measure different aspects of model performance and are applicable to different types of models, it is not meaningful to directly compare them. However, we able to obtain the R-squared score for each model, and we can see that the KNN model's R-squared score of 0.51 far exceeds the linear regression model's R-squared score of 0.16. For this reason, we will choose to continue forward using the KNN modeling approach.

Conclusion & Recommendations

The models improved quite variably to each other, but there is still room for improvement in each one. Looking first at the KNN model, although the 96.11% is an impressive starting point, there is still room for improvement. There are more steps that could be performed to increase accuracy for this model, including but not limited to feature engineering, ensemble methods, and feature selection. By exploring these strategies and experimenting with different approaches, we can hope to increase the accuracy of diabetes prediction in this model.

For our linear regression model, we obtained an R-squared score of 0.16. Optimally, we would like a higher R-square, score, indicating the model accounting for more of the variability in the model. If one would decide to move forward with this modeling technique, some adjustments that can be made to increase this score is feature engineering, checking for multicollinearity amongst predictor variables, and residual analysis, analyzing the residuals to identify patterns or unequal variance, also known as heteroscedasticity. By using these methods, it can potentially improve the performance of the linear regression model.

Going forward, we advocate for the utilization of the K-Nearest Neighbors Classification model and making the recommended adjustments to improve and refine the predictive capabilities. With these enhancements, we hope for a more model to aid in the prediction of diabetes.

Our final diagnosis of our model using the KNN Classification upon our revisit highlighted our hypothesized improvement on model prediction, as a result we saw a significant increase in our R^2 value from 16% to 51% as well as an increase in accuracy of 96.3% with the newly utilized model. While the R^2 value still isn't of statistically relevant significance this would suggest further variable refinement within our data may be warranted to successfully prove the efficacy of diabetes predictability in our population for a more promising result.

References

Mustafa, M. (2023, April 8). *Diabetes prediction dataset*. Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset