HIDS-7006: Project Report

Predicting Patients Diabetes Status

Joan Mattle

1. Introduction

Diabetes is a chronic health condition where the body struggles to efficiently convert food into energy due to insufficient insulin production or ineffective use of insulin. Normally, the body converts food into glucose, which is released into the bloodstream, prompting the pancreas to produce insulin. Insulin facilitates the entry of glucose into cells for energy use. However, in diabetes, either insufficient insulin is produced or cells become resistant to its effects, resulting in elevated blood sugar levels. This prolonged condition can lead to serious health complications such as heart disease, vision loss, and kidney disease. While there is no cure for diabetes, adopting a healthy lifestyle through weight management, nutritious eating habits, and regular physical activity can significantly improve the management of the condition.

Understanding the root causes of diabetes is imperative for effective prevention strategies. My project leverages the CDC Diabetes Health Indicators dataset to predict a patient's diabetes status and identify the key features that contribute to accurately predicting this status. This approach is pivotal as it enables us to prioritize interventions and devise personalized plans for individuals with identified risk factors.

2. Previous Work

The inspiration for this project was the paper "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques". This study aimed to develop predictive models to identify risk factors for type 2 diabetes, aiding in early diagnosis, intervention, and reducing medical costs. Using data from the 2014 Behavioral Risk Factor Surveillance System (BRFSS), the study analyzed 138,146 participants, including 20,467 with type 2 diabetes. Several machine learning models were built, including support vector machine, decision tree, logistic regression, random forest, neural network, and Gaussian Naive Bayes classifiers.

Results indicated that all models achieved high accuracy and area under the curve (AUC). The neural network model performed the best overall, while the decision tree model was preferred for initial screening due to its high sensitivity. Additionally, the study identified new potential risk factors for type 2 diabetes, including both under-sleeping (≤6 hours per day) and over-sleeping (≥9 hours per day), as well as infrequent checkups. Overall, the study highlighted the utility of machine learning in predicting type 2 diabetes risk and identifies novel factors that could inform early detection and intervention strategies.

3. Methods

3.1 Data Source

The data source is called CDC Diabetes Health Indicators from UC Irvine Machine Learning Repository. The dataset contains healthcare statistics and lifestyle survey information about people in general along with their diagnosis of diabetes. It includes 253680 instances and 21 features.

3.2 Pre-processing the data

The first step to preprocessing was one-hot encoding the categorical variables in order for the machine learning model to interpret categorical variables. The second step was to balance the data. In this dataset 86 percent of the people did not have diabetes, while 14% did. The package SMOTE was used to generate synthetic samples for the minority class by interpolating between existing minority class samples to balance the class distribution in the training data, ensuring equal representation of minority and majority classes.

3.3 Models

Four models were run: random forest, logistic regression, SVM, and neural network. For the random forest model and logistics regression model hyperparameter tuning was conducted using grid search with five-fold cross-validation in order to optimize the model’s recall. For random forest parameters explored included criteria for splitting nodes ('gini' or 'entropy'), the number of estimators (trees) in the forest, whether bootstrap samples were used, maximum depth of the trees, maximum number of features considered for splitting, and minimum number of samples required to split a node. The parameter tuning for logistic regression revealed that the regularization strength (C) has been set to 0.001, employing L2 regularization penalty, and utilizing the 'liblinear' solver algorithm for optimization. Due to GPU limitations parameter tuning was not able to be conducted for the SVM model and default parameters were used. The neural network model consists of three fully connected (dense) layers followed by a dropout layer to prevent overfitting. The first hidden layer has 32 units, employs the ReLU activation function, and takes input dimensions matching the shape of the training data. The first hidden layer has 32 units, employs the ReLU activation function, and takes input dimensions matching the shape of the training data. The output layer contains a single unit with sigmoid activation, suitable for binary classification tasks. The model is compiled using the Adam optimizer and binary cross-entropy loss function and accuracy is monitored as a metric during training. Training is conducted over 100 epochs with a batch size of 16.

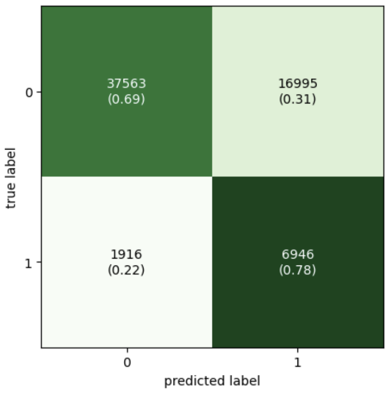
4. Results

4.1 Classification Report

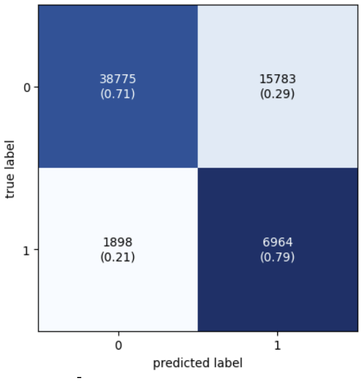
Method	Accuary	Recall 0	Recall 1	Precision 0	Precision 1	F1-Score 0	F1-Score 1
Random Forest	0.70	0.69	0.78	**0.95**	0.29	0.80	0.42
Logistic Regression	0.72	0.71	**0.79**	**0.95**	0.31	0.81	0.44
SVM	0.72	0.71	**0.79**	**0.95**	0.30	0.81	0.44
Neural Network	**0.80**	**0.93**	0.38	0.83	**0.61**	**0.88**	**0.47**

4.2 Confusion Matrices

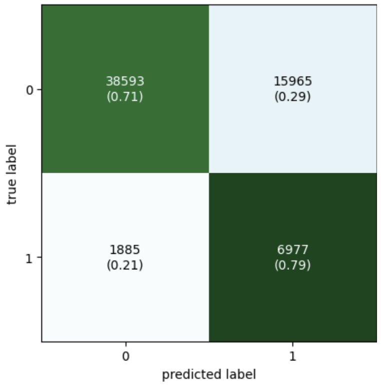
Random Forest



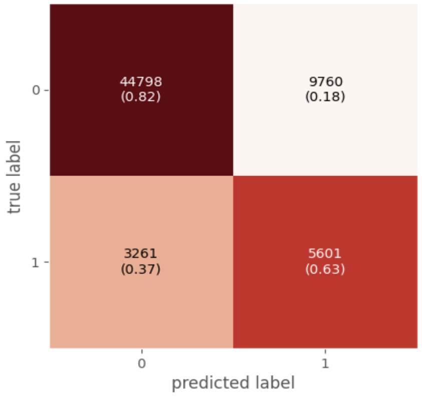
Logistic Regression



SVM

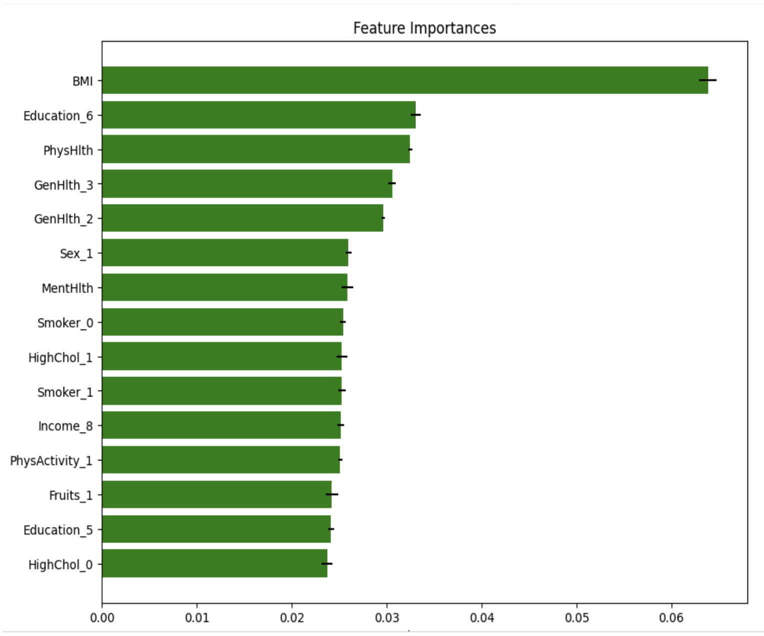


Neural Network

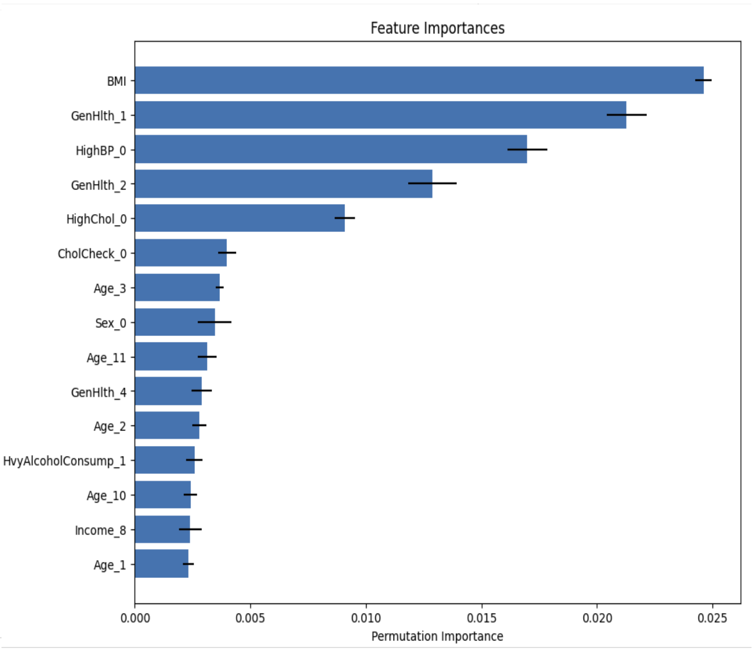


4.3 Feature Importance

Random Forest



Logistic Regression



5. Discussion and Future Work

Although the neural network model has the highest accuracy, the logistic regression and SVM models have the highest recall for positive predictions. Given the goal of predicting diabetes the goal is to optimize the true positive rate. Therefore the best models here are logistic regression and SVM. For both the random forest and logistic regression models BMI is the most important feature in predicting diabetes status. Other features with importances include general health (excellent, very good, and good), high blood pressure, high cholesterol, education level (college graduate), physical health in the past 30 days, and if cholesterol has been checked in the past five years.

Future work that could be done for this project to improve its performance. First additional models could be run such as naive bayes and Extreme Gradient Boosting. These models could potentially improve recall and better predict when a person has diabetes. Moreover, looking at the important features used to predict if a person has diabetes could help discover the factors that play into a person getting diabetes and help people prevent getting the disease. In addition, feature engineering could be done in order to create new features or transform existing ones that may better capture underlying patterns in the data and improve a models performance. Finally, error analysis visualization could help identify patterns or trends in misclassifications, providing insights into areas where the model struggles and guiding further improvements.

Work Cited

1.17. neural network models (supervised). scikit. (n.d.-a). https://scikit-learn.org/stable/modules/neural_networks_supervised.html

CDC diabetes health indicators. UCI Machine Learning Repository. (2023, September 25). https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

Feature importances with a forest of trees. scikit. (n.d.-b). https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Probability calibration curves. scikit. (n.d.-c). https://scikit-learn.org/stable/auto_examples/calibration/plot_calibration_curve.html#sphx-glr-auto-examples-calibration-plot-calibration-curve-py

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019, September 19). *Building risk prediction models for type 2 diabetes using Machine Learning Techniques*. Centers for Disease Control and Prevention. https://www.cdc.gov/pcd/issues/2019/19_0109.htm