*Predicting diabetes from silent symptoms using machine learning*

# Introduction

Diabetes is a chronic disease that occurs either when pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation. The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014. Prevalence has been rising more rapidly in low- and middle-income countries than in high-income countries. (*Diabetes*., n.d.)

According to the World Health Organization symptoms of diabetes include the need to urinate more often than usual, feeling very thirsty, constant hunger, weight loss, vision changes and fatigue. These symptoms may occur suddenly. These are early signs but are subtle and as a result, disease may be diagnosed several years after onset, after complications have already arisen.

The starting point for living well with diabetes is an early diagnosis – the longer a person lives with undiagnosed and untreated diabetes, the worse their health outcomes are likely to be. Hence an early detection of diabetes is critical to mitigate its effects and to improve the life expectancy.

This paper intends to utilize machine learning models to predict the likelihood of having diabetes based on the silent symptoms and offers a potentially more proactive solution to individuals who may benefit from early intervention and lifestyle modifications. This non-invasive machine learning solution benefits in identifying individuals at risk of diabetes particularly in low-income countries, where the access to comprehensive medical care is often constrained by economic, infrastructural, and cultural factors.

# Dataset Overview

Dataset was obtained from UCI repository. This data has been originally collected using direct questionnaires from patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. The dataset includes 17 attributes and 520 observations. There are no missing values.

All attributes except age are categorical. The attributes include 14 silent symptoms of diabetes with values ‘Yes’ or ‘No’ depending on prevalence of symptom in the individual. While most attributes are easily understood, a few require additional description and it is given below:

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Polyuria | Excessive urination. |
| Polydipsia | Excessive thirst |
| Polyphagia | Excessive hunger |
| Genital Thrush | Discomfort and itching in genital area |
| Partial Paresis | Weakness in a part of the body, resulting in reduced muscle function. |
| Alopecia | Hair-loss |

The value of the class variable is ‘Positive’ for individuals diagnosed with diabetes and ‘Negative’ for those without diabetes. The dataset exhibits an imbalance in its distribution, with approximately 62% of observations belonging to the 'Positive' class, while the remaining 38% are associated with the 'Negative' class.

A graph of a number of classes

Description automatically generated

The number of females who belong to class ‘Negative’ is very less as per the below chart.

A graph of a class grouped by gender

Description automatically generated

When there is an imbalance in the dataset it can lead to misinterpretations and the model might not generalize well on unseen data. To overcome this issue, stratified cross validation technique is used in this study.

The minimum age of individuals available in the dataset is 16 and maximum age is 90. Based on the below boxplot the individuals having an age greater than 80 are outliers. The recommended screening for prediabetes and type 2 diabetes in adults are from ages 35 to 70 years.  (Jonas *et al.*, 2022). Based on this it is decided to remove outliers from dataset.

A graph with a box plot

Description automatically generated with medium confidence

The presence of duplicate records is noted during the analysis, but there was no means to identify if these observations are actual duplicates, as the dataset doesn’t have information about the individual name or time at which the observation is recorded. Hence it was decided to retain the duplicate records assuming they are from different individuals.

As part of Exploratory data analysis, chi-squared test is conducted to understand the relationship between each feature and class variable (Brownlee, 2018). The chi-squared value indicates the strength of association between two variables, and the p-value tells whether the association is statistically significant. A low p-value (typically less than 0.05) suggests a significant association between variables.

The results of chi-squared test are arranged based on strength of association between the Feature and Class variable. Polyuria shows the strongest association with class variable. For the features ‘Obesity’, ‘delayed healing’ and ‘Itching’ p value is higher, and it indicates there is no statistical significance in the association between class variable and these features, as per the given dataset and these variables might not contribute to classification.

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Feature** | **Chi-Squared** | **p-value** |
| 1 | Polyuria | 227.866 | 0.0000 |
| 2 | Polydipsia | 216.172 | 0.0000 |
| 0 | Gender | 103.037 | 0.0000 |
| 3 | sudden weight loss | 97.296 | 0.0000 |
| 11 | partial paresis | 95.388 | 0.0000 |
| 5 | Polyphagia | 59.595 | 0.0000 |
| 9 | Irritability | 45.208 | 0.0000 |
| 13 | Alopecia | 36.064 | 0.0000 |
| 7 | visual blurring | 31.808 | 0.0000 |
| 4 | weakness | 29.768 | 0.0000 |
| 12 | muscle stiffness | 7.289 | 0.0069 |
| 6 | Genital thrush | 5.792 | 0.0161 |
| 14 | Obesity | 2.327 | 0.1271 |
| 11 | delayed healing | 0.962 | 0.3267 |
| 9 | Itching | 0.046 | 0.8297 |

The multivariate analysis uncovers the relationship among features and here the chi squared test is used as all features are categorical (Brownlee, 2019). According to the test results, ‘Polydipsia’ and ‘Polyuria’ show strong association and these two features are strongly associated with the feature ‘partial paresis’. The statistical significance of these associations is verified by checking the p-value.

A diagram of a number of objects

Description automatically generated with medium confidence

# Data Processing Steps

1. Data Loading and basic validations
2. Removed outliers (observations with Age>80)
3. As all categorical variables have only 2 distinct values it is replaced with numerical values. (Yes as 1 and No as 0, Male as 1 and Female as 0 , Positive as 1 and Negative as 0)
4. Min Max Normalization: The dataset has the continuous variable ‘Age’ and all other features have value 0 or1. The difference in ranges might cause algorithms to give more weight to "age" compared to binary features. Scaling helps to treat both Age and other binary features equally.
5. Split dataset into two sets.
   1. Training (80%) – Used for training and cross validation and to compare and finalize the best model.
   2. Test dataset (20%) – To evaluate the performance of the best model on unseen data.
6. Fit various models on training data.
7. Perform stratified cross validation using training data for each model, compare performance metrics and select best model. It enforces the class distribution in each split of the data to match the distribution in the complete training dataset. (Brownlee, 2020)
8. Interpret the best model and evaluate performance on unseen test data.

# Algorithm Selection

This use-case is a supervised learning problem with features (mainly symptoms) and labels (diabetic or non-diabetic). The objective is binary classification – categorizing individuals as diabetic or not based on symptoms.

Algorithms for binary classification is identified by considering below factors.

1. Categorical nature of input features.
2. Relatively small dataset size.
3. Model Complexity and Interpretability.
4. KNN

k Nearest neighbors is one of the simplest predictive models. For classification, k-nearest neighbor method assigns most common class of nearest neighbors of an instance. It doesn’t build a model explicitly; in other words, it is nonparametric. As it doesn’t have a hypothesis, to classify a new instance, it needs to calculate the distances to other instances every time. Hence it is computationally expensive. The value of k and the distance metric decides the performance of model. KNN is not interpretable on a modular level as there are no parameters to learn. But to explain a prediction, we can retrieve the k neighbors that were used for the prediction. So, when the number of features is minimal the model becomes explainable. (Molnar, n.d.) As kNN is suitable for small data sets and it is interpretable with minimal features, it has been decided to consider this model.

1. Naïve Bayes

Naïve Bayes algorithm is simple, efficient, and fast. It is suitable for small datasets. It is based on Bayes’ probability theorem. The probability of an event (class) can be better estimated if some prior knowledge of conditions (features) that might be related to the event is available. Naïve Bayes classifier is a generative model as it classifies the data by assuming probability distributions within the data. (Raschka, 2017) It is interpretable as we can understand how much each feature contributes towards the class prediction based on the conditional probability. (Molnar, n.d.)

1. Logistic Regression – It is a simple and effective model. The predicted parameters give inference about the importance of each feature.
2. Decision Tree and Random Forest

Decision Tree is a simple nonparametric model. It is easy to implement, cheap and highly interpretable. But it is subject to overfitting.

Random Forest is an ensemble learning method that combines the predictions of multiple individual models to create a stronger, more accurate model. It is suitable for categorical input variables. The random forest method helps to reduce overfitting by combining multiple decision trees, each trained on a slightly different subset of the data. Each subset considers a random group of features. But it is difficult to interpret the model and feature importance is one way by which the model can be understood.

**Algorithm not considered.**

1. Support Vector Classifier – SVC lacks interpretability and it is more effective for high dimensional problems and hence it is not considered for this study.

# Performance Evaluation Methodology

The performance of various models can be evaluated using accuracy, precision, recall and F1 score. For use cases related to disease prediction accuracy alone might not help to evaluate the model because even though accuracy is high if number of false negatives (Individual has diabetes but model predicted it as non-diabetic) is more it can be dangerous. Precision describes how accurate positive predictions are. Recall provides fraction of positives the model identified. F1 score is the harmonic mean of precision and recall. False Negative Rate measures the proportion of actual positive cases that were incorrectly predicted as negative by the model. The area under the curve (ROC-AUC score) is calculated to evaluate the performance. (E Alpaydin, 2014, *19.7 Measuring Classifier Performance*)

Accuracy = (TruePositive + TrueNegative)/ N

Precision (P)= TruePositive/(TruePositive + FalsePositive)

Recall (R) = TruePositive/(TruePositive + FalseNegative)

F1Score = (2\*P\*R)/(P+R)

False Negative Rate (Miss Rate) = False Negative/ (False Negative + True Positive) or 1-Recall

Here, weighted average F1 score and false negative rate are used to compare the performance of various models. For this use case the Miss Rate is important as it provides a clear idea of cases that are predicted as non-diabetic while it is diabetic. It emphasizes models’ tendency to miss positive cases.

# Model Selection

The models are trained using training dataset and it is evaluated using stratified cross validation using training data. The classification report for each algorithm is available in the Appendix. The report summary is given below:

|  |  |  |
| --- | --- | --- |
| **Model** | **Weighted F1 Score** | **Miss Rate** |
| NaiveBayes | 0.87 | 14% |
| kNN | 0.92 | 11% |
| Logistic Regression | 0.92 | 6% |
| Decision Tree | 0.96 | 3% |
| Random Forest | 0.98 | 1% |

As per the report summary, Random Forest model is the best performing model considering F1 score and miss rate. The Miss rate for interpretable models is relatively high when compared to random forest and standard-deviation from the cross-validation results is comparatively less (0.023757) for Random-Forest, meaning less variance. Hence it is decided to select random forest for evaluation.

To address the class imbalance, an additional analysis was carried out by oversampling the training dataset using Synthetic Minority Over-sampling Technique (SMOTE) and all models were trained and tested using the resampled training data. However, it did not give any noticeable improvement in model performance and miss rate. Hence the existing training dataset is used.

Predictions are made using Random Forest model on unseen data (Test Data) and performance metrics is given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Forest** | **F1 Score** | **Miss Rate** | **Precision** | **Recall** |
| Train | 0.98 | 1% | 0.98 | 0.98 |
| Test | 0.98 | 1.60% | 0.98 | 0.98 |

The miss rate is slightly higher in test results. But test predictions are comparatively accurate and similar and there is no deviation from the training results.

# Model Interpretation using Feature Importance.

The feature importance score is calculated by aggregating the contributions of feature across all trees in forest. It is based on Gini Impurity. Features that consistently lead to greater reduction in impurity when used for splitting tend to receive higher importance scores.

The features in the order of importance used by random-forest model are given below. This is complimenting the results of EDA indicating the reliability of the model.

A graph with blue and white bars

Description automatically generated

|  |  |
| --- | --- |
| **Feature** | **Importance Score** |
| Polydipsia | 0.2368 |
| Polyuria | 0.1739 |
| Gender | 0.0938 |
| Age | 0.0861 |
| partial paresis | 0.0683 |
| Irritability | 0.0563 |
| sudden weight loss | 0.0478 |
| Alopecia | 0.0403 |
| delayed healing | 0.0324 |
| Polyphagia | 0.0302 |
| Itching | 0.0283 |
| muscle stiffness | 0.0277 |
| visual blurring | 0.023 |
| Genital thrush | 0.0207 |
| Obesity | 0.0174 |
| weakness | 0.0168 |

From feature importance and EDA “Obesity” is considered as least important factor as per given dataset. But it is one of the important causes for diabetics (Watts, 2022). Hence this needs to be verified by domain experts.

# Conclusion

To conclude, by utilizing the power of machine learning techniques diabetes can be detected early just from silent symptoms and without any medical interventions. The model was able to correctly predict 98% of cases. However, it's worth noting that a small fraction of individuals (1.6%) was falsely identified as non-diabetic, potentially requiring further refinement. Addressing this concern could involve refining the model with new data, ensuring that it accurately caters to both positive and negative cases.

Word Count: 2198

The code is available in Gitlab. link: https://gitlab.uwe.ac.uk/n4-jose/ml.git

Appendix

Classification Report for various Models

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

REFERENCES

*Diabetes*. (no date) [online]. Available from: <https://www.who.int/health-topics/diabetes> [Accessed 24 August 2023].

UCI Machine Learning Repository (2020) *Early stage diabetes risk prediction dataset.* [online]. Available from: <https://archive.ics.uci.edu/dataset/529> [Accessed 23 August 2023].

Jonas, D.E., Schaaf, E.B.V., Riley, S., Allison, B., Middleton, J.C., Baker, C., Ali, R., Voisin, C.E. and LeBlanc, E. (2022) Introduction. In: *Screening for Prediabetes and Type 2 Diabetes Mellitus in Children and Adolescents: An Evidence Review for the U.S. Preventive Services Task Force [Internet]* [online]. Agency for Healthcare Research and Quality (US). Available from: <https://www.ncbi.nlm.nih.gov/books/NBK584383/> [Accessed 23 August 2023].

Brownlee, J. (2018) A Gentle Introduction to the Chi-Squared Test for Machine Learning *MachineLearningMastery.com*. 14 June 2018 [online]. Available from: <https://machinelearningmastery.com/chi-squared-test-for-machine-learning/> [Accessed 24 August 2023].

Brownlee, J. (2019) How to Perform Feature Selection with Categorical Data *MachineLearningMastery.com*. 24 November 2019 [online]. Available from: <https://machinelearningmastery.com/feature-selection-with-categorical-data/> [Accessed 24 August 2023].

Molnar, C. (no date) *5.7 Other Interpretable Models | Interpretable Machine Learning* [online]. Available from: <https://christophm.github.io/interpretable-ml-book/other-interpretable.html> [Accessed 25 August 2023].

Raschka, S. (2017) *Naive Bayes and Text Classification I - Introduction and Theory* [online]. Available from: <http://arxiv.org/abs/1410.5329> [Accessed 25 August 2023].

Watts, M. (2022) The UK is the fattest country in Europe. The number of obese adults is forecast to rise by 73% over the next 20 years from to 26 million people, resulting in more than a million extra cases of type 2 diabetes, heart disease and cancer. *Diabetes*. 8 September 2022 [online]. Available from: <https://www.diabetes.co.uk/diabetes-and-obesity.html> [Accessed 26 August 2023].

Brownlee, J. (2020) How to Fix k-Fold Cross-Validation for Imbalanced Classification *MachineLearningMastery.com*. 12 January 2020 [online]. Available from: <https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/> [Accessed 28 August 2023].

Alpaydin, Ethem. *Introduction to machine learning*. MIT press, 2014.