BIRMINGHAM CITY
University

Diabetes Prediction Using Machine Learning

Student Name: Aditya Adhikari

Student ID: 25123781

Date: 30th May 2025

Module Code: CMP4294

Module Title: Introduction to AI

Co-ordinator: Umesh Regmi

Page count: 11

Word count: Approximately 2500

Table of Contents

1. Introduction

Diabetes mellitus is a medical care problem that is severe all over the world and is possessed by millions of people. World Health Organization (2024) confirms that the adult population with diabetes has been increasing speedily in the recent years, largely attributable to sedentary life styles, unhealthy diets and an ageing population. The early detection of diabetes is highly important, because the lack of its control may lead to some severe health consequences, including heart disease, kidney failure, blindness and amputations.

Most of the conventional diagnostic tools rely on the performance of clinical tests such as blood sugar and hemoglobin test, which, however, require medical resources and cannot be applied in resource-limited locations. Machine learning (ML) has emerged as a productive technology in healthcare to enhance and accelerate disease identification (Shickel et al., 2018) ML algorithms analyze past health data to discover underlying relationships among several clinical variables and this enables them to detect diabetes early and identify the individuals at risk.

This report aims at determining whether one can use health data to predict diabetes in patients using supervised machine learning. Using the Pima Indians Diabetes dataset we experiment with two different classifiers; Logistic Regression and Random Forest in order to identify significant variables and compare the output of both models by calculating accuracy, recall, precision and F1 score. It is aimed at emphasizing that machine learning may assist physicians in making more decent decisions and improving health outcomes through delivering rapid and trustful diagnostics.

2. Background

Because so many people now have diabetes, healthcare systems are struggling to cope. Frequently, patients in developing regions do not get diagnosed quickly because they cannot easily get tested in clinical labs. As a result of machine learning, it will now be possible to use quick, organized and accurate screening tools to address the gap in diagnoses.

Historical data can be used by ML to spot patterns and confidently predict outcomes. In predicting diabetes, important features are blood pressure, blood glucose levels, BMI, insulin levels and age of the patient. Working these features alongside analytics can help discover a patient's risk of diabetes.

Because Logistic Regression is a well-known technique, it helps in solving classification problems where the response can be either 'diabetic' or 'non-diabetic.' Also, Random Forest Classifiers depend on grouping many decision trees to create predictions that are both accurate and can be used outside the training sample. Having different frameworks gives users an advantage in deciding which model to use and starting its implementation.

3. Aim and Objectives

Aim:

Designing a machine learning model to anticipate if a patient has diabetes by looking at certain clinical data.

Objectives:

- Feature patient health data grouped by labels.
- Take care of any missing, unevenly distributed or improperly scaled data.
- Look for ways to visually understand and study how data is related.
- Make use of classification algorithms, for example Logistic Regression and Random Forest.
- Check a model's performance using accuracy, precision, recall, F1-score and the confusion matrix.

4. Dataset Description

The Pima Indians Diabetes Dataset which is often applied in healthcare ML fields, is used in this example. It includes 450 records and 7 columns along with 1 target attribute(Kahn, n.d.).

Features include:

- Pregnancies
- Glucose
- Insulin
- BMI
- DiabetesPedigreeFunction - Age -Outcome

Target:

- Outcome – 1 indicates diabetic, 0 indicates non-diabetic

Pandas were used to read in the data and its shape and head were examined.

5. Problem to be Addressed

It tries to find solutions for predicting diabetes using as little medical equipment as possible. Particular issues being looked at are:

-Feature Relevance: Knowing which medical signs have the greatest impact on the chance of developing diabetes.

-Occasionally, clinical data have entries that are missing or don't make sense and these must be managed to keep the model strong.

-When Predictive models encounter data with an uneven distribution (1 sample to 10 samples), they might show bias favoring the majority group (non-diabetic).

-Model Transparency: Medical experts rely on trusting and acting on predictions which is why interpretability is important in clinical models.

-Using machine learning to create a pipeline overcomes these issues, making it possible for screening to join electronic medical charts and phone apps.

6. Machine Learning Model

Summary of the Machine Learning Approach

It uses the typical approach for a monitored machine learning pipeline:

- Handling data by exporting it and then cleaning it
- Examining the data through exploratory data analysis (EDA)
- Changing the features of data
- Model training and evaluation are important to perform.
- Deciding which classifier is the best by considering a range of evaluation metrics

Two kinds of classifiers were used:

- Logistic Regression is helpful when the outcome can only be yes or no.(scikit-learn.org, n.d.)

-Random Forest Classifier uses multiple machines to improve how it works.

7. Data Processing

Handling Missing Values:

The dataset included Insulin, BMI (Body Mass Index) and SkinThickness features that mostly contained zero values. Although statistically a zero is allowed, in healthcare it is rather surprising and usually reflects either a missing or unrecorded value during data collection. Insulin levels or BMI of zero are not possible for a living person.

For this reason, missing values were filled with proxy or temporary, entries of zero. It was crucial in data preprocessing to find and treat these correctly, as that kept the model from receiving wrong information. Several approaches were tried, depending on the pattern seen in the data of each feature. Normally distributed features were filled in by mean imputation and outlier-affected or uneven features were filled in by median imputation. Giving particular attention to pseudo-missing data helped both keep the data clean and make the model more effective.

Data Cleaning:

It is very important to clean data well in machine learning, especially for medical data, because even little errors in the data can lead to major problems with predictions. Thoroughly examining the data revealed inconsistent entries, outliers and duplicate records in this project.

Great care was taken to recognize and exclude measurements that are not possible in nature (e.g., BloodPressure was never interpreted as “zero”). As these were considered outliers, they were handled by marking them as missing or correcting them using what is known for the type of data and analysis. The data was checked for misspelled words, any unusual formatting and wrong column structures that could disturb the extraction and training of features.

No real null (NaN) values showed up in the data, though several zeros were understood as missing when considering the clinical relevance of each field. After dealing with the problems, the data in the dataset was rebuilt, checked and ready for more study. It made sure that all inputs were thorough and clinically correct which is necessary in healthcare machine learning projects.

8. Exploratory Data Analysis

Class Distribution:

This dataset has 172 diabetic patients and 277 non-diabetic patients which is evidence of moderate class imbalance. It follows that of the patients in the dataset, around 38% have diabetes and the rest do not. Such a gap may lead to poor classification results for models, much more so if accuracy is the single metric used. The model might favor examples from the class with the most cases which are non-diabetic. Because of this, the models were evaluated based on precision, recall and F1-score and stratified splitting was applied to guarantee that all classes were treated the same while training the model.

Correlation Matrix:

A heatmap was created using a correlation matrix to see how features are related to the target variable. Correlation coefficients between every pair of numerical features are shown in this matrix which can suggest which features should come first in the model.

When examining the strongest relationship, Glucose significantly and positively connected with Outcome, meaning that high glucose values are a clear sign of diabetes. BMI and Age were closely linked to a higher target score, agreeing with medical data that people with more body fat and advanced age are generally more likely to develop diabetes. DiabetesPedigreeFunction showed low, yet noticeable correlations together with Pregnancies, compared to Insulin's weak correlation, most likely because of its high variability and zeros. Thanks to these findings, some features were chosen and confirmed as clinically meaningful.

Feature Distribution:

To know the features' behavior, each one was examined with histograms and box plots. Looking at the plots made it possible to notice the characteristics of every variable, look for skewness and find potential problems with the data.

Most values for Glucose, BMI and Insulin occupied the smallest part of the range, while just a few high values appear far at the other end. The age of patients seemed normally distributed, except for a long tail on the high end, meaning that significantly more middle-aged people than elderly are present. Counts of pregnancies followed a clear pattern and the DiabetesPedigreeFunction usually had small values with only a few very high ones. It was revealed through these patterns that scaling and handling zeros or strange inputs should be done before any training takes place.

9. Outlier Detection:

Removing outliers during preprocessing is vital because they can wrongly affect statistics and negatively impact models which depend on scale or variance.

Here, box plots were mainly used to find outliers in the numerical continuous features Insulin, BMI, Glucose, Age and DiabetesPedigreeFunction. Box plots can quickly display the

IQR and mark any data points that are greater than 1.5 times the IQR away from the first or third quartile—usually these are considered outliers.

Several features, especially Insulin and DiabetesPedigreeFunction, had values that stood out as much higher than the rest. Some readings for Insulin went far higher than what is normal in the body and this could be an extreme case or lead to faulty results. BMI was found to have some very high figures and these could not represent most of the population.

Before removing outliers, experts in the field discussed if these values matched what was expected. Because unusual values in some health factors might indicate diabetes risk, they were kept in the data. Therefore, models became robust by applying algorithms such as Random Forest which perform well even at values near the extremes. Also, it was necessary to adjust the data using scaling to not let outliers greatly influence the training process.

10. Data Transformation

Scaling:

Applying StandardScaler made sure the numerical features had a mean of zero and variance of 1 which the Logistic Regression model relies on.(scikit-learn.org, n.d.)

11. Feature Selection:

This project used a mix of understanding the problem, looking at the data closely and testing different models to determine which features to include. Although the Recursive Feature Elimination (RFE) algorithm was not used, features were chosen after considering their relation to the target, their different distribution patterns and what the features represented in a clinical setting.

According to the correlation matrix, the presence of Glucose, BMI and Age indicated a strong correlation with the outcome and thus their own importance in predicting diabetes. We chose to keep DiabetesPedigreeFunction which is not very correlated, since it gives important clinical information about the likelihood of diabetes.

Besides making predictions, the Random Forest classifier was also applied to figure out what features are most important. It judges the effect of every feature on decision-making using the Gini impurity or information gain. Based on this model, Glucose was the top factor, followed by BMI, Age and Pregnancies. Because of many zero readings and other missing-like values, the importance of insulin lowered. (Towards Data Science, 2019)

All the features were kept here because they all mattered and could interact with each other in unexpected ways that could change the outcome. Yet, the highest importance ranking led people to focus first on the most informative parts during analysis and interpretation. Other feature reduction techniques, like Recursive Feature Elimination.

12. Model Training

Choosing a Model:

- Logistic Regression is often used for being easy to understand and simple to implement.
- Random Forest serves to deal with non-linearity and provides strong results.

The data was divided into a train set (80%) and a test set (20%) in both models.

13. Model Evaluation

Metrics Used:

Multiple evaluation metrics were used to check the performance of all models:

-Accuracy shows how often the predictions are correct which is easy to see but can sometimes be meaningless when some classes are more common than others.

-Precision tells us what proportion of the positive predictions by the model are accurate. Doctors rely on high precision to distinguisher healthy patients from people diagnosed with diabetes.

-Recall (Sensitivity) reflects the accuracy of the model in recognizing actual diabetic cases. Failing to recognize real positive cases in health screening could have serious results.

-F1 Score makes it easier to compare precision and recall which is needed when the dataset contains more of one kind of data than another.(Towards Data Science, 2019)

-With a Confusion Matrix, you can see the actual positive predictions, actual negative predictions, incorrect positives and incorrect negatives in the model's predictions. (scikit-learn.org, n.d.)

All of the evaluation metrics showed that the Random Forest model was the strongest performer. The fact that it can handle relationships that are not always straightforward gave boost to this method, making it better than Logistic Regression.

14. Hyperparameter Tuning:

Although Optuna or GridSearchCV was not required, the default parameters were enough to show the performance. Optimizations could be made after through future tuning.

15. Result

The Results Favor: Random Forest Classifier

After comparing the two models, the Random Forest Classifier was seen as the most accurate and dependable one:

-It can make accurate predictions 81% of the time which demonstrates strong performance.

-There is a 79% chance for the test to correctly identify true cases of diabetes.

-Recall is 72% which suggests that it catches the majority of positive examples.

-F1 Score: 75% which means the model does well in balancing precision and recall.

-The confusion matrix which is great for healthcare, showed fewer wrong reports of healthy people as diabetic patients than logistic regression.

So, Random Forest meets the basic needs for an important role in a medical prediction system.

16. Recommendations

- The Random Forest model should be used as the main approach.
- Fill in missing data points by imputing.
-Try to use models like XGBoost that are more complex.
-Try using GridSearch or Optuna for tuning your hyperparameters.

17. Future Work

- If this feature is available, include things like HbA1c.
- Deployment: Put together a web site where users can enter data and get an immediate prediction.
- Carry out systematic tuning to improve how many cases are retrieved.

-Show model predictions in terms that are easy to understand with (Shapley Additive Explanations)SHAP or LIME(Local Interpretable Model-agnostic Explanations).

18. Conclusion

In brief, this study found that machine learning is very effective, especially the Random Forest Classifier, at determining diabetes risk in patients. The entire process which starts with preparing data and ends with assessing results, is a good demonstration of ML's value for public health.

The project achieved success because data science principles were used such as cleaning data, picking suitable features, scaling numbers and choosing appropriate algorithms for binary classification. Even with positive findings, the study points out that there is still work to do on hyperparameter tuning, adding a variety of features and deploying the system in real time.

In many cases, detecting diabetes before it causes problems greatly helps patients and lessens serious outcomes. Using models like Random Forest in healthcare, mobile apps and decision-support tools for doctors could shape how diabetes is handled, mainly in places with few resources.

With the evolution of machine learning, it is now expected to have a major positive effect on diagnostics, preventive care and outcomes for patients—making the technologies featured in this report increasingly needed as time goes by.

References

Shickel, B., Tighe, P.J., Bihorac, A. and Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), pp.1589–1604. doi:https://doi.org/10.1109/jbhi.2017.2767063.

Kahn, M. (n.d.). *UCI Machine Learning Repository*. [online] archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/dataset/34/diabetes.

scikit-learn.org. (n.d.). *scikit-learn: machine learning in Python — scikit-learn 0.22.2 documentation*. [online] Available at: https://scikit-learn.org.

Towards Data Science. (2019). *Towards Data Science*. [online] Available at: https://towardsdatascience.com

World Health Organization (2024). *Diabetes*. [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes.