Diabetes Classification Using Machine Learning

Helle Skjøth Sørensen (HS)
School of Communication and Culture, Aarhus University
Data Science, Prediction, and Forecasting
Chris Mathys

Lina Elkjær Pedersen (LP)
School of Communication and Culture, Aarhus University
Data Science, Prediction, and Forecasting
Chris Mathys

Link to Github: https://github.com/orgs/Data-science-exam/repositories

Abstract

Diabetes is a chronic disease which affects blood glucose regulation. Millions of people worldwide are affected by the disease, which is developed due to a combination of genetic and environmental factors. In this paper, two different machine learning algorithms, logistic regression and XGBoost, are used to predict diabetes. The data used for the classification task contain demographic information along with various diabetes risk factors such as smoking habits, haemoglobin levels, hypertension, and BMI. The accuracy scores of the classification algorithms are evaluated and discussed. It was found that the XGBoost algorithm performed better than the logistic regression algorithm. The hyperparameter tuning method, Bayesian optimisation, was employed on XGBoost to further improve classification accuracy. However, the difference between the different XGBoost classifiers was very small and could be attributed to chance. The best performing model, the Bayesian optimised XGBoost, had an accuracy score of 97.33%. The features which were identified as most influential in this model were: blood glucose level, age, BMI, HbA1c level and smoking history. These features are in line with what medical research considers particularly influential risk factors of diabetes.

Table of contents

Introduction

Diabetes is a chronic, metabolic disease which affects a large number of people worldwide, and its prevalence is rising rapidly (Mayer-Davis et al., 2017; Zimmet et al., 2014). In 2021 it was estimated that approximately 537 million adults were living with diabetes worldwide (International Diabetes Federation, 2021). Diabetes is an illness which is developed as a result of genetics and a variety of environmental factors. Therefore, while some individuals are more inclined to develop the disease than others, diabetes can to a large extent be prevented by living a healthy lifestyle (WHO, 2023). Machine learning methods can be used to detect individuals, who have the disease. The current paper employs the machine learning algorithms logistic regression and XGBoost to a dataset of demographic and medical information in order to classify whether people have diabetes or not. The performances of the two different algorithms are compared and it is evaluated whether it is preferable to have a transparent model (logistic regression) or have a black-box model (XGBoost) with slightly improved accuracy. Moreover, a hyperparameter tuning is performed on the XGBoost model, and the performance of the tuned model is compared to a model with default hyperparameter settings.

Diabetes – HS

Diabetes affects the regulation of blood glucose. The disease occurs when the body either cannot produce enough, or effectively use, insulin. Insulin is the hormone that regulates the level of glucose in the blood. Although it is possible to live with diabetes, the disease can have profound consequences, and should be discovered as soon as possible, so it can be treated and cause minimum damage to the body.

Diabetes can have severe to lethal consequences. Untreated diabetes can lead to hyperglycaemia, i.e., raised levels of blood sugar, which can negatively affect many of the body's systems. Especially the blood vessels in the heart, eyes, and kidneys are exposed to damage, which over time can cause permanent loss of vision and kidney and heart failure (WHO, 2023). Hypoglycaemia, i.e., low levels of blood sugar, is also a common effect of diabetes, and can lead to potentially life-threatening seizures and loss of consciousness (DiMeglio et al., 2018). Additionally, the poor blood flow and nerve damage caused by diabetes can cause foot ulcers, which may lead to amputation (WHO, 2023). In general, it is estimated that up to 1.5 million people each year die from diabetes, and a further half million

deaths are due to kidney diseases caused by diabetes. Furthermore, around 20 % of cardiovascular deaths are caused by raised blood glucose levels due to diabetes (WHO, 2023).

Diabetes is a worldwide health issue, but it is especially prevalent and growing in low- and middle-income countries (WHO, 2023). Furthermore, a study by Mayer-Davis and colleagues (2017) showed that diabetes amongst youths in the United States has increased significantly between 2002 and 2012, particularly among racial and ethnic minorities.

Types of Diabetes – LP

There are several types of diabetes. They all have similar consequences but arise from different issues. The two most common variations of diabetes are type 1 and type 2 (NIDDK, n.d.), which will be introduced thoroughly, but there are also other variations, which will only be mentioned briefly.

Type 1

Type 1 diabetes is caused by an autoimmune β-cell destruction, which usually leads to an absolute insulin deficiency (American Diabetes Association, 2017). It is commonly diagnosed by having unusually high blood glucose concentrations (DiMeglio et al., 2018).

Type 1 diabetes has traditionally been thought to have a juvenile onset and to be primarily caused by genetics. However, as knowledge about the disease has increased, it has been found that this is not an accurate paradigm, in fact up to 50% of type 1 diabetes onsets occur in adulthood (American Diabetes Association, 2017; DiMeglio et al., 2018).

Nonetheless, genetics do play an important role for type 1 diabetes, as it is a heritable polygenic disease. Thus, being born in a family with type 1 diabetes increases the lifetime risk of developing type 1 diabetes, with a 3% risk if the mother has it, 5 % if the father has, and 8% risk if a sibling has the disease (Pociot & Lernmark, 2016). Furthermore, there is a large variation of overall lifetime risk by country. The incidence rate is highest in Scandinavian countries, followed by Europe, North America and Australia, whereas type 1 diabetes is a rare disease in Asian countries. The disease is also slightly more common in males than females (DiMeglio et al., 2018; Katsarou et al., 2017).

It has been found that many environmental exposures also affect and are associated with the development of type 1 diabetes. Some highly relevant environmental risk factors are: diet, vitamin D sufficiency, and a decreased biodiversity in gut-microbiomes. Thus, development of type 1 diabetes is both influenced by genetic predispositions and susceptibility, as

well as lifestyle factors including diet, hygiene, and childhood infections (DiMeglio et al., 2018; Katsarou et al., 2017).

Type 1 diabetes is treated with regular blood glucose level checks and insulin injections when necessary. This treatment makes type 1 diabetes a treatable and liveable disease, however, the disease is still associated with considerable financial, medical and psychological burdens (DiMeglio et al., 2018).

Type 2 – HS

Type 2 diabetes is the most common type of diabetes, and it is estimated that approximately 90% of diabetics have this type of the disease (Chatterjee et al., 2017). Type 2 diabetes is caused by building up an insulin resistance, which leads to a progressive loss of insulin secretion from the β-cells (American Diabetes Association, 2017). This means that the body either cannot optimally use the insulin or cannot produce enough of it, which leads to raised blood glucose levels (Diabetesforeningen, n.d.-a). Development of type 2 diabetes depends on a mixture of being genetically predisposed to it and living an unhealthy lifestyle (Diabetesforeningen, n.d.-b; Zimmet et al., 2014). Being genetically predisposed means having a weak pancreas, which is the organ that produces insulin. However, it is the lifestyle factors that place a strain on the pancreas, which trigger whether or not you develop type 2 diabetes. Such factors include being overweight, having an inactive lifestyle, and smoking. Age is also a large risk factor, because the pancreas naturally produces less insulin with age (Diabetesforeningen, n.d.-b). Because of these risk factors, type 2 diabetes was traditionally thought to be a disorder of middle-aged and elderly people. However, the disease has become more common in children and adolescents too, especially in conjunction with the increase in childhood obesity (American Diabetes Association, 2017; DiMeglio et al., 2018; Zimmet et al., 2014).

Since type 2 diabetes is primarily triggered by external factors, it is possible to prevent or delay the onset of the disease with regular physical activity, a healthy diet, avoiding tobacco use, and by maintaining a healthy body weight. Furthermore, physical activity and diets can be used as treatment along with medication and regular screening, surveillance and treatment of complications (WHO, 2023)

Rare types of diabetes –LP

As mentioned, there are other, more rare types of diabetes. Gestational diabetes (GDM) is the third most common type, yet it is a temporary kind, since women can develop it during pregnancy, but it usually resolves itself shortly after delivery. However, having had gestational diabetes is associated with an increased risk of developing type 2 diabetes later in life. With detection and treatment this disease is usually harmless, however untreated GDM has been associated with an increased perinatal morbidity due to hyperglycaemia during pregnancy (NIDDK, n.d.; Reece et al., 2009). Other types of diabetes are: Maturity onset diabetes of the young, neonatal diabetes, Wolfram Syndrome, Alström Syndrome, Latent Autoimmune diabetes in adults, type 3c diabetes, steroid-induced diabetes, and cystic fibrosis diabetes. It is estimated that only about 2% of diabetics have these types (Diabetes UK, n.d.-a).

Machine learning – HS

Machine learning plays a large part in today's society. Machine learning is utilised in the autocorrect of your phone, in language translation apps, it is used to predict what you want to watch on Netflix and in a multitude of other technologies that are used by millions of people every day (Brown, 2021). Machine learning is a vast, growing subfield of the even larger field, artificial intelligence (AI). A survey by Deloitte from 2020 found that while 67% of responding companies were already using machine learning, 97% of respondents planned to do so by 2021 (Ammanath et al., 2020). Machine learning employs data and algorithms in order to make computers imitate the way humans learn. By using machine learning the computers are able to "learn" hidden patterns within the data without being explicitly programmed to recognize these patterns. Similar to humans, the computer will improve its accuracy gradually as a part of the learning process (IBM, 2016).

Supervised learning – LP

The field of machine learning typically distinguishes between four different techniques: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Which approach to use depends on the data available and on the specific problem you want to solve (Ray, 2019). Supervised learning requires labelled datasets in order to train the algorithm (Janiesch et al., 2021). Unsupervised learning algorithms on the other hand are able to analyse and cluster unlabelled data (IBM, 2023b). Semi-supervised learning entails that

only some of the input dataset is labelled (IBM, 2022a). Like supervised learning, reinforcement learning maps input to output. However, it does so using an interactive learning style in which it employs feedback from its surroundings via its own actions and experiences (Bhatt, 2019).

In the current paper we will be utilising supervised machine learning to classify whether people have diabetes or not based on selected biomarkers and relevant demographic information. Classification is a standard task in supervised machine learning in which the algorithm predicts which class each datapoint belongs to. The output is categorical, meaning it consists of a finite number of classes (Keita, 2022). Classification is simply a regression problem with a threshold determining which class something belongs to. In the current paper a binary classification task is used. Thus, if the prediction by the model is above the threshold, the datapoint will be classified as diabetic, if the prediction is below the threshold, it will be classified as not-diabetic (Gong, 2022).

As aforementioned, performing supervised machine learning requires that you have data input as well as labelled "targets" (i.e., the correct answer) to train the model. The model will learn hidden patterns in the data that connect the input data to the correct target outputs. After the training process the model should be able to correctly label most unseen data (Janiesch et al., 2021).

Supervised machine learning can be applied to a vast variety of algorithms. Some of the most widely used groups of supervised learning algorithms include linear models, neural networks, and tree-based algorithms. Examples of linear models include, linear regression, in which the relationship between variables is investigated, and logistic regression, which is similar to linear regression, but used on categorical dependent variables (IBM, 2023a) . Neural networks are models that are organised in networks of *neurons* which are supposed to replicate the learning capabilities of the human brain (IBM, 2023b). Neural networks can, like the linear models, belong to the branch of machine learning called *shallow learning*, however, they often have multiple layers of neurons making them a *deep learning* mechanism (Janiesch et al., 2021). Tree-based algorithms consist of tree-like structures of if-else statements used in order to decide which target label is most likely. The decision happens at the bottom of the if-else statements in the *leaves* of the decision tree (Gong, 2022). For the purpose of this assignment, we will compare a logistic regression model and a tree-based algorithm called XGBoost, on their performance and interpretability on a binary classification task.

Parameters and hyperparameters – HS

In machine learning, there are values which you choose for the model before training it, and values which the model should estimate during training. The first are called hyperparameters, while the latter are called parameters (Rouse, 2023). The goal of machine learning is to find the model that fits your data best, while also generalising to new data. This means that a machine learning process aims to find the optimal model, and thus the optimal parameters for this model. Parameters are e.g., model coefficients, weights, and biases. When the model is training, it continuously updates the parameters, as the algorithm tries to learn the pattern between the input variables and the target variable. Which parameters are the best ones is estimated by using a loss function e.g., gradient descent or sum of squared errors. When training is finished, the output is a model with the best parameter values. This is the model, which is evaluated on your test dataset, to give a final performance score (Nyuytiymbiy, 2020).

How the model should estimate, update and learn the best parameters depends on which hyperparameters the learning algorithm has. As the prefix 'hyper-' suggests, these are parameters at a level above the model parameters. Hyperparameters control the learning process of the algorithm, and they are values which are set before training on data begins. Common examples of hyperparameters include: train-test split ratio, tree depth, number of hidden layers in a neural network etc. Hyperparameters are not a part of the final model, and from looking at the final model it is not possible to know which hyperparameters it was trained with. Importantly, hyperparameters are not updated during training, they have the same values before and after the learning process (Nyuytiymbiy, 2020). The number of hyperparameters in a machine learning algorithm can vary vastly and there can be a great number of combinations of them (IBM, 2022b). Since the process of finding the optimal parameters depends on which hyperparameters your learning model has, it is important to choose the best hyperparameters for your model. The process of finding and choosing the right ones is called hyperparameter tuning or hyperparameter optimisation (Nyuytiymbiy, 2020; Rouse, 2023).

Description of dataset – LP

The dataset used for this paper was made with the purpose of predicting diabetes. It was created by Mohammed Mustafa (2023a) and was retrieved from Kaggle.com on April 17th 2023. The data contains nine different variables, and has 100,000 observations, where each observation corresponds to a person. The variables are a mixture of demographic variables, diabetes risk factors and a binary variable indicating whether or not a person has diabetes.

The demographic variables are gender and age. The diabetes risk factors are hypertension, heart disease, smoking history, body mass index (BMI), haemoglobin A1c level, and blood glucose level. Hypertension and heart disease are binary variables, indicating whether or not a person has these medical conditions. BMI, blood glucose level, and haemoglobin A1c level — which is a person's average blood glucose level measured over 2-3 months (Diabetes UK, n.d.-b)— are all numerical variables within their respective scales. The last risk factor, smoking history, is a categorical variable with six categories: never, former, not current, current, ever, and no information. The five first categories mentioned are ordered to span from never having smoked to smoking a lot. Please note that type of diabetes is not specified in the dataset.

Research statement

Considering that diabetes is such a prevalent worldwide health issue that has severe and lethal consequences, as well as being both a personal and societal expense, it is important to try preventing the development of the disease. Given that machine learning is used to find patterns in complex data, it can be a useful tool in predicting whether people have diabetes or not, based on information about known risk factors of developing the disease. Accurate prediction models can then be used as a tool to detect people who are at risk of developing, in order to help prevent this from happening. There are many machine learning methods and algorithms, and some are easier to understand than others. The aim of this paper is to try to make the most accurate prediction model for predicting diabetes, as well as to discuss the trade-off between having a very accurately predicting model that you do not necessarily understand the underlying mechanisms of, versus a model you do understand with a slightly worse prediction accuracy, when working with sensitive medical data.

Methods – HS

When performing supervised machine learning the data should be parted into a training dataset and a test dataset. The training set is used to fit the model, subsequently, the test set is used to evaluate whether the model is merely describing the training set or whether the parameters found can be used efficiently for unseen data.

When performing hyperparameter tuning the dataset will be split into an additional set. Thus, there will be a training dataset, a validation dataset, and a test dataset. The training

dataset is used to fit the model. After the model has learned from the training data, the validation data is used to evaluate the model fit and to decide whether to change the hyperparameters of the algorithm. This process is repeated until you are convinced you have found optimal hyperparameters that allows the model to learn appropriate parameters, which are both good at describing the training data but more importantly can generalise to new data. Only then, once you have settled on a set of hyperparameters, should the test data be used to evaluate model performance. Thus, the test dataset is not a part of the development of the model. This process is mainly performed to avoid overfitting (Shah, 2020).

Logistic regression – LP

Logistic regression is often used for binary classification tasks. A logistic regression algorithm first produces a linear regression, then uses this as input in a sigmoid function. The sigmoid function squishes real numbers to map between 0 and 1, so 0 in the log space is equal to $-\infty$ and 1 is equal to ∞. Now that the output is restricted between 0 and 1 it can be used as a probability that the datapoint belongs to a specific category (Geeks for Geeks, 2023). When logistic regression is used in machine learning there are two phases: training and test.

During the training phase, stochastic gradient descent is used to learn an intercept and which weights to give each of the features in the dataset. The weights are organised as a vector of real numbers. Each number is tied to a feature and signifies how important the given feature is for the classification decision. The weights work as the beta coefficients in a regular linear regression. The optimal weights are learned by minimising a loss function. Gradient descent is a method for finding the minimum of a function. The algorithm is "placed" on the loss function and based on the slope of the gradient it "decides" whether to move one way or the other. The algorithm then takes a *step* in the descending direction. The size of the step depends on a learning rate. If the learning rate is too high the algorithm might miss the minimum, while if it is too small it will be very time consuming to reach the minimum (Jurafsky & Martin, 2023). Therefore, gradient descent integrates an adaptive learning method, so the *step size* is dependent on the slope of the gradient. The learning rate is simply multiplied by the gradient. This means that the steeper the slope, the larger the step size, once the function approaches the minimum and the slope becomes more horizontal the algorithm will take smaller steps (Srinivasan, 2019). The algorithm might take a step, which is too large and end on the opposite slope. If this happens, it will take a smaller step back from where it came.

This technique could lead to getting stuck in a local minimum, however since the logistic regression uses a convex loss-function, the loss-function has only one minimum, the global minimum.

In the testing phase the algorithm multiplies the weights with each of the features of the data point and adds this to the intercept. The output is a real number between 0 and 1, which signifies how probable it is that the data point belongs to the target class. Note that while the decision boundary is usually set to 0.5, it can take any value between 0 and 1. The performance of the model can then be estimated based on the fraction of correct classifications (Jurafsky & Martin, 2023).

XGBoost – HS

XGBoost (Chen & Guestrin, 2016) stands for Extreme Gradient Boosting and is a tree-based model of the type *ensemble decision trees*. As aforementioned, a tree-based model is a nested arrangement of if-else statements. Ensemble decision trees entail that there are more than one decision tree, and that the trees complement each other and all inform the final decision (Guestrin & Chen, 2022b). This combination of several models into an ensemble is also called “boosting” hence the name XGBoost (Data Base Camp, 2023).

XGBoost consists of a combination of trees called CART (Classification and Regression Trees). CARTs are different from ordinary decision trees since they assign a real score to each of the leaves in the tree. This produces a more detailed interpretation than a simple classification. The training process in XGBoost is used to learn the optimal structure of trees and the optimal score for each leaf. The training is performed additively, meaning the model learns the structure and scores of one tree, it holds this structure fixed and then adds new trees one by one (Guestrin & Chen, 2022b). A new tree is only trained on the data which was misclassified using the previous ensemble of trees. This chain is continued until all training data is classified correctly. Thus, the ensemble of trees compensates for what the individual models are lacking by attempting to predict the error of the previous model. In order to build the optimal next model to add to the ensemble a loss function is minimised. This process is called gradient boosting (Data Base Camp, 2023). As aforementioned loss functions are often involved in machine learning to minimise the error during training of a model, however, the mechanisms involved in learning the tree structure are far more complex than an ordinary optimization problem where you take the gradient, such as the method used in logistic regression (Guestrin & Chen, 2022b). XGBoost is often referred to as a *black-box* model, because

of its complexity and how difficult it is to interpret how it makes predictions (Data Base Camp, 2023).

Another important feature of XGBoost is that it contains a regularisation term which controls for model complexity. In other tree-based models, complexity is often left to be described by heuristics, however XGBoost includes a mathematical term to define it formally. The complexity term is included in a "structure score" which is used to evaluate how good a tree structure is. The structure score is equivalent to the impurity measure on a regular decision tree except for the fact that it takes model complexity into account (Guestrin & Chen, 2022b). Thus, the lower structure score, the better the split, and the lower probability of misclassification (Loaiza, 2020).

Bayesian optimisation and HGBoost – LP

As aforementioned, hyperparameter tuning aims to find the optimal architecture for the model to learn parameter values. There are several types of hyperparameter tuning, the most common of which are manual search, random search, and grid search. Manual search is simply manually manipulating the hyperparameters using trial and error. Random search explores the hyperparameter space using random sampling. Grid search, on the other hand, is a very systematic way of exploring the hyperparameter space in which all possible combinations of selected hyperparameter values are evaluated (Jordan, 2017). Bayesian optimisation is a more complex method for hyperparameter tuning. Bayesian optimisation is more efficient than the standard methods mentioned above because it takes past performance into account when determining which hyperparameters to evaluate. Bayesian optimization uses Bayes theorem (Joyce, 2021) to build a probability model, which it then uses to search the hyperparameter space and ultimately select the ideal hyperparameter settings (Wang, 2022). At every step, the model updates the probability model according to Bayes' rule in order to determine which point in the parameter space is most useful to evaluate next. Bayesian optimisation is ideal to use for black-box functions, such as XGBoost, because it is unclear which effect changing the different hyperparameters has (Agnihotri & Batra, 2020).

In the current paper, HGBoost was used to apply Bayesian optimisation to XGBoost. HGBoost (Taskesen, 2020) stands for Hyperoptimized Gradient Boosting and is a python library using Bayesian optimisation to tune the hyperparameters in gradient boosting models. An inner loop uses *k*-fold cross validation on the training dataset to choose the hyperparameter settings which best fit the training data. The *n* best models are then evaluated using *k*-fold

cross validation on the validation dataset. This is done to avoid overfitting by finding out which one of the *n* best models is the most generalisable. The best model is chosen and is evaluated on the test dataset. The accuracy of the model is then assessed. The function outputs the hyperparameter settings of the best model. Finally, XGBoost should be fit on the entire dataset with the given hyperparameter settings and evaluated as normal (Taskesen, 2022).

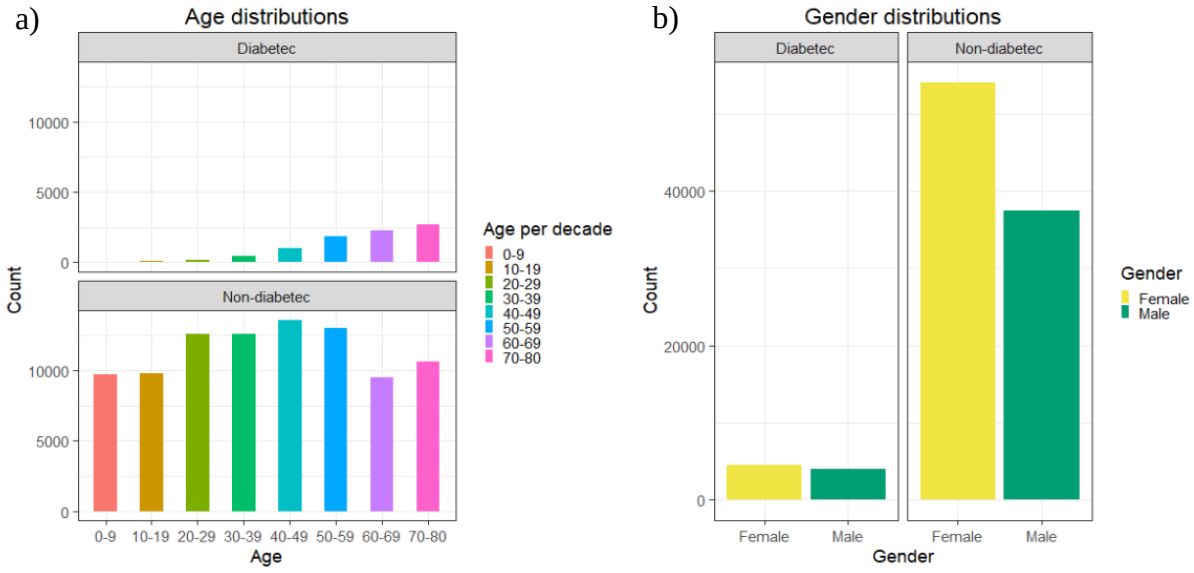
Analysis

Data exploration and cleaning – HS

After downloading the data, the data was inspected and explored in RStudio (R Core Team, 2022). It was found that there is an uneven distribution between diabetics and non-diabetics. Out of the 100,000 people in the dataset 8,500 are diabetics, while the remaining 91,500 people are not diabetic. When inspecting the gender variable, it was found that there are three categories: male, female, and other. However, only 18 people identified as other, and none of those were diabetics, so it was decided to exclude these data points from the dataset to simplify the gender variable. This resulted in 91,482 data points of non-diabetics. In the whole dataset, the two remaining gender variables contained 58,552 females, hereof 4,461 diabetics, and 41,430 males, hereof 4,039 diabetics. See Figure 1b for a visualisation of the gender distribution. The age variable was distributed across a whole lifespan for both diabetics and non-diabetics. The age for diabetics had a minimum of 3 years and a maximum of 80 years (mean ≈ 61, sd = 14.55). Non-diabetics had a minimum age of 0.08, corresponding to 1 month old (Mustafa, 2023b), and a maximum age of 80 years (mean ≈ 40, sd = 22.30). See Figure 1a for a visualisation of the age distribution. The cleaned dataset was saved as a .csv file to be used for the classification analysis, which was performed using Python 3.10.4 (Van Rossum & Drake, 2022).

Figure 1

Age and gender distributions for diabetics and non-diabetics.



Procedure – LP

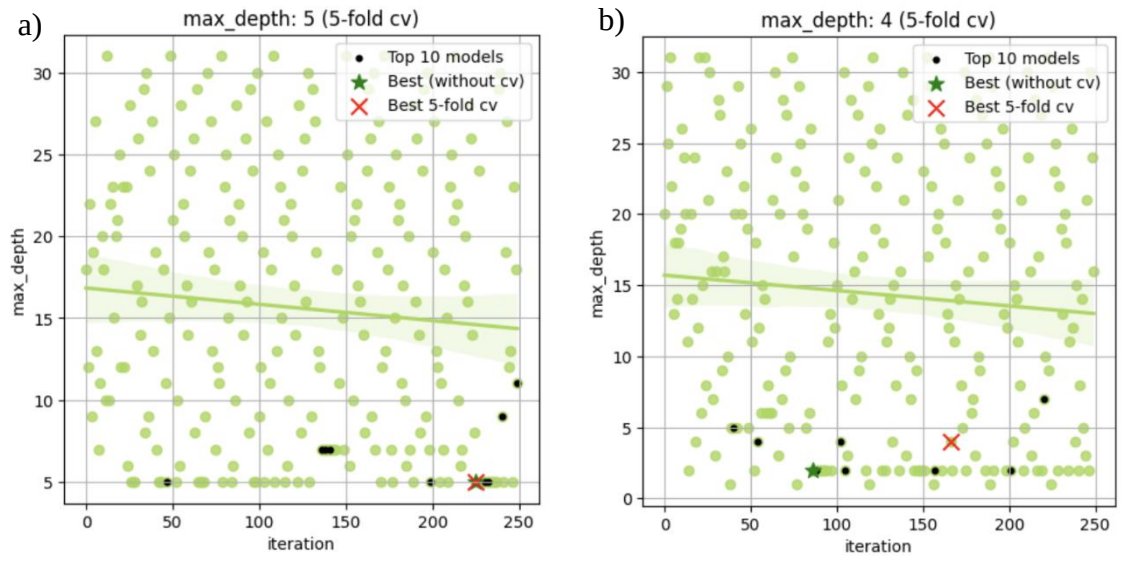
The cleaned dataset was imported in Visual Studio Code (Microsoft, 2023), where a few instances of additional data cleaning was necessary. The dataset was first parted into an array of predictor variables, X, and a one-dimensional target variable, Y. Both arrays were then split into a test and a training version using scikit-learn (Pedregosa et al., 2011). A random seed was specified in order to have reproducible results. Two instances of XGBoost were run: One using the default settings, and one using the settings used by the developers of XGBoost in their guide "Get started with XGBoost" (Guestrin & Chen, 2022a). Moreover, one instance of logistic regression was run using default settings. Performance of the classifiers were assessed, where both XGBoost classifiers outperformed the logistic regression classifier (see details in the Results section below).

In order to further improve the XGBoost classifier, Bayesian optimisation was used to tune the hyperparameters of the algorithm with HGBoost. The hyperparameters in HGBoost were held to the default except it was chosen to evaluate the top 20 models with cross validation instead of the standard 10 best models. This means that 250 different combinations of hyperparameters were run using 5-fold cross validation on a data split where 20% was used for validation and 20% was used for testing, leaving 60 % for the training. A random seed was set in order to have reproducibility of results. After fitting all 250 models, the hyperparameter

tuning was plotted. The plots show the space in which each of the hyperparameters were explored. All models were plotted as a point in the space with the 10 best models highlighted. Most of the plots looked fine, however for the hyperparameter "max_depth" it seemed like the model wanted to explore space below the limit which was set by the developers of HGBoost (see figure 2a). In the HGBoost package it is not possible to manually set the hyperparameter space to be explored. Therefore, it was necessary to copy and slightly change the source code of the package in order to extend the search space. Thus, the hgboost_mod.py file in the Python repository is an exact replica of the HGBoost package, except that the exploratory space of the max_depth hyperparameter has been extended from the original 5-32 to 1-32. Figure 2b shows the exploration of the max_depth space after the package was modified. It is clear from the plot that the model benefitted from this extension of the search space, as 8 out of the 10 best models have a max_depth setting below 5 (figure 2b). HGBoost produces a set of hyperparameter settings which was evaluated to be the best performing model on the unseen validation set. These hyperparameter settings were then used to train a final XGBoost classifier. This XGBoost classifier was run on the same data split which was used to train and evaluate the initial classifiers.

Figure 2

Hyperparameter space of max depth



Note: The figure shows the model iterations with two different settings of search space for the max_depth hyperparameter: a) shows the original settings, b) shows the modified settings. NB: there is a change of scale on the y-axis between the two plots.

Results – HS

Classification performances

The classification performances of the algorithms reported in prose are also reported in Table 1 for clarity. The logistic regression classification algorithm performed with an accuracy score of 96.1994 %. The XGBoost classification algorithm with default hyperparameter settings performed with an accuracy score of 97.2136 %. The XGBoost classification algorithm with the developer suggested hyperparameter settings performed with an accuracy score of 97.2996%. The XGBoost classification algorithm with the optimised hyperparameter settings performed with an accuracy score of 97.3346%.

Table 1

Model	**Performance**	**Percentage of incorrect predictions**	**Number of incorrect predictions in test dataset**
Logistic regression	96.1994 %	3.8006 %	760
Default XGBoost	97.2136 %	2.7864 %	557
Developer suggested XGBoost	97.2996 %	2.7004 %	540
Optimised XGBoost	97.3346 %	2.6654 %	533

Logistic regression model

After training the logistic regression classifier, the resulting model has the following intercept and weights:

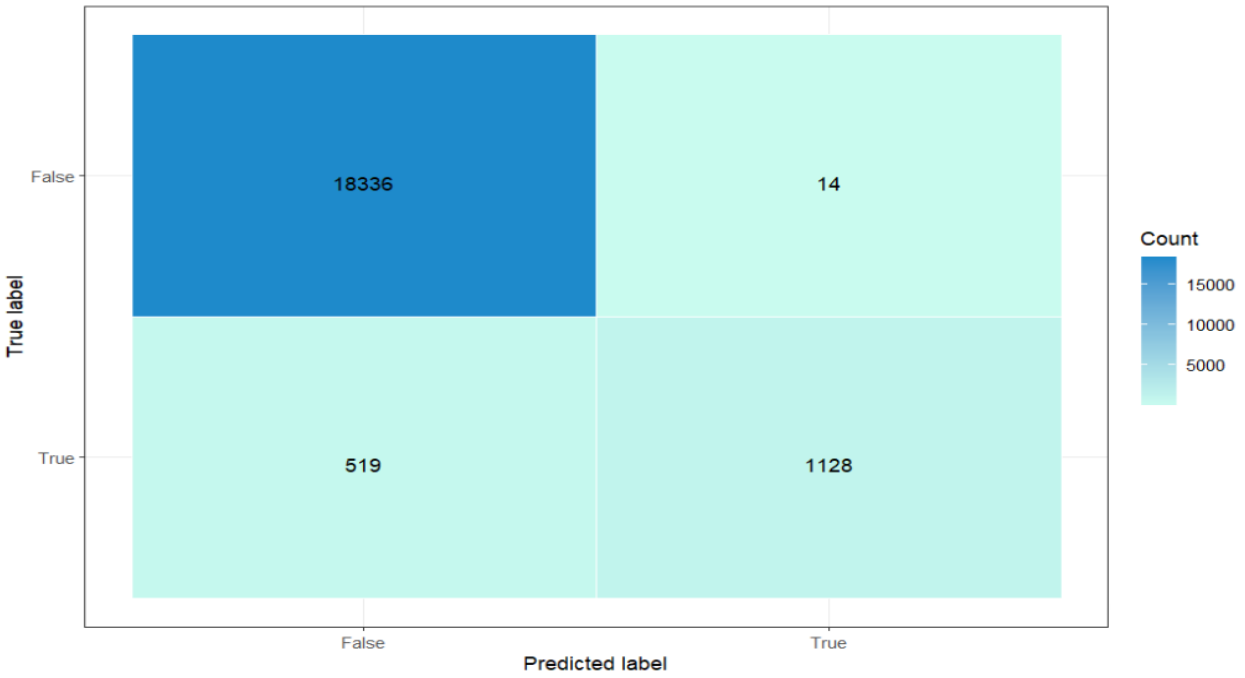
$$Diabetes \sim -27.325 + 0.306 \cdot gender + 0.047 \cdot age + 0.760 \cdot hypertension + 0.770 \cdot heart\ disease - 0.076 \cdot smoking\ history + 0.092 \cdot BMI + 2.345 \cdot HbA1c\ level + 0.033 \cdot blood\ glucose\ level$$

Bayesian optimised XGBoost

Figure 3 showcases how the optimised XGBoost model's predictions are labelled, as well as the true labels on the test dataset. From this it is clear that the model makes more false negatives (2.60%) than false positives (0.07%).

Figure 3

Bayesian optimised model performance on test dataset

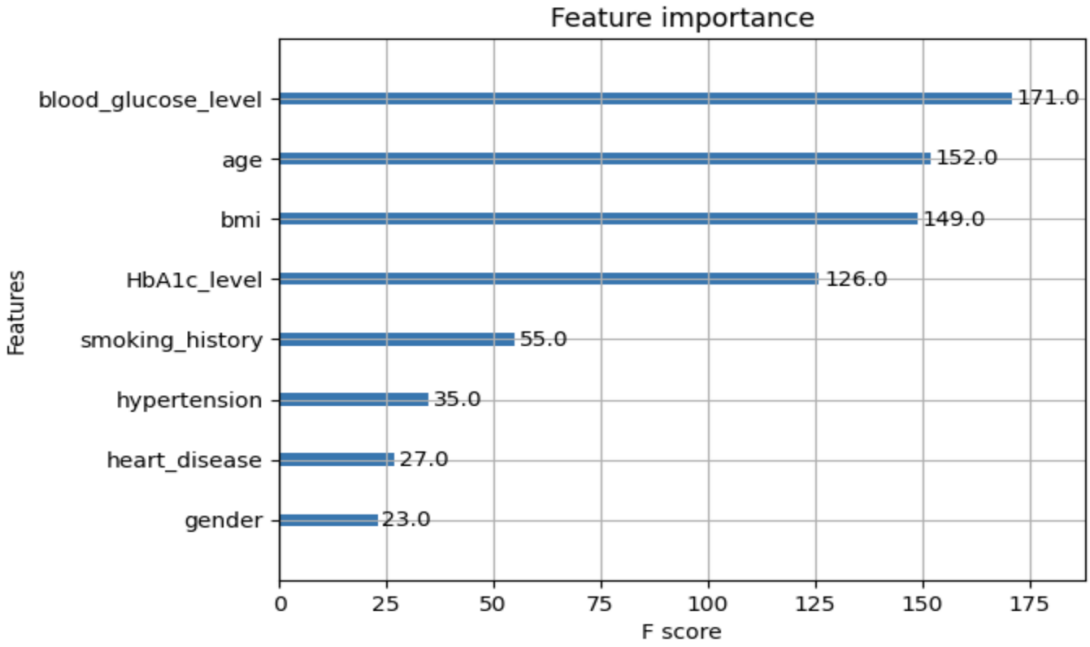


Note: Confusion matrix displaying predicted vs. true labels

Figure 4 shows how important the different variables were for the construction of the model. The larger the number, the more often a variable was used to make key decisions in splits of the decision tree (Brownlee, 2016).

Figure 4

Feature Importance for Bayesian optimised model



Discussion

Discussion of results – LP

As reported in the results section (Table 1), all four classifiers presented in this paper classified diabetes patients with between 96% and 97.4% accuracy. There was slightly more than 1.1% difference in performance between the best (the Bayesian optimised XGBoost) and the worst performing classifier (the logistic regression), corresponding to 1135 more subjects being classified correctly using the optimised XGBoost. In fact, all the XGBoost classifiers outperformed the logistic regression classifier with at least 1%. However, it should be noted that we did not attempt to optimise the hyperparameters of the logistic regression classifier, since this was out of the scope of this paper. It is therefore a possibility that the performance of the logistic regression classifier could have been improved enough to level with those of XGBoost. Between the XGBoost classifiers the difference in performance was very small, all lying between 2.66 and 2.79 percent incorrect classifications. This corresponds to a difference of 120 subjects out of the total dataset. Since the difference in performance is so small, it should be considered whether the results were merely due to chance and therefore not robust across other splits in data.

For medical datasets like the one used in the current paper it is essential whether the incorrect classifications are false negatives or false positives. In cases such as this one, where an illness is the subject of the analysis it is preferable to have false positives as compared to false negatives. The emotional stress of falsely being classified as suffering from an illness is not as great an issue as having the disease and it not being discovered. If a patient is falsely classified as healthy, they might not receive treatment, which could have detrimental effects for the individual. Whereas if a patient is falsely classified as ill, more tests will be performed, and it will be discovered that the initial result was wrong. As reported in the results section (figure 3), the best performing classifier made 533 wrong classifications when tested on the test dataset. Of the 533 wrong classifications, 519 were false negatives, while only 14 were false positives. This means that the classifier is not particularly sensitive to positive cases of diabetes since it much more frequently classifies ill people as being healthy than vice versa. This could be a result of the unbalanced dataset, as there are many more instances of healthy individuals. Therefore, the classifier will easily gain prediction accuracy by simply guessing that the individual is healthy. As explained above this is an undesired bias, as the

model has only captured 1,128 out of 1,647 corresponding to 68.5 % of diabetics in the test dataset.

Feature importance – HS

Figure 4 showcases feature importance for the different variables in the dataset for the best performing model. The highest-ranking feature (blood glucose level) and the fourth (HbA1c level) most important feature are related, as they both pertain to blood glucose levels. It is a common feature of diabetics to have higher blood glucose levels, since their body is not able to effectively use the glucose, and it therefore accumulates in the blood (Diabetes UK, n.d.-b; Seery, 2023). The fact that the model finds these features relevant for predicting diabetes fits well with what is known about diabetes and the risk factors as well as symptoms of the disease.

Age ranks second highest, meaning it is also a highly influential variable. This is sensible, since it is known that age in itself is a risk factor of developing diabetes, especially type 2, as mentioned in the introduction. In general, diabetes is more common for people older than 45 than in children and young adults (*Type 2 Diabetes*, 2023) and as can be seen in figure 1a, there are also more adults and elderly in the dataset who have diabetes.

BMI is the third most important feature. Overweight is a substantial risk factor of developing diabetes (Polsky & Ellis, 2015; WHO, 2023). Being obese strains the body and organs, as it elevates non-esterified fatty acid, plasma leptin, and tumour necrosis factor-α levels, which are all influential in causing insulin resistance (Leong & Wilding, 1999). Up to 90% of type 2 diabetes cases can be attributed to excess weight (Hossain et al., 2007), and it is therefore sensible that the BMI variable ranks highly in prediction importance.

Although smoking is a large risk factor of diabetes too (Fagard & Nilsson, 2009; WHO, 2023), it only ranks as the fifth most important variable in the model. However, this might be due to the data, since 35.8 % of the people in the dataset were categorised as “no info” in the smoking variable.

It seems the model has picked up some patterns in the data which align well with the research on diabetes. However, given that XGBoost is a bit of a black box, it is uncertain how the model has learned this, and what exactly it means. Furthermore, it is important to be aware that a feature can be valued as important for a model for several reasons. If a variable contains some very influential extremes, this might influence how important it is for the model, whereas it might not generally be a large risk factor. As mentioned with the smoking

variable, it is also partly a reflection of the data structure, so given a new dataset, the variables might rank differently in importance. The different variables might also work in interaction with each other, which a feature importance measure might not capture.

Advantages and limitations of the algorithms – LP

As aforementioned, the XGBoost classifiers outperformed the logistic regression classifier. An additional advantage of XGBoost, apart from the improved accuracy, is that it can capture nonlinear relationships between the input and output variables, whereas the logistic regression algorithm assumes linearity. Moreover, XGBoost is very flexible and can be applied to a variety of different machine learning problems, whereas logistic regression is limited to categorical output (Cotton, 2022; Data Base Camp, 2023).

Both for the logistic regression and for the XGBoost there is a chance of overfitting (Cotton, 2022). XGBoost is particularly delicate to overfitting when the hyperparameters are not properly tuned. Another issue with using a black-box model such as XGBoost is that the model and its predictions are not easily interpretable (Data Base Camp, 2023). This is particularly important in some medical machine learning tasks, where interpretability is central and could inform the treatments and understanding of symptoms. The logistic regression classifier on the other hand is based on straightforward mathematical weights, which are more easily understood. XGBoost has also been found to perform worse on imbalanced datasets, i.e., when there is an unequal number of instances of each of the target classes. This is because the model can become biased and perform poorly on the class with a low number of occurrences (Data Base Camp, 2023). While the current paper presented an imbalanced dataset, the performance of XGBoost was still very good. However, others working with the same dataset have achieved an even higher accuracy (Area under the curve (AUC) score of up to 99.5%) when balancing the data by omitting most of the non-diabetics from the data (MSS-Suman, 2023). For the current analysis this was purposefully avoided. A balanced dataset is not realistic in the case of diabetes prediction. If doctors were to rely on a machine learning algorithm to help in preventing diabetes, the data would have an imbalance like the one found in the current dataset as there are more healthy individuals than diabetics. In fact, the dataset represents the real-world balance of the two target categories quite well; there are 8.5% diabetics in the dataset, while it was estimated in 2019, that 9.3% of the world population were diabetic (Saeedi et al., 2019).

Finally, it should be mentioned that using XGBoost requires rigorous pre-processing of data as it is very sensitive to missing values and relies on categorical variables being encoded appropriately. Deficient data pre-processing could therefore lead to poor results (Data Base Camp, 2023). This is not an issue for the current analysis, however, as the dataset contained no missing values, and pre-processing entailed ensuring that the categorical variables were marked as factors.

Limitations of the dataset – HS

The dataset which was used for this paper was quite clean and well structured. However, there are some limitations and missing information, which could improve the machine learning process even further. Firstly, it is not optimal that the dataset does not include information about which type of diabetes the people have. Even though the types share a lot of characteristics, there are some differences. BMI e.g., is a more indicative variable for type 2 diabetes, than type 1 (Johns Hopkins, 2023). Having this information would enable the model to learn this distinction too. Secondly, there are other risk factors, which are important, that would have been preferable to have included. As mentioned in the introduction, the development of diabetes, and especially type 1 diabetes, is influenced by genetic history, i.e., are there other people in your family with diabetes, or with weak pancreas. Combining this with information regarding diabetes types, could enable more distinct patterns in the data to be found. Lastly, the models could benefit from additional demographic information, such as racial and geographic information, since there are variations in how common diabetes is across these variables, as mentioned in the introduction. In general, machine learning benefits from large datasets with meaningful variables (Chawla, 2020), so having more data points, especially in the diabetic category, and additional important variables included in the dataset, could improve predictions further.

Discussion of research statement – LP

When training models to predict diabetes, both the logistic regression and XGBoost models had an accuracy above 96%. However, all versions of the XGBoost algorithm performed at least 1% better than the logistic regression classifier. In this dataset, 1% corresponds to 1000 people, which means that by using an XGBoost classifier, there would be approximately 1000 more people, who were classified correctly, than by using the logistic regression model. However, the logistic regression model is easy to understand. The training process outputs a

model with meaningful parameters: an intercept and weights for the variables (see Results section). Given that the variables of this dataset are on various scales, the weights are not directly interpretable. However, if all variables were standardised before training the logistic classifier, it would be possible to see which variables were afforded most weight in the classification and are therefore important variables.

XGBoost does not output an equation for the model since the structure of the model is too complex. Therefore, it is more difficult to know what the model looks like, and how it was created. However, it is possible to access information about feature importance and rates of false positives and negatives (see Results), which offer an insight into the underlying model, and how important each variable is.

Consequently, it seems there is a trade-off between having a classifier you understand the underlying mechanisms of and having the most accurate classification model. What is preferable depends on the purpose of the model. For the current analysis, the purpose is to correctly predict diabetes, rather than discover possible risk factors of diabetes. The dataset contains solely well-established risk factors of developing diabetes, so there has not been a focus on understanding all the weightings. Therefore, XGBoost is preferable for this cause since it provides more accurate predictions than the logistic regression model.

HS – The three XGBoost models performed almost equally well, regardless of the hyperparameter tuning. Hyperparameter tuning is quite time consuming and computationally heavy. Therefore, it should be considered whether the process and computational costs of Bayesian optimisation is worth doing. Using the current data, the classifier with hyperparameter tuning did perform slightly better than the default classifier, however the difference was so small that it could just be due to chance. Thus, in this case it did not improve the model considerably to perform Bayesian optimisation. It could be argued that there is something gained from the process as it is preferable to know that the hyperparameters are sensible, as opposed to not doing hyperparameter tuning and knowing that the model could be substantially improved. Bayesian optimisation is preferable to other methods of hyperparameter tuning. It is not as exhaustive as grid search and is more computationally heavy than random search, however the search of the hyperparameter space in Bayesian optimisation is guided by Bayes theorem and is therefore more efficient with its use of computational power.

A diabetes classifier could potentially be used as a prediction tool to detect people who are at risk of developing diabetes. The people at risk could be informed of prevention methods to avoid ever developing the disease. However, the classifier presented in this paper was not very sensitive to positive cases and predicted a great deal of false negatives, which is

not optimal. It is important to note that such a classifier should not stand alone but could be used to flag the at-risk person's electronic health record, so their doctor is informed of the increased risk of developing diabetes.

Suggestions for future research – HS

Future research into predicting diabetes could be improved with a larger dataset. Having more variables in the data, which are meaningful for diabetes detection, could also improve the accuracy score. Furthermore, XGBoost might not be the best model for this dataset due to the unbalanced target variable, so it could be an option to try with additional machine learning algorithms and tune these hyperparameters too. In line with this, it could also be beneficial to perform hyperparameter optimisation on the logistic regression classifier, to further improve its performance.

Conclusion

In this paper, machine learning was used to classify diabetes. Two different algorithms, logistic regression and XGBoost, were applied. In total four different classifiers were trained: A logistic regression classifier and three XGBoost classifiers. The difference between the XGBoost classifiers was solely the hyperparameter settings: one model was run with default hyperparameter settings, one was run with settings suggested by the authors of XGBoost, and the final model was tuned using Bayesian optimisation to find the optimal hyperparameter settings. The performance and applicability of the algorithms were compared and discussed. It was found that the XGBoost classifiers outperformed the logistic regression classifier, and it was argued that for the current analysis, the additional performance accuracy of XGBoost outweighed the fact that this model is more difficult to interpret. Between the XGBoost classifiers the difference in performance was very small and could be due to chance. It was therefore discussed whether hyperparameter tuning was worthwhile. The best performing classifier was further explored, and it was found that it had identified the most prominent risk factors. Thus, the features utilised for classifications were in line with medical research of diabetes.

References

Agnihotri, A., & Batra, N. (2020). Exploring bayesian optimization. *Distill*, *5*(5), e26.

American Diabetes Association. (2017). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. *Diabetes Care*, *41*(Supplement_1), S13–S27. https://doi.org/10.2337/dc18-S002

Ammanath, B., Jarvis, D., & Hupfer, S. (2020). Thriving in the era of pervasive AI. *Deloitte, Rep*.

Bhatt, S. (2019, April 19). *Reinforcement Learning 101*. Medium. https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

Brown, S. (2021, April 21). *Machine learning, explained | MIT Sloan*. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Brownlee, J. (2016, August 30). Feature Importance and Feature Selection With XGBoost in Python. *MachineLearningMastery.Com*. https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/

Chatterjee, S., Khunti, K., & Davies, M. J. (2017). Type 2 diabetes. *The Lancet*, *389*(10085), 2239–2251. https://doi.org/10.1016/S0140-6736(17)30058-2

Chawla, V. (2020, September 16). *Is More Data Always Better For Building Analytics Models?* Analytics India Magazine. https://analyticsindiamag.com/is-more-data-always-better-for-building-analytics-models/

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cotton, R. (2022). *Machine Learning Cheat Sheet*. https://www.datacamp.com/cheat-sheet/machine-learning-cheat-sheet

Data Base Camp. (2023, February 25). *What is XGBoost? | Data Basecamp*. https://database-camp.de/en/ml/xgboost-en

Diabetes UK. (n.d.-a). *Types of diabetes*. Diabetes UK. Retrieved 22 May 2023, from https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes

Diabetes UK. (n.d.-b). *What is HbA1c?* Diabetes UK. Retrieved 28 May 2023, from https://www.diabetes.org.uk/guide-to-diabetes/managing-your-diabetes/hba1c

Diabetesforeningen. (n.d.-a). *Fakta om type 2-diabetes | Hvad er type 2-diabetes?* Retrieved 22 May 2023, from https://diabetes.dk/diabetes-2/fakta-om-type-2

Diabetesforeningen. (n.d.-b). *Gener og livsstil kan udløse type 2-diabetes*. Retrieved 22 May 2023, from https://diabetes.dk/diabetes-2/fakta-om-type-2/symptomer-og-arsager/ar-sager-til-type-2-diabetes

DiMeglio, L. A., Evans-Molina, C., & Oram, R. A. (2018). Type 1 diabetes. *The Lancet*, *391*(10138), 2449–2462. https://doi.org/10.1016/S0140-6736(18)31320-5

Fagard, R. H., & Nilsson, P. M. (2009). Smoking and diabetes—The double health hazard! *Primary Care Diabetes*, *3*(4), 205–209. https://doi.org/10.1016/j.pcd.2009.09.003

Geeks for Geeks. (2023, March). *Logistic Regression in Machine Learning*. https://www.geeksforgeeks.org/understanding-logistic-regression/

Gong, D. (2022, July 12). *Top 6 Machine Learning Algorithms for Classification*. Medium. https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

Guestrin, C., & Chen, T. (2022a). *Get Started with XGBoost—Xgboost 1.7.5 documentation*. https://xgboost.readthedocs.io/en/stable/get_started.html#get-started-with-xgboost

Guestrin, C., & Chen, T. (2022b). *Introduction to Boosted Trees—Xgboost 1.7.5 documentation*. https://xgboost.readthedocs.io/en/stable/tutorials/model.html

Hossain, P., Kawar, B., & El Nahas, M. (2007). Obesity and Diabetes in the Developing

World—A Growing Challenge. *New England Journal of Medicine*, *356*(3), 213–215. https://doi.org/10.1056/NEJMp068177

IBM. (2016). *What is Machine Learning? | IBM*. https://www.ibm.com/topics/machine-learning

IBM. (2022a, November 15). *Supervised vs. Unsupervised Learning: What's the Difference?* https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

IBM. (2022b, December 8). *IBM Hyperparameter tuning with Deep Learning Impact*. https://www.ibm.com/docs/en/wmla/1.2.3?topic=features-hyperparameter-tuning

IBM. (2023a). *What is Supervised Learning? | IBM*. https://www.ibm.com/topics/supervised-learning

IBM. (2023b). *What is Unsupervised Learning? | IBM*. https://www.ibm.com/topics/unsupervised-learning

International Diabetes Federation. (2021, December 9). *Facts & figures*. https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695.

Johns Hopkins. (2023, February 13). *Overweight and Obesity in People With Type 1 Diabetes Nearly Same as General Population | Johns Hopkins | Bloomberg School of Public Health*. https://publichealth.jhu.edu/2023/overweight-and-obesity-in-people-with-type-1-diabetes-nearly-same-as-general-population

Jordan, J. (2017, November 2). *Hyperparameter tuning for machine learning models.* https://www.jeremyjordan.me/hyperparameter-tuning/

Joyce, J. (2021). Bayes' Theorem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem/

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing [preprint]* (3. Edition draft). https://web.stanford.edu/~jurafsky/slp3/

Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B. J., Jacobsen, L. M., Schatz, D. A., & Lernmark, Å. (2017). Type 1 diabetes mellitus. *Nature Reviews Disease Primers*, *3*(1), 17016. https://doi.org/10.1038/nrdp.2017.16

Keita, Z. (2022). *Classification in Machine Learning: A Guide for Beginners*. https://www.datacamp.com/blog/classification-machine-learning

Leong, K. S., & Wilding, J. P. (1999). Obesity and diabetes. *Best Practice & Research Clinical Endocrinology & Metabolism*, *13*(2), 221–237. https://doi.org/10.1053/beem.1999.0017

Loaiza, S. (2020, March 23). *Gini Impurity Measure*. Medium. https://towardsdatascience.com/gini-impurity-measure-dbd3878ead33

Mayer-Davis, E. J., Lawrence, J. M., Dabelea, D., Divers, J., Isom, S., Dolan, L., Imperatore, G., Linder, B., Marcovina, S., Pettitt, D. J., Pihoker, C., Saydah, S., & Wagenknecht, L. (2017). Incidence Trends of Type 1 and Type 2 Diabetes among Youths, 2002–2012. *New England Journal of Medicine*, *376*(15), 1419–1429. https://doi.org/10.1056/NEJMoa1610187

Microsoft. (2023). *Visual Studio Code* (1.78.0).

MSS-Suman. (2023). *Random Forest 97%—AUC 99.5% on balanced data*. https://kaggle.com/code/msssuman/random-forest-97-auc-99-5-on-balanced-data

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

Mustafa, M. (2023b, April). *Diabetes prediction dataset—Age Discussion*. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

NIDDK. (n.d.). *What is Diabetes?* National Institute of Diabetes and Digestive and Kidney

Diseases. Retrieved 22 May 2023, from https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

Nyuytiymbiy, K. (2020, December 30). *Parameters and Hyperparameters in Machine Learning and Deep Learning*. Medium. https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pociot, F., & Lernmark, Å. (2016). Genetic risk factors for type 1 diabetes. *The Lancet*, *387*(10035), 2331–2339. https://doi.org/10.1016/S0140-6736(16)30582-7

Polsky, S., & Ellis, S. L. (2015). Obesity, insulin resistance, and type 1 diabetes mellitus. *Current Opinion in Endocrinology, Diabetes and Obesity*, *22*(4). https://journals.lww.com/co-endocrinology/Fulltext/2015/08000/Obesity,_insulin_resistance,_and_type_1_diabetes.5.aspx

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ray, S. (2019). *A quick review of machine learning algorithms*. 35–39.

Reece, E. A., Leguizamón, G., & Wiznitzer, A. (2009). Gestational diabetes: The need for a common ground. *The Lancet*, *373*(9677), 1789–1797. https://doi.org/10.1016/S0140-6736(09)60515-8

Rouse, M. (2023, May 4). Hyperparameter. *Techopedia*. https://www.techopedia.com/definition/34625/hyperparameter-ml-hyperparameter

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., Shaw, J. E., Bright, D., & Williams, R.

(2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice*, *157*, 107843. https://doi.org/10.1016/j.diabres.2019.107843

Seery, C. (2023, January 5). Normal blood sugar ranges and blood sugar ranges for adults and children with type 1 diabetes, type 2 diabetes and blood sugar ranges to determine people with diabetes. *Diabetes*. https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

Shah, T. (2020, July 10). *About Train, Validation and Test Sets in Machine Learning*. Medium. https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Srinivasan, A. V. (2019, September 7). *Stochastic Gradient Descent—Clearly Explained !!* Medium. https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31

Taskesen, E. (2020). *HGBoost* (1.1.0).

Taskesen, E. (2022, October 24). *A Guide to Find the Best Boosting Model using Bayesian Hyperparameter Tuning but without…*. Medium. https://towardsdatascience.com/a-guide-to-find-the-best-boosting-model-using-bayesian-hyperparameter-tuning-but-without-c98b6a1ecac8

Type 2 Diabetes. (2023, April 18). Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/type2.html

Van Rossum, G., & Drake, F. L. (2022). *The Python Language Reference—Python 3.10.4 documentation* (3.10.4). https://docs.python.org/release/3.10.4/reference/index.html

Wang, W. (2022, March 22). *Bayesian Optimization Concept Explained in Layman Terms*. Medium. https://towardsdatascience.com/bayesian-optimization-concept-explained-in-layman-terms-1d2bcdeaf12f

WHO. (2023, April 5). *Diabetes*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

Zimmet, P. Z., Magliano, D. J., Herman, W. H., & Shaw, J. E. (2014). Diabetes: A 21st century challenge. *The Lancet Diabetes & Endocrinology*, *2*(1), 56–64. https://doi.org/10.1016/S2213-8587(13)70112-8