

Project 1: Supervised Learning - Classification

Alexander Svarfdal Gudmundsson

Jan Babin

October 2, 2024

1 Introduction

1.1 Background

Diabetes is a chronic condition characterised by elevated blood sugar levels. While genetic factors can play a role, a significant amount of diabetes Type 2 cases is associated with lifestyle choices. According to the World Health Organization [WHO2016], 422 million people worldwide are living with diabetes, with numbers continuing to rise. The societal impact of diabetes is significant: a diminished quality of life and a higher risk of serious health complications which in turn may lead to increased healthcare costs. Therefore, there is medical merit in early predictions of diabetes to refine treatment and management strategies.

1.2 Objective

The goal of this project is to build a machine learning model that can identify patterns in lifestyle and demographic data associated with diabetes and thereby predict whether a given person has diabetes / is very likely to develop it. While the model is purely intended for academic purpose, the project allows us to explore how machine learning can be used to analyse health-related data and provide insights into potential risk factors.

1.3 Dataset

The dataset at hand, "Diabetes, Hypertension and Stroke Prediction," is from Kaggle and contains health-related data in CSV format intended for predicting diabetes, hypertension and stroke. For this project, we are solely focussing on diabetes. The dataset consists of 70,692 observations and 18 features, some of which are outlined as follows (for a full overview, see Table 3 in the Appendix):

- **Age:** Coded in 13 age groups (see Figure 1).
- **Sex:** Binary variable representing male (0.0) and female (1.0).
- **BMI:** Body Mass Index, a continuous variable.
- **Lifestyle indicators:** Such as smoking status (whether the individual has smoked at least 100 cigarettes in their lifetime), physical activity in the past 30 days (excluding job-related activity), daily fruit and vegetable consumption, and heavy alcohol consumption (based on defined weekly limits for men and women).

- **General health and mental/physical health:** Self-reported general health on a scale from 1 (excellent) to 5 (poor), along with the number of days with poor mental and physical health over the past 30 days.
- **Pre-existing conditions:** Information on high cholesterol, coronary heart disease or myocardial infarction, and difficulty walking.

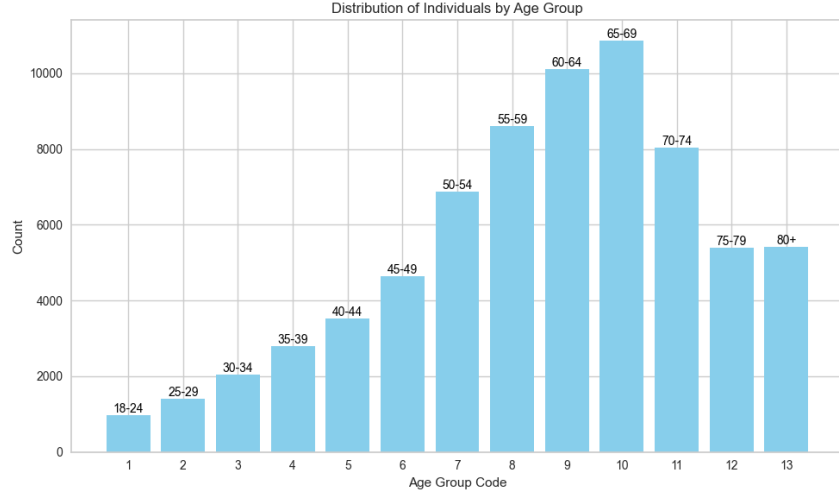


Figure 1: Distribution of Age Groups in the Dataset: Counts of individuals across thirteen defined age groups, ranging from 18 to 24 years to 80 years and older, according to the AGE5YR scheme.

The target variable is **Diabetes**, which is represented as a binary category (0.0 for no diabetes, 1.0 for diabetes). No missing or null values were identified in the dataset, ensuring that data cleaning was minimal and straightforward. Furthermore, the dataset is perfectly balanced, in that, 50% of observations correspond to diabetes, the other 50% do not.

2 Process

The steps involved in the analysis follow a standard supervised machine learning pipeline, from data preparation to model selection and evaluation.

2.1 Data Loading and Exploration

The dataset was loaded into a pandas DataFrame. Initial exploration was conducted to understand the structure of the data:

- **Data Integrity:** We confirmed that the dataset contained no missing or null values.
- **Feature Exploration:** The features were analysed for their unique values and data types, revealing that many variables were binary, and the rest were either categorical or continuous. Continuous variables included features like BMI, Age, MentHlth, and PhysHlth, while categorical variables included GenHlth (self-reported health scale from 1 to 5) and binary indicators for lifestyle factors like Smoking, Physical Activity, and Alcohol Consumption.

2.2 Preprocessing

2.2.1 One-Hot Encoding

WE SHOULD DISCUSS THIS AGAIN

Since some categorical features (e.g., Sex, GenHlth) were non-numeric, one-hot encoding was applied to convert these categories into numeric values to give the distance based models a chance too:

- Sex: Encoded as a binary variable with one-hot encoding and `drop_first=True` to prevent perfect multicollinearity (i.e., avoiding redundancy with one binary variable sufficing to represent gender).
- GenHlth: This feature, originally a categorical variable on a scale from 1 to 5, was one-hot encoded to allow for better model interpretation and training.

2.2.2 Scaling

(SCALING SEEMS TO BE DEACTIVATED IN THE NOTEBOOK? SHOULD WE NOTE THAT IT DIDNT MAKE A DIFFERENCE SINCE PREPROCESSING TAKES CARE OF IT ALSO) Certain features with a wide range of values were standardized using `StandardScaler`. These included: BMI, MentHlth, PhysHlth, and Age: Standardization ensures that the models (e.g., logistic regression, support vector machines) are not biased towards features with larger numeric ranges. This is important for algorithms that rely on the magnitude of features during decision-making.

2.2.3 Train-Test Split

The dataset was split into a training set (80%) and a test set (20%) using `train_test_split` from `sklearn`. We decided on this split for our relatively large data set in order to retrieve both substantial training and test set sizes. This reduces overfitting (model has enough training to learn nuances without memorising specific instances) as well as yielding a robust set of unseen data for unseen data and is a common industry practice. Stratification was applied to ensure that the class distribution of the target variable (Diabetes) remained balanced across the training and test sets.

2.3 Feature Correlation Analysis

To explore potential multicollinearity among features, a correlation matrix was generated using Seaborn's heatmap function. The goal was to identify highly correlated features that could negatively impact model performance by redundancy.

Findings: The GenHlth and PhysHlth features exhibited some degree of linear correlation but it was not high enough to justify omitting one of two would outweigh having more features / data to train on. WE MIGHT WANNA REFERENC SOMETHING REGARDING CORRELATION VALUES HERE (when to throw away a feature / when not)

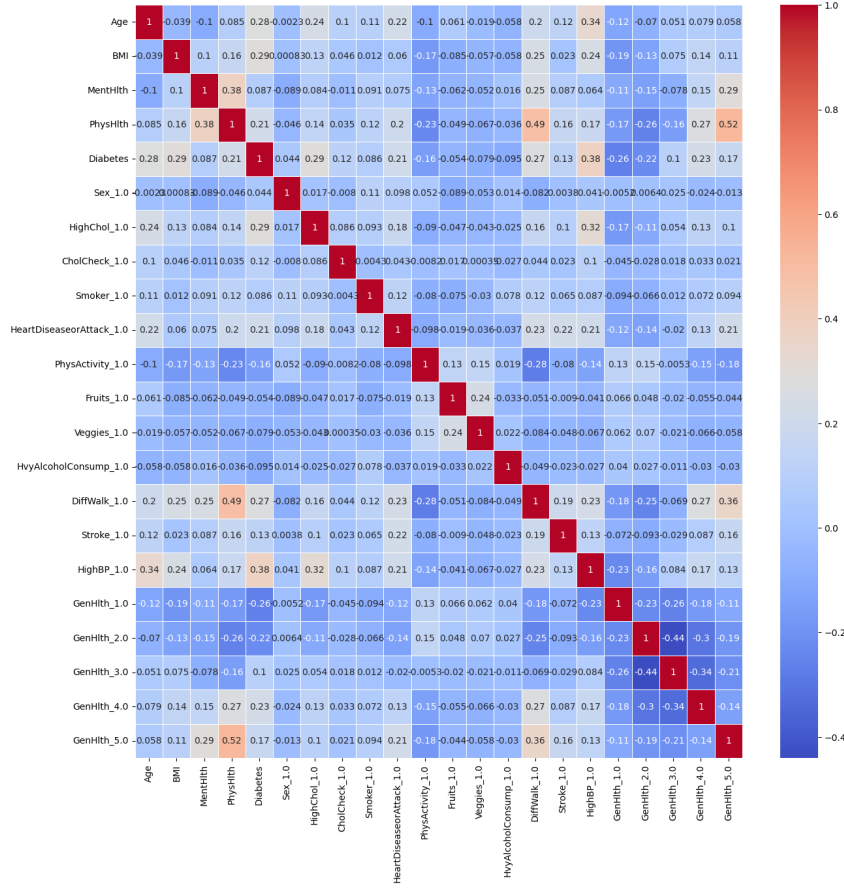


Figure 2: Correlation matrix showing relationships between features in the dataset (Pearson correlation). Higher values indicate stronger correlations.

2.4 Model Selection and Evaluation

2.4.1 PyCaret Experiment Setup

For model selection, we utilised PyCaret’s ClassificationExperiment to quickly get a comparison between multiple classification models. We then selected the 5 best performing models based on default hyperparameters to fine-tune further via hyperparameter optimisation.

PyCaret’s built-in random grid-search was used to tune the hyperparameters of the selected models, optimising for accuracy.

2.4.2 Model Stacking and Ensembling

Ensemble Methods: In an effort to improve prediction accuracy, ensemble learning methods were applied. A hard voting ensemble (WHAT DOES THIS EXACTLY DO?) (blend_models) was created using the top-performing models. Stacking: A stacking model was also built, However, the stacking model did not outperform the ensemble.

2.5 Model Stacking and Ensembling

After experimenting with feature selection (also implemented via PyCaret), we found that applying feature selection resulted in slightly lower model performance. Therefore, the

final models were built without this step.

3 Results

By using PyCaret’s built-in `ClassificationExperiment`, we can quickly evaluate which classifiers perform best on our test (???) dataset.

Model	Classifier	Accuracy	Recall	Precision	F1
<code>gbc</code>	Gradient Boosting Classifier	0.7508	0.7897	0.7329	0.7601
<code>lightgbm</code>	Light Gradient Boosting Machine	0.7495	0.7926	0.7298	0.7598
<code>ada</code>	Ada Boost Classifier	0.7482	0.7731	0.7366	0.7543
<code>lr</code>	Logistic Regression	0.7467	0.7737	0.7342	0.7534
<code>ridge</code>	Ridge Classifier	0.7467	0.7799	0.7314	0.7549
<code>lda</code>	Linear Discriminant Analysis	0.7467	0.7798	0.7314	0.7548
<code>rf</code>	Random Forest Classifier	0.7273	0.7627	0.7124	0.7366
<code>nb</code>	Naive Bayes	0.7246	0.7251	0.7244	0.7247
<code>et</code>	Extra Trees Classifier	0.7108	0.7330	0.7019	0.7171
<code>knn</code>	K Neighbors Classifier	0.7019	0.7196	0.6951	0.7071

Table 1: Performance of Top 10 Models (Accuracy, Recall, Precision, F1) with Highlighted Values

3.1 Summary of Best Model Performance

The table below shows the best results for the top 5 models (sorted by accuracy) for the key metrics: Accuracy, Recall, Precision, and F1. The best value for each metric is highlighted in yellow. WE NEED TO CHECK IF THIS IS RIGHT

Model	Classifier	Accuracy	Recall	Precision	F1
<code>lightgbm</code>	Light Gradient Boosting Machine	0.7515	0.7930	0.7336	0.7603
<code>gbc</code>	Gradient Boosting Classifier	0.7508	0.7897	0.7329	0.7601
<code>lr</code>	Logistic Regression	0.7499	0.7737	0.7342	0.7534
<code>ada</code>	Ada Boost Classifier	0.7482	0.7731	0.7366	0.7543
<code>ridge</code>	Ridge Classifier	0.7467	0.7799	0.7314	0.7549

Table 2: Best Results of Top 5 Models (Sorted by Accuracy, with highlights for best metrics)

- 3.2 Ensemble (stacking and blending)
- 3.3 Testing the models with the test data
- 3.4 Overfitting/Underfitting
- 3.5 Addressing over and underfitting?
- 4 Future Work

Placeholder

Appendix

Feature	Description
Age	Coded in 13 age groups (e.g., 1: 18-24, 2: 25-29, etc.)
Sex	Sex of the individual (0: Male, 1: Female)
HighChol	High cholesterol (0: No, 1: Yes)
CholCheck	Checked cholesterol in the last 5 years (0: No, 1: Yes)
BMI	Body Mass Index (continuous variable)
Smoker	Smoked at least 100 cigarettes in their lifetime (0: No, 1: Yes)
HeartDiseaseorAttack	History of coronary heart disease or myocardial infarction (0: No, 1: Yes)
PhysActivity	Engaged in physical activity in the past 30 days, excluding work (0: No, 1: Yes)
Fruits	Consumes fruit 1 or more times per day (0: No, 1: Yes)
Veggies	Consumes vegetables 1 or more times per day (0: No, 1: Yes)
HvyAlcoholConsump	Heavy alcohol consumption (men: 14+ drinks/week, women: 7+ drinks/week) (0: No, 1: Yes)
GenHlth	Self-reported general health (1: Excellent, 2: Very good, 3: Good, 4: Fair, 5: Poor)
MentHlth	Days of poor mental health in the past 30 days (0 to 30)
PhysHlth	Days of poor physical health in the past 30 days (0 to 30)
DiffWalk	Difficulty walking or climbing stairs (0: No, 1: Yes)
Stroke	History of stroke (0: No, 1: Yes)
HighBP	High blood pressure (0: No, 1: Yes)
Diabetes	Presence of diabetes (0: No, 1: Yes)

Table 3: Full list of dataset features used in the analysis.