Project Proposal for “Comparison of different Machine Learning Techniques for Type 2 Diabetes Detection”

Student 1: Rafael Almeida Albuquerque (40230226)
Student 2: Bobae Jeon (40239836)
Student 3: Kim-Carolin Lindner (40256554)
Student 4: Tung Vu (40223712)
Date submitted: October 19, 2022

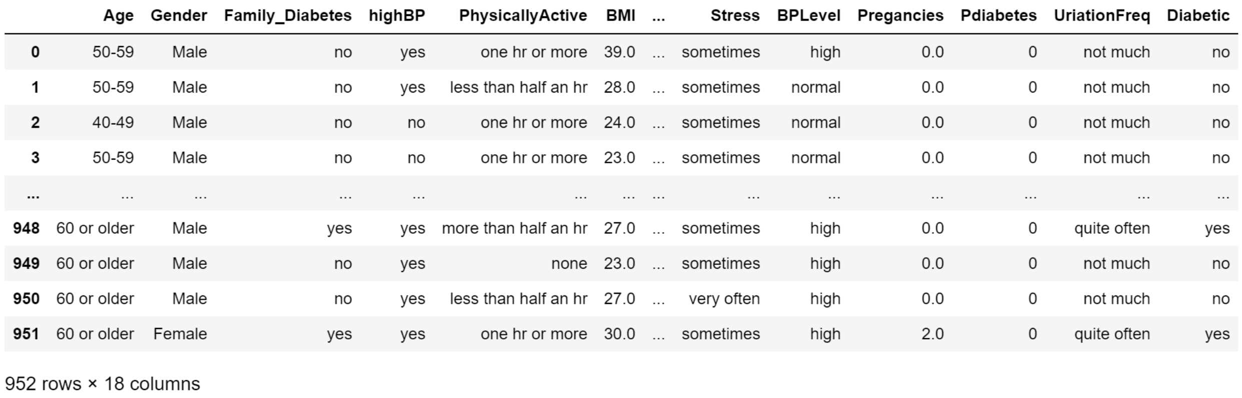
Goal of the project. More than 500 million people worldwide suffer from type 2 diabetes and even more live at an increased risk of developing the disease. Type 2 diabetes is the most common type of diabetes and can lead to heart diseases, strokes and other severe conditions [2]. Therefore, an early diagnosis and risk assessment is key for the prevention of serious damage. Nevertheless, people with this type of diabetes often live many years without being diagnosed although there exist specific factors that could indicate a high risk of a patient [1]. Machine learning techniques can help to predict type 2 diabetes based on given risk factors and hence lead to earlier and better treatment.

Following, the aim of this work is to investigate the performance of different machine learning models for the prediction of type 2 diabetes based on given risk factors. Once we have found the best model for the given task by comparing the models’ performances against each other, we will show how fine-tuning can help increase the performance even more. Since missing a diagnosis of type 2 diabetes in early stages can have life-threatening consequences, we want to take a closer look at false negatives in our models and minimize them as a secondary goal.

Data we plan to use. To achieve our goal, we will use the Diabetes-Dataset-2019 dataset [3]. This tabular dataset was collected from online and offline surveys by Tigga and Garg. It was used in their evaluation of machine learning models [4]. The dataset contains 952 observations with 17 attributes, where each attribute is either numerical or categorical. Our target variable is “diabetic” with two possible values, namely “no” and “yes”. Hence, our task will be to perform binary classification from given features. A detailed example of our dataset is provided in Figure 1.

For preprocessing, we will first need to perform an exploratory data analysis. We will clean our dataset from missing or false values and duplicates. Since our dataset is remarkably imbalanced (72% “no”, 28% “yes”), we will execute sampling techniques to achieve balanced data. After that, we will apply feature engineering and either convert categorical string values to integers or perform one-hot encoding as well as normalize numerical values. Moreover, we aim to create new features in order to increase the performance of our models. Of course, feature correlation will also be investigated in this step.

Once preprocessing is done, we will split the samples into 85% training and 15% testing. Since our dataset is not big enough, we chose this split to maximize the training data while preserving the test data to some extent. Particularly, we will perform the split with a specific random seed to keep the result constant. It will be easily trained on a laptop since the dataset is not extremely big and we aim to apply basic machine learning models.



	Age	Gender	Family_Diabetes	highBP	PhysicallyActive	BMI	...	Stress	BPLevel	Pregancies	Pdiabetes	UriationFreq	Diabetic
0	50-59	Male	no	yes	one hr or more	39.0	...	sometimes	high	0.0	0	not much	no
1	50-59	Male	no	yes	less than half an hr	28.0	...	sometimes	normal	0.0	0	not much	no
2	40-49	Male	no	no	one hr or more	24.0	...	sometimes	normal	0.0	0	not much	no
3	50-59	Male	no	no	one hr or more	23.0	...	sometimes	normal	0.0	0	not much	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...
948	60 or older	Male	yes	yes	more than half an hr	27.0	...	sometimes	high	0.0	0	quite often	yes
949	60 or older	Male	no	yes	none	23.0	...	sometimes	high	0.0	0	not much	no
950	60 or older	Male	no	yes	less than half an hr	27.0	...	very often	high	0.0	0	not much	no
951	60 or older	Female	yes	yes	one hr or more	30.0	...	sometimes	high	2.0	0	quite often	yes

952 rows × 18 columns

Figure 1. Examples of our dataset. Each row will be randomly assigned to training set or test set

How we measure "success." The measurement of our models' performance will be done in two steps, since we must first compare all techniques implemented against each other, and later evaluate the final fine-tuned mode. For both parts, our measures will be based on comparisons of standard metrics (accuracy, precision, recall, F1 score) alongside the AUC (Area Under the ROC Curve), the latter being selected because it makes it easier to visualize the rate at which the model is correctly classifying each instance into its respective class.

A ten fold cross-validation will be used to make sure the models have a lesser variation between themselves and are not biased by the selection of our validation set. After that, the optimal model will be selected and the process of fine-tuning will be done with a grid-search optimization to increase the performance even further. Finally, our final result will be compared with the models developed in [4] as it was done using the same dataset and would make a great point of comparison. We do not expect to surpass this model but having values to compare it to will give some perspective in the final evaluation of our model.

Moreover, it is important to point out that since our focus is mainly on the metrics presented above, we will not be taking into account training/prediction time since this model would not be used in a real-time system.

Division of work. Kim will perform exploratory data analysis and preprocessing, as well as train a LogisticRegression model. Tung will train a Neural Networks model and SVMs. Rafael will train the RandomForestClassifier as well as a DecisionTreeClassifier and work on our final visualization of results. Bobae will train a XGBoost model and perform fine-tuning for our best model. All team members will evaluate and analyze the models and write the project report.

Python packages we expect to use. Pytorch, scikit learn, pandas, numpy, matplotlib, seaborn.

References:

[1] Centers for Disease Control and Prevention. Type 2 Diabetes. https://www.cdc.gov/diabetes/basics/type2.html. Last accessed: 2022-10-19.

[2] IDF Diabetes Atlas. https://diabetesatlas.org/. Last accessed: 2022-10-19.

[3] Open ML. Diabetes-Dataset-2019. https://www.openml.org/search?type=data&status=active&id=43341&sort=runs. Last accessed: 2022-10-19.

[4] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, *167*, 706-716.