**Diabetes prediction**

**Saška Topalović SW56/2019**

1. Motivation

Diabetes is a chronic condition characterized by high blood sugar levels, which, if left unmanaged, can lead to severe health complications such as cardiovascular disease, kidney problems, nerve damage, and vision loss. In recent years, there has been a significant increase in the number of people diagnosed with diabetes worldwide. Sedentary lifestyles, unhealthy dietary habits, and a lack of awareness about the risk factors contribute to this growing epidemic. The prediction and prevention of diabetes has never been more important and that is why I chose this topic.

2. Research questions

Prediction of diabetes based on a dataset [1] that includes features: age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level.

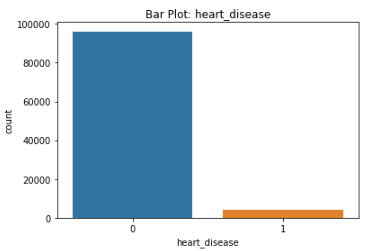
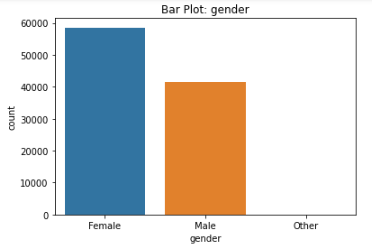
The dataset contains 9 columns and 100 000 rows. It also includes 4 numerical (age, bmi, HbA1c\_level, blood\_glucose\_level) and 5 categorical features (gender, hypertension, heart\_disease, smoking\_history, diabetes). The dataset is unbalanced.

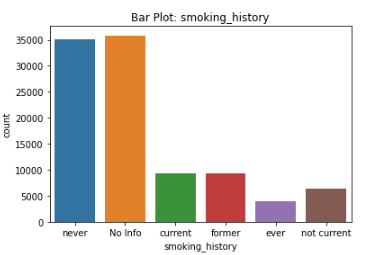
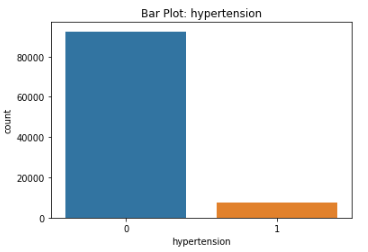
3. Related work

Since I couldn't find papers that specifically cover kNN, SVM, and logistic regression while using the micro f1 performance measure, in this section, I will discuss a paper [2] that I found which utilizes multiple classification models, including kNN, SVM, and logistic regression. The paper employed label encoding for categorical feature encoding. Subsequently, the multicollinearity between features was checked. The f1 score was used as the performance measure. The results obtained using kNN, SVM, and logistic regression will be of interest. The dataset was divided into training (80%) and test (20%) sets. StandardScaler was used for normalization on both the training and test datasets. The f1 performance measure obtained for the kNN classifier was 0.94, for logistic regression it was 0.94, and for the SVM classifier it was also 0.94.

4. Methodology

After loading the mentioned dataset, I created barplots for all categorical features.





From the barplots, I noticed that the 'hypertension' and 'heart disease' features had a significantly higher number of '0' values, so they were removed for that reason. Additionally, all rows where the value of the categorical feature 'gender' was equal to 'other' were removed, as there were only 18 instances of such cases.

I checked for outliers. Firstly, I verified if there were any negative values for the 'age' feature, but that was not the case. Next, I examined the 'blood\_glucose\_level' feature. Since a blood glucose level greater than 126 indicates diabetes, the idea was to remove all rows where the 'blood\_glucose\_level' exceeded 126 and the corresponding 'diabetes' column value was 0. However, this would result in the exclusion of a significant number of rows, which did not yield good results after classification.

As the dataset was unbalanced, I also tried under-sampling, over-sampling, and smoothing techniques, but they did not provide satisfactory results.

Then, I performed encoding on the remaining categorical features ('gender' and 'smoking\_history') using label and one-hot encoding (using encode\_categorical\_with\_label\_encoding(data) and encode\_categorical\_with\_one\_hot\_encoding(data)).

The final preprocessing step was the normalization of numerical features ('age', 'bmi', 'HbA1c\_level', 'blood\_glucose\_level') using StandardScaler on both the training and test datasets.

5. Discussion

The dataset was split into training (80%) and test (20%) sets using the train\_test\_split() method from the sklearn library.

For classification, SVM classifier, kNN classifier, and logistic regression were used.

As a performance measure for the given problem of predicting diabetes due to the unbalanced dataset, the micro f1 score was used. The micro f1 score takes into account the number of correctly classified instances for each class, meaning both classes (having diabetes - 1 and not having diabetes - 0) are treated equally important, which is crucial for diabetes prediction regardless of the class imbalance. Additionally, the micro f1 score evaluates both false positives and false negatives and assesses the overall accuracy of the model.

The results obtained before hyperparameter optimization were as follows:

I used RandomizedSearchCV for hyperparameter optimization. The following hyperparameter values yield the best results for the classifier:

* SVM: C=10, gamma=1, kernel='rbf'
* kNN: algorithm='kd\_tree', n\_neighbors=6, weights='uniform'
* Logistic regression: C=0.6465976998829948, penalty='l2', solver='saga'

After hyperparameter optimization for all three classifiers, the following values were obtained for the micro f1 score: **SVM:** **0.9684**, **kNN:** **0.9607** and **logistic regression:** **0.9592**. How label encoding performed better than one-hot encoding, all the presented results were obtained using label encoding for encoding categorical features.

In conclusion, for the problem of predicting diabetes, the SVM classifier performed the best.

6. References

[1] <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>

[2] <https://www.kaggle.com/code/moumoni/diabetes-prediction-2>