Machine Learning
Exercise 3 - Kaggle competition 2/2
Data Set: Stroke

Student: se21m024, Thomas Stummer, matriculation number: 1425616

Introduction

This document describes the process to participate at the Kaggle competition for the data set 'stroke' and summarizes the results.

Related Files and Links

- Locally developed and tested notebook:***se21m024_Stummer_ml_ex3_stroke.ipynb***
- Kaggle submission description:***se21m024_Stummer_ml_ex3_stroke_kaggle.pdf***
- Online Kaggle notebook: ***https://www.kaggle.com/code/se21m024/notebook-se21m024-stroke***
- Submitted Kaggle prediction: ***stroke_prediction_2022-06-05_08-26-39_utc.csv***

Kaggle Submission

The notebook *se21m024_Stummer_ml_ex3_stroke.ipynb* was locally developed and tested. The notebook was than imported in Kaggle. The only change in the Kaggle notebook after the import was setting the flag use_kaggle_paths to 'True' to adjust the path for the input data files. Then the notebook was run online and the solution file was submitted directly via the Kaggle web GUI. The solution file *stroke_prediction_2022-06-05_08-26-39_utc.csv* was then downloaded and added to the Moodle upload. The document *se21m024_Stummer_ml_ex3_stroke_kaggle.pdf* contains exactly the information that was provided in the description field of the Kaggle submission.

Data Set Analysis

The data set contains different parameters about persons including the information whether a person suffered a stroke or not. Some features like hypertension and heart_disease have a significant shift towards one of the binary possibilities (factor of 100 to 200). Other features are more equally represented in the data set like the gender split which is slightly shifted towards females (59% female, 41% male) or the residence type (50% urban and 50% rural).

Utilized Software

Operating System:
Windows 10

Software Version:
Python 3.8.5

sklearn Version:
0.23.2

Data Preparation and Algorithms

Processing applied to the data before training / predicting:
The non-numeric features in the data sets have been mapped to numeric values before further processing (nan values were filled with '0'). The learning data was split into the input features (everything except of the 'ID' and 'stroke' column) and the target feature (the 'stroke' column).

Then the learning data was split into 2/3 training set and 1/3 test set. With this split 4 different algorithms with different parameters were trained:

- k-NN (1-NN)
- k-NN (2-NN)
- k-NN (3-NN)
- Decision Tree (max features: None)
- Decision Tree (max features: sqrt)
- Decision Tree (max features: log2)
- SVM (SVC classifier)
- Random Forests (num trees: 10, max features: sqrt)
- Random Forests (num trees: 10, max features: log2)
- Random Forests (num trees: 100, max features: sqrt)
- Random Forests (num trees: 100, max features: log2)

The best performing algorithm on this test split was than retrained with the whole learning data and used to create the prediction on the test set. The exact results for all algorithms can be seen in the notebook itself.

Results

Local Results

The best performing setting was the k-NN algorithm with k=2, which accomplished an accuracy of 0.954 and an F1 measure of 0.936 on the learning set.

Kaggle Results

The results on the learning set achieved by the local notebook was exactly the same when executing the notebook online in Kaggle (accuracy of 0.954 and an F1 measure of 0.936). The submitted solution for the Kaggle test set achieved a score of only 0.11594 (public leaderboard) which is surprisingly very significantly lower than the accuracy on the learning set.