Applied MSc in Data Analytics

Applied MSc in Data Science & Artificial Intelligence

Applied MSc in Data Engineering & Artificial Intelligence

Project: Stroke Prediction Model

Instructor: Hanna Abi Akl

Project Summary:

According to the World Health Organisation (WHO), **stroke** is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

The dataset provided is used to predict whether a patient is likely to get a stroke based on input parameters like gender, age, various diseases and smoking status.

Below is the information you have regarding the dataset attributes:

1) **id:** unique patient identifier
2) **gender:** "Male", "Female" or "Other"
3) **age:** age of the patient
4) **hypertension:** 0 (if the patient doesn't have hypertension) or 1 (if the patient has hypertension)
5) **heart_disease:** 0 (if the patient doesn't have a heart disease) or 1 (if the patient has a heart disease)
6) **ever_married:** "No" or "Yes"
7) **work_type:** "children", "Govt_job", "Never_worked", "Private" or "Self-employed"
8) **Residence_type:** "Rural" or "Urban"
9) **avg_glucose_level:** average glucose level in the blood
10) **bmi:** body mass index
11) **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown" (in this case the information for the patient is not available)
12) **stroke:** 1 (if the patient had a stroke) or 0 (if the patient didn't have a stroke)

Project Objectives:

Using the provided dataset, you are asked to train a model that predicts whether a patient has a stroke or not. The project can be submitted as a Jupyter Notebook and should include exploratory analysis of the data, feature engineering and selection, model training and evaluation.

You may use additional resources as you see fit (provided you can justify how they can serve your solution). You can even consult similar solutions from the Internet. **However, this comes with a big responsibility: any submission that is over-plagiarised or does not reflect personal work will not be accepted**.

Project Evaluation:

The project will be evaluated using the following rubric. It contains the required items for a complete submission as well as bonus elements. The grading system is over 5 and the final grade will be transformed to a grade over 100.

- Data analysis (data processing, data cleaning, exploratory analysis, plots of relevant attributes) **[1 point]**
- Feature selection (feature engineering, feature pruning, choice justification) **[1 point]**
- Model training (motivation for selected model, comparison of different models) **[1 point]**
- Model evaluation (evaluation metric, results interpretation) **[1 point]**
- Project report (short report explaining the approach and results) **[1 point]**
- **BONUS:** Project reproducibility (requirements file with necessary packages, README file for running the project) **[1/2 point]**
- **BONUS:** Project hosting (Github, Docker, AWS, Heroku or any other method) **[1/2 point]**

Project Timeline:

The deadline for the project is **60 days** from the project start date. Additionally, you are free to set a meeting with the instructor to discuss possible approaches, problems or other points pertaining to the project.