Week 4 Assignment - MSc AIDA UoM

Professor Nikolaos Samaras

Academic Year: 2020-2021

On Kaggle (`https://www.kaggle.com`), there is a dataset named ”Machine Learning for Diabetes with Python” (`https://www.kaggle.com/rahulsah06/machine-learning-for-diabetes-with-python/data`).

This dataset includes medical data (anonymized) for diabetes from 768 subjects. Approximately one in seven American adults has diabetes today, and by 2050 this ratio is expected to be one in three.

The following information pertains to the columns of the file `diabetes_data.csv`:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age
- Outcome (Class Variable)

Tasks:

1. Select 80% of the records as the training dataset randomly, ensuring the same ratio between the two classes as seen in the entire dataset. The remaining 20% will be used as the test dataset.
2. Use the KNeighborsClassifier algorithm to develop prediction models with $k = 5$ and $k = 7$ and use the following metrics for computing distance: Euclidean, Manhattan, and Chebyshev.
3. Perform a significance test between the models generated for $k = 5$ and $k = 7$.