Bishop's University

CS 405/CS 505 – Data Mining

Final project: Diabetes prediction

1. Introduction

Diabetes is one of the most common and hazardous diseases on the planet. It requires a lot of care and proper medication to keep the disease in control. In this data mining project, this project teaches you to develop a classification system to detect whether the patient has diabetes or not. As part of this project, you will practice the usage of the classifiers that we have seen during data in class to predict diabetes for patients of a given dataset. You have to implement the three classifiers in Python and to validate them using the provided diabetes dataset by following the four steps that are described in 2).

2. Tasks

Step 1. Choose three of the following classifiers to use them for the prediction of diabetes:

- Decision Trees
- Naïve Bayes
- Neural Networks
- K-Nearest neighbors
- Support Vector machine

The choice of three classifiers should be justified.

Step 2. Assess the chosen three classifier in Step 1). The principle is to compare between the three classifiers and to use the concept of model validation seen in the assignment 2. The overfitting problem should be taken into consideration during the experiments. A

section of experimental results must be added and detailed. You can support your arguments with curves and scores.

Step 3. Possibility to increase the scores. You have to check the possibility to increase the obtained scores in Step 2. Innovation in the experiment to improve the obtained scores will be an asset.

Step 4. A conclusion has to be added at the end. It should mention what was achieved in this final project, drawbacks, and the possibility of the improvement of the obtained results.

3. Submission

You must submit the *pdf* file of your report. In fact, it must contain detailed explanation of each step. In addition, you have to submit your *.py* or *.ipynb* if you prefer using *Jupyter Notebook*. ***Please, do not submit your final projet in .zip or .rar files.*** If you do a transformation to the dataset, you will have to mention it on your report and you will have to submit it.