Aprendizagem 2024
Homework I – Group 016
(ist1106022, ist1106720)

Part I: Pen and paper
Part II: Programming

1.) To compare the performance of a *kNN* classifier with $k = 5$ and a naive Bayes classifier, a 5-fold stratified cross-validation was performed on the on a heart disease dataset:

a.) For comparing accuracies of the two classifiers, a box plot was generated for each:

The performance of the *kNN* is less consistent as the box plot is wider, while the Naive Bayes classifier has a more consistent performance, as the box plot is narrower.

This is due to the fact that the *kNN* is non-parametric and therefore is more sensitive across folds, and with non-normalized data, different scaled features, redudant, irrelevant or noisy features can affect the performance of the classifier.

The Naive Bayes classifier, on the other hand, is parametric and is not sensitive to these factors.

b.) However, when choosing to scale the feature data with a Min-Max Scaler, the box plots look as follows:

This happens because the *kNN* classifier, when choosing to normalize feature data, it adresses the previously mentioned problems and leads to heavily reduced variability, which in turn leads to a more consistent and accurate classification.

For *Naive Bayes*, the scaling of the data does not generally affect the performance, since it is a parametric model.

c.) Perfoming a paired t-test on the accuracies of the two classifiers, using `scipy`'s method `ttest_rel` that considers H_0 to be the hypothesis that the two classifiers have the same stastical significance regarding accuracy, and H_1 to be the hypothesis that *kNN* is more accurate than Naive Bayes, the results are as follows:

- When not scaling feature data: P-value = 0.998415501126768
  - Considering a 1% threshold, *kNN* is not statistically superior than Bayes
  - Considering a 5% threshold, *kNN* is not statistically superior than Bayes
  - Considering a 10% threshold, *kNN* is not statistically superior than Bayes
- When min-max scaling feature data: P-value = 0.7532332545792753
  - Considering a 1% threshold, *kNN* is not statistically superior than Bayes
  - Considering a 5% threshold, *kNN* is not statistically superior than Bayes
  - Considering a 10% threshold, *kNN* is not statistically superior than Bayes

We can therefore easily conclude that the hypothesis "the *kNN* model is statistically superior to naïve Bayes regarding accuracy" is **false / rejected**.

2.) To compare the performance of uniform and distance-based weights *kNN* classifiers with varying amounts of neighbors (k) used in classifications, a 80-20 train-test split was performed for each combination:

a.) The following plots showcase the obtained results:

b.) Generally, the bigger the value of k on a *kNN* classifier, the more accurate the predictions are, until a certain point. This point, for either uniform or distance-based weights, is when the accuracies on both training and testing data is the highest, which seems to be around $k = 30$ for this dataset.

In terms of generalization capabilities, the distance-based weighting system seems to be more overfitted, with the training accuracy being unchanged and the testing accuracy flatlining. The uniform weighting system, however, seems to be more general, with the training and testing accuracy approaching each other as the number of neighbors increases.

Before the suggested point, for the uniform weights, the model is overfitting on the training data. Past $k = 30$ it may be at the risk of underfitting, a phenomenon not observable for the dataset at hand.

3.) Some properties from the dataset that may justify the shortcomings of Naive Bayes classifier on this dataset are:

- Since Naive Bayes assumes that all features are independent, it ignores the relantionships and correlation that these features may have in relation to heart disease diagnosis.
- The dataset has features which are not normally distributed or are not numerical, which is a violation of the Gaussian Naive Bayes conditions. Some of these features include the categorical features, such as cp, fbs, $resteg$, $exang$ and $slope$, for example. This leads to incorrect probability calculation, and therefore, incorrect classification.