Lab 2: Breast Cancer Prediction

1 Files included

dataR2.csv – csv data for training classifiers
dataR2.arff – arff data for training classifiers

2 Background

Breast cancer screening is an important strategy to allow for early detection and ensure a greater probability of having a good outcome in treatment. Robust predictive models based on data that may be collected in routine consultation and blood analysis are sought to provide an important contribution by offering more screening tools. In this lab the aim is to assess how machine learning techniques may be applied to data that can be collected in routine blood analyses - namely, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age, and Body Mass Index (BMI) - and be used to predict the presence of breast cancer. These parameters may be a good set of candidates, as it has been verified deregulation in their profile in obesity-associated breast cancer.

3 The assignment

The goal of this lab is to develop and assess prediction models which can potentially be used as a biomarker of breast cancer, based on anthropometric data and parameters which can be gathered in routine blood analysis. For each of the 116 participants, several clinical features were observed or measured, including age (years), BMI (kg/m2), Glucose (mg/dL), Insulin (µL/mL), HOMA, Leptin (ng/mL), Adiponectin (µg/mL), Resistin (ng/mL), and MCP-1 (pg/dL). These features should be used to predict the presence of breast cancer (1=healthy controls, 2=patients).

Your task is to apply machine learning algorithms (logistic regression, decision trees, artificial neural networks, k-NN) to the data to obtain the best possible model. You have to compare and report the accuracy, precision, recall, and F-measure (using 10-fold cross-validation) of the different algorithms. You should experiment with the parameters of the different algorithms to maximize their accuracy. Check if your models are overfitting. What is the accuracy of the baseline classifier (i.e. a classifier predicting the majority class)? You can present your results in a table of the form:

Algorithm	**C.C.I.%**	**Precision**	**Recall**	**F-measure**
Baseline classifier				
Log Reg				
k-NN (k=…)				
Decision Trees				
…				

Using the most promising algorithms, a **second task** is to explore different feature sets (obtained by applying filter and wrapper selection methods) with the aim to 1) determine which are the most relevant features for this problem, and 2) find out if feature selection improves or not the accuracy of the classifiers. As part of your exploration, you should plot a learning curve where the x-axis is the number of features in order of relevance (1 to 9), and the y-axis is the correctly classified instances percentage (C.C.I.%).

You may use the algorithms provided by the WEKA machine learning library (https://www.cs.waikato.ac.nz/ml/weka), Scikit-Learn (https://scikit-learn.org/) or any other machine learning library.

Submitting your answer

The lab may be solved in teams of two people (1 submission per team) or individually. Submission is through the Aula Global where you can find the submission deadline. Submissions consist of a PDF file containing an explanation of the procedure you followed, the code if you wrote some, and the plots and tables with your results. The lab will be evaluated taking into account the report and the best model accuracy.