UNIVERSIDADE DA CORUÑA

# First Machine Learning model to predict cancer proteins with Weka.

## Fundamentals in Bioinformatics.

Pedro Sánchez García

MASTER IN BIOINFORMATICS FOR HEALTH SCIENCES

TEACHER: DR. CRISTIAN ROBERT MUNTEANU

## PRACTICE 0: CLASSIFICATIONS

**You have a dataset with protein descriptors for proteins involved in two types of cancers: HCC and HBC. Using Practise/CancerDataSet.csv database and Weka, find the best classification model that can predict if a protein is related with HCC or HBC. You should use 10-fold cross validation, at least one method from *BayesNet, Decision Table*, and *Random Forest*. Order the performance of the models using TP Rate.**

### Random Forest

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,998    0,002    0,998      0,998    0,998      0,996    1,000     1,000     HBC
                 0,998    0,002    0,998      0,998    0,998      0,996    1,000     1,000     HCC
Weighted Avg.    0,998    0,002    0,998      0,998    0,998      0,996    1,000     1,000

=== Confusion Matrix ===

    a    b   <-- classified as
 1046    2 |    a = HBC
    2 1052 |    b = HCC
```

### Decision Table

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,948    0,491    0,657      0,948    0,777      0,509    0,789     0,764     HBC
                 0,509    0,052    0,908      0,509    0,652      0,509    0,789     0,827     HCC
Weighted Avg.    0,728    0,271    0,783      0,728    0,714      0,509    0,789     0,796

=== Confusion Matrix ===

   a   b   <-- classified as
 994  54 |   a = HBC
 518 536 |   b = HCC
```

### Naive Bayes

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,169    0,078    0,683      0,169    0,271      0,139    0,573     0,595     HBC
                 0,922    0,831    0,527      0,922    0,671      0,139    0,573     0,529     HCC
Weighted Avg.    0,547    0,456    0,605      0,547    0,472      0,139    0,573     0,562

=== Confusion Matrix ===

   a   b   <-- classified as
 177 871 |   a = HBC
  82 972 |   b = HCC
```

## DISCUSSION

Based on the results obtained, Random Forest is the best method, since if we look at it, the TP rate is very close to 1, which indicates a good prediction of the model for both the type of HCC and HBC cancer. In addition, unlike the Decision Table and Naive Bayes methods, the weighted average in Random Forest is high, in such a way that confirms that this is the best model for this problem treated in practice. This highlights its robustness, one of the characteristics for which it is widely applied in gene expression classification, mass spectrometry data analysis, prediction of interactions between proteins ...