SEMINAR 4

1. OBJECTIVES

- Work with input/output streams, read/write data from/to files.
- Work with inheritance and polymorphism.

2. PROBLEM STATEMENT

Implement a very simple machine learning library allowing classification, according to the given UML diagram (next page). The library will offer various algorithms for classification, allowing the user to choose any of them. All classifiers will provide training and testing capabilities and the library will also offer various evaluation measures to test the performance of the algorithms.

Test the algorithms using the *Pima Indians Diabetes dataset* (https://www.andreagrandi.it/2018/04/14/machine-learning-pima-indians-diabetes/) and report the obtained performances. Which is the best classifier?

Classifier

+train(instances: Instance[])
+predict(instances: Instance[]): int[]

Distance

+compute(v1: double[], v2: double[]): double
+copy(): Distance

-uses

KNearestNeighbours

-k: int

+train(instances: Instance[])
+predict(instances: Instance[]): int[]

RuleBasedClassifier

+train(instances: Instance[])
+predict(instances: Instance[]): int[]

NaiveBayes

+train(instances: Instance[])
+predict(instances: Instance[]): int[]

EuclideanDistance

+compute(v1: double[], v2: double[]): double
+copy(): EuclideanDistance

ManhattanDistance

+compute(v1: double[], v2: double[]): double
+copy(): ManhattanDistance

-1..*

Instance

-input: double[]
-output: int

+get_input(): double[]
+get_output(): int
-reset_instance()

EvaluationMeasure

+evaluate(instances: Instance[], predictions: int[]): double

Accuracy

MeanSquaredError