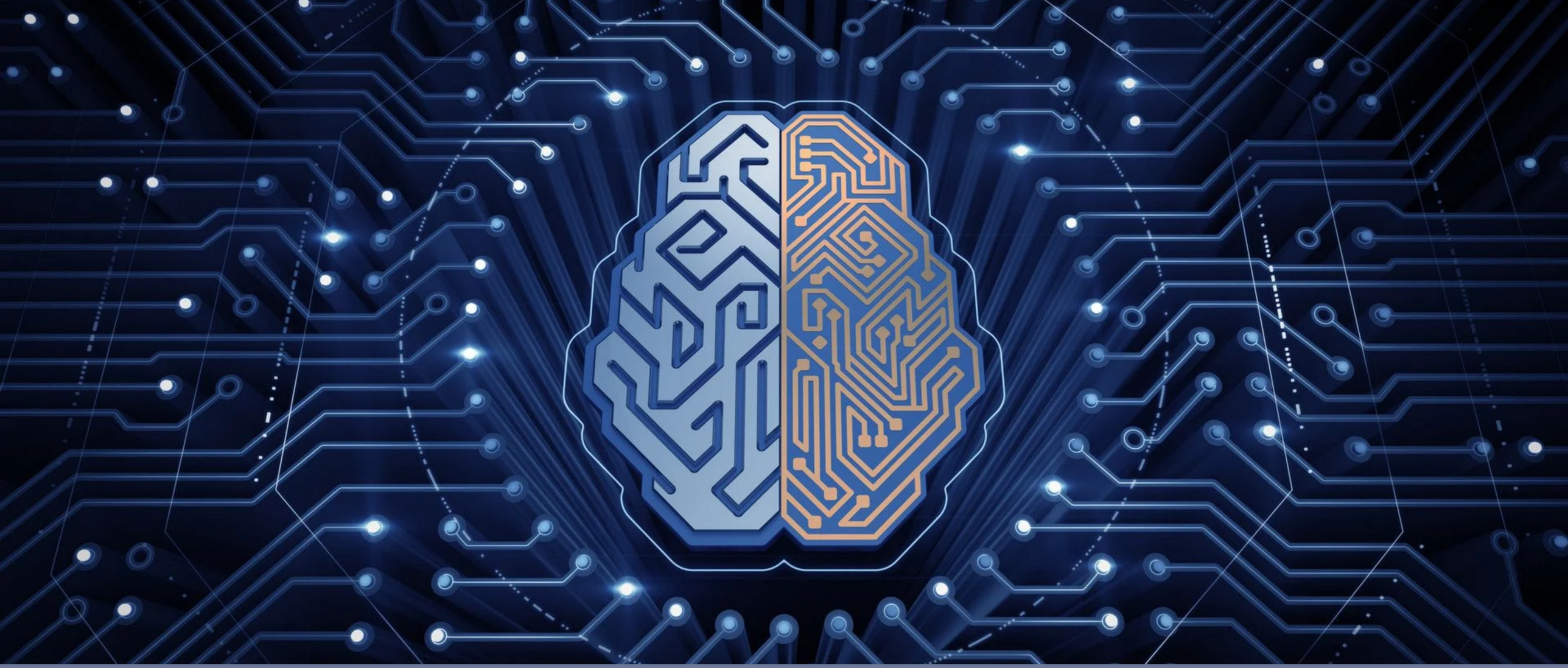


MACHINE LEARNING PROJECT

U.PORTO
FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

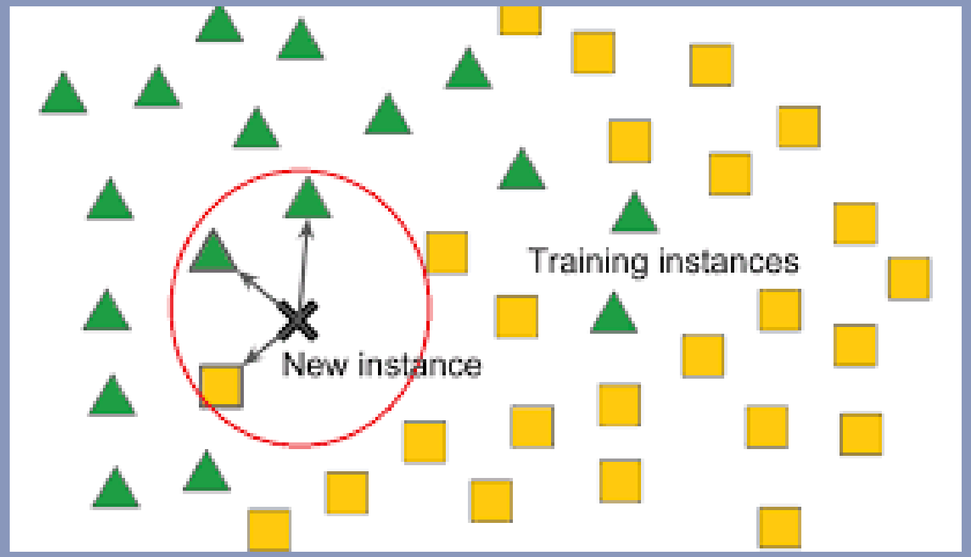
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

BEATRIZ SILVA - 202107955
CATARINA MONTEIRO - 202105279
LARA SOUSA - 202109782

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

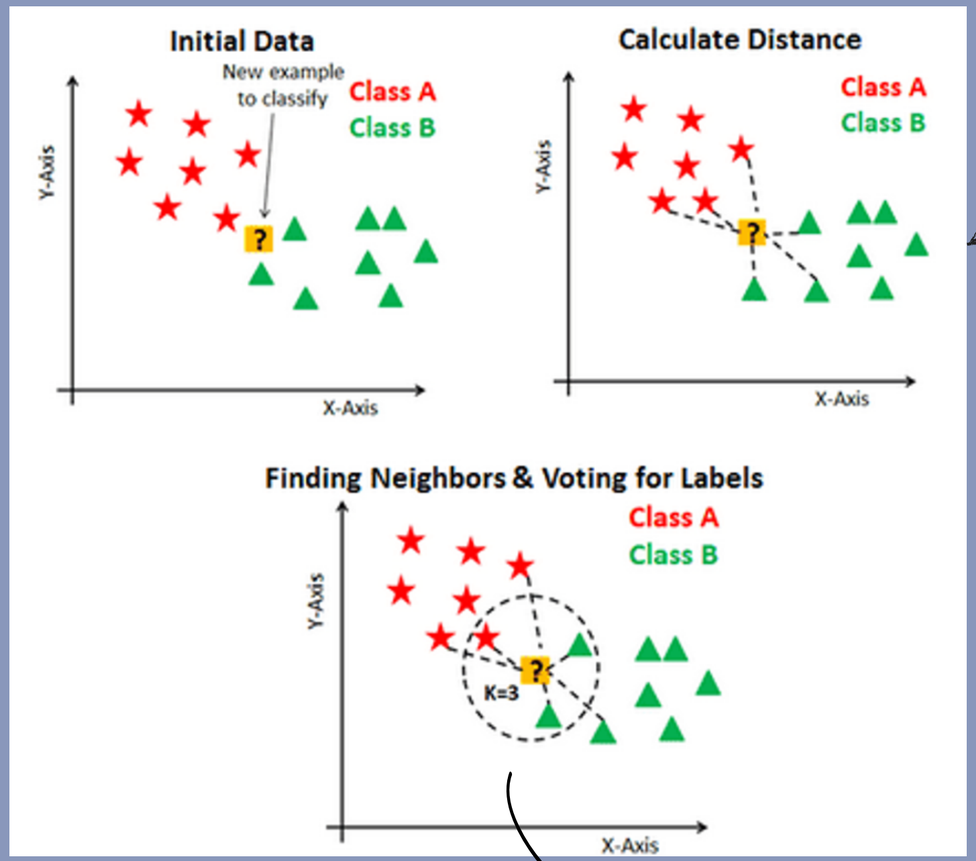
K-NEAREST NEIGHBOURS

CLASSIFICATION TASKS



BASED ON THE CLASSES OF THE K NEAREST NEIGHBOURS IN THE TRAINING SET, THE KNN ALGORITHM ASSIGNS A CLASS TO A NEW EXAMPLE.

THE DISTANCE BETWEEN EXAMPLES: CALCULATED USING A DISTANCE METRIC.



VALUE OF K: HOW MANY NEIGHBOURS ARE CONSIDERED IN THE CLASSIFICATION

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

K-NEAREST NEIGHBOURS

THE BASE CASE

ADVANTAGES

- SIMPLE AND INTUITIVE;
- NON-LINEAR DECISION BOUNDARIES;
- CAN HANDLE MULTI-CLASS CLASSIFICATION AND CAN ALSO MAKE PROBABILISTIC PREDICTIONS.

DISADVANTAGES

- COMPUTATIONALLY EXPENSIVE;
- REQUIRES STORING THE ENTIRE TRAINING DATASET IN MEMORY;
- SENSITIVE TO THE SCALE OF FEATURES;
- LACK OF INTERPRETABILITY.

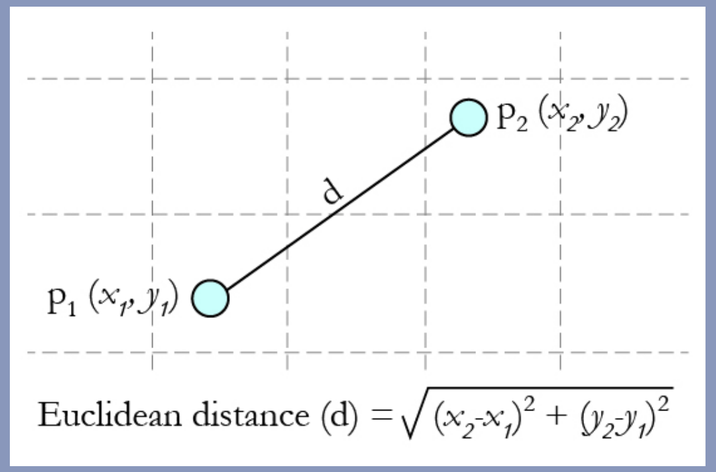
OUTLIERS

- SENSITIVE TO OUTLIERS BECAUSE IT CALCULATES DISTANCES BETWEEN DATA POINTS.

OVERFITTING

- WITH A SMALL K, THE ALGORITHM BECOMES SENSITIVE TO NOISE AND OUTLIERS IN THE TRAINING DATA, LEADING TO POOR GENERALIZATION ON UNSEEN DATA.

METRIC



SOLUTION IDEA

OUTLIERS

- USING DISTANCE METRICS THAT ARE LESS SENSITIVE TO OUTLIERS

OVERFITTING

- BY USING DIFFERENT SUBSETS OF FEATURES, EACH MODEL FOCUSES ON DIFFERENT ASPECTS OF THE DATA, REDUCING THE RISK OF OVERFITTING TO SPECIFIC FEATURES OR RELATIONSHIPS.

BAGGING

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

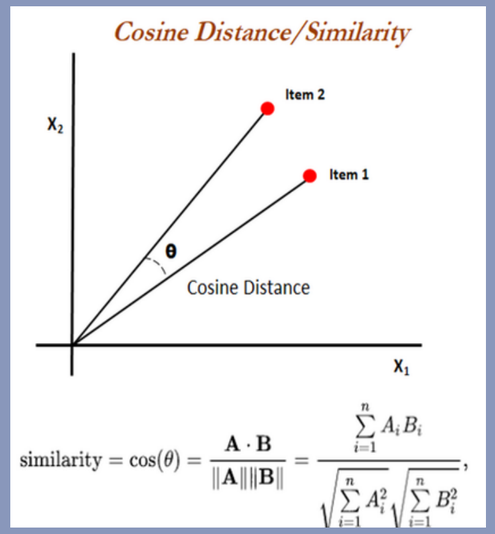
K-NEAREST NEIGHBOURS - BAGGING

OUTLIERS

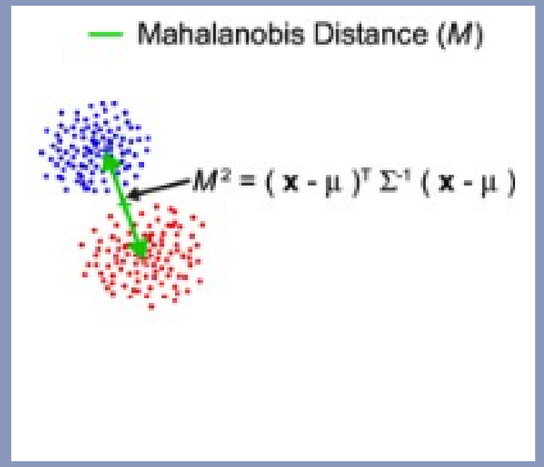
MITIGATE THE IMPACT OF OUTLIERS

EMPLOYING DIFFERENT DISTANCE METRICS: EACH MODEL IN THE ENSEMBLE WILL ASSIGN DIFFERENT WEIGHTS TO OUTLIERS BASED ON THEIR RESPECTIVE DISTANCE CALCULATIONS.
THIS DIVERSIFICATION CAN HELP REDUCE THE IMPACT OF OUTLIERS AND MAKE THE OVERALL BAGGED KNN MODEL MORE ROBUST.

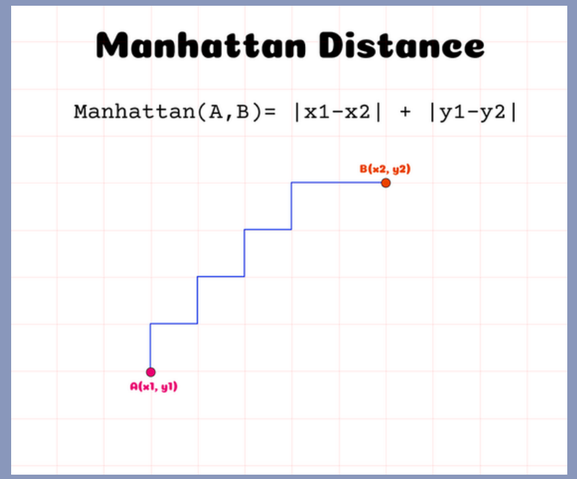
DISTANCE METRICS



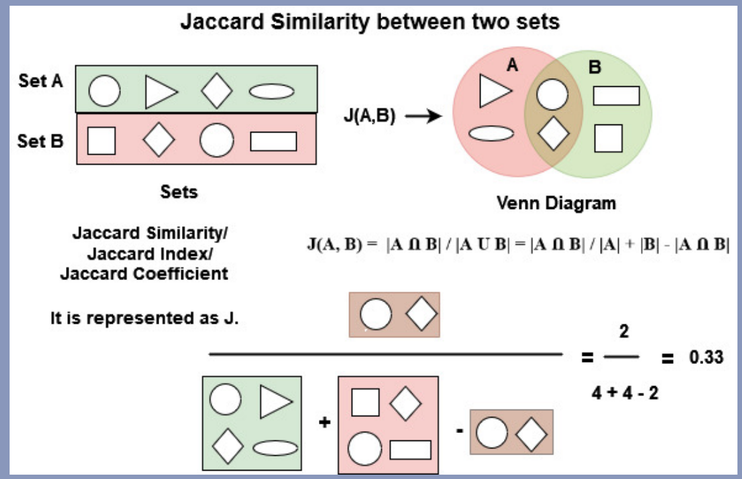
- QUANTIFY THE SIMILARITY BETWEEN TWO VECTORS IN A HIGH-DIMENSIONAL SPACE;
- MEASURES THE ANGLE BETWEEN TWO VECTORS RATHER THAN THEIR ABSOLUTE MAGNITUDE;
- IT RANGES FROM -1 TO 1.



- MEASURE OF THE DISTANCE BETWEEN A POINT AND A DISTRIBUTION;
- CONSIDERS THE CORRELATIONS BETWEEN VARIABLES, ALLOWING FOR A MORE ACCURATE REPRESENTATION OF THE TRUE DISTANCE BETWEEN POINTS IN A DATASET.



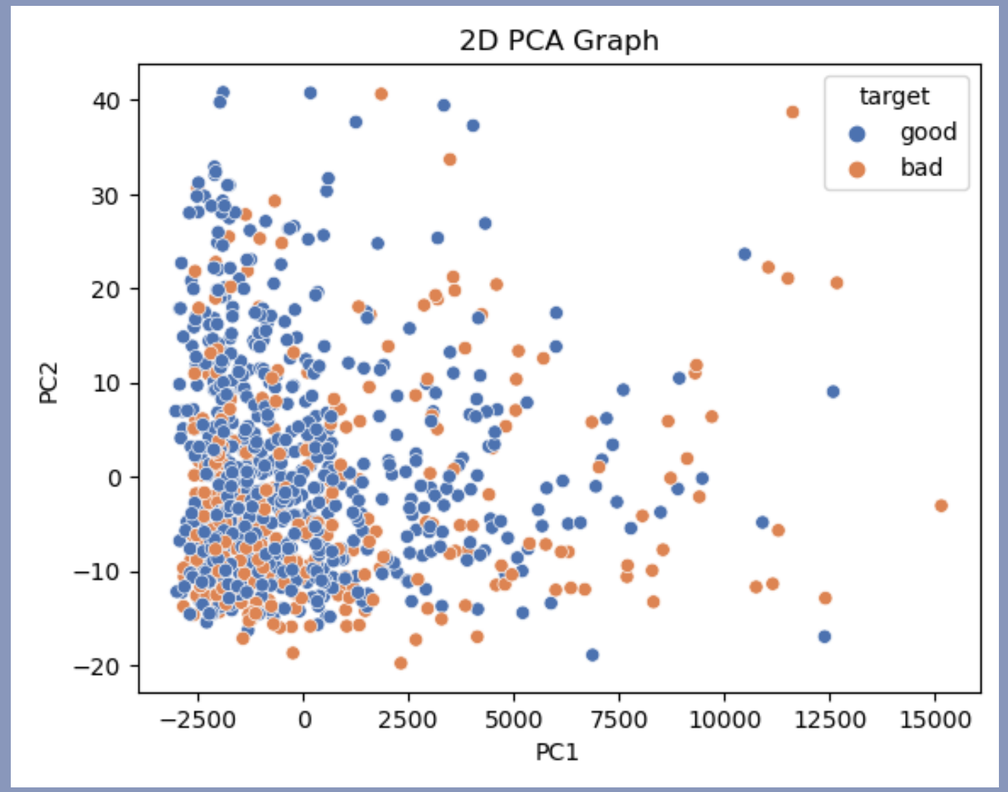
- CAN BE A SUITABLE CHOICE IN CERTAIN SCENARIOS, ESPECIALLY WHEN DEALING WITH FEATURES THAT HAVE A GRID-LIKE OR CATEGORICAL NATURE;
- DISTANCE BETWEEN TWO POINTS IN A GRID-LIKE SPACE;
- CAN BE DEFINED IN A HIGHER-DIMENSIONAL SPACE.



- A MEASURE OF SIMILARITY BETWEEN TWO SETS;
- IN THE CONTEXT OF KNN ALGORITHM, CAN BE USED AS A DISTANCE METRIC TO DETERMINE THE SIMILARITY BETWEEN TWO DATA POINTS;
- IS DEFINED AS THE SIZE OF THE INTERSECTION OF THE SETS DIVIDED BY THE SIZE OF THEIR UNION.

K-NEAREST NEIGHBOURS - BAGGING

DATASET WITH MORE OUTLIERS



CHOICE OF THE DATASET

WE STARTED BY ANALYZING ALL AVAILABLE DATASETS.
AFTER THIS SORTING, THE ONE WITH THE MOST OUTLIERS WAS CHOSEN FOR THE STUDY IN QUESTION WHOSE ID IS 31.

THIS DATASET CONSISTS OF:

- 1000 ROWS * 21 COLUMNS;
- TWO CLASSES ON THE TARGET: 'GOOD' AND 'BAD'.

THE NOTEBOOK

OUR NOTEBOOK CONSISTS OF 3 MAIN PARTS:

- KNN CLASSES
- STATISTICS
- TRAIN FUNCTIONS

TO BEGIN WE SPLITTED THE DATASET AND USED ONLY THE NUMERIC FEATURES.

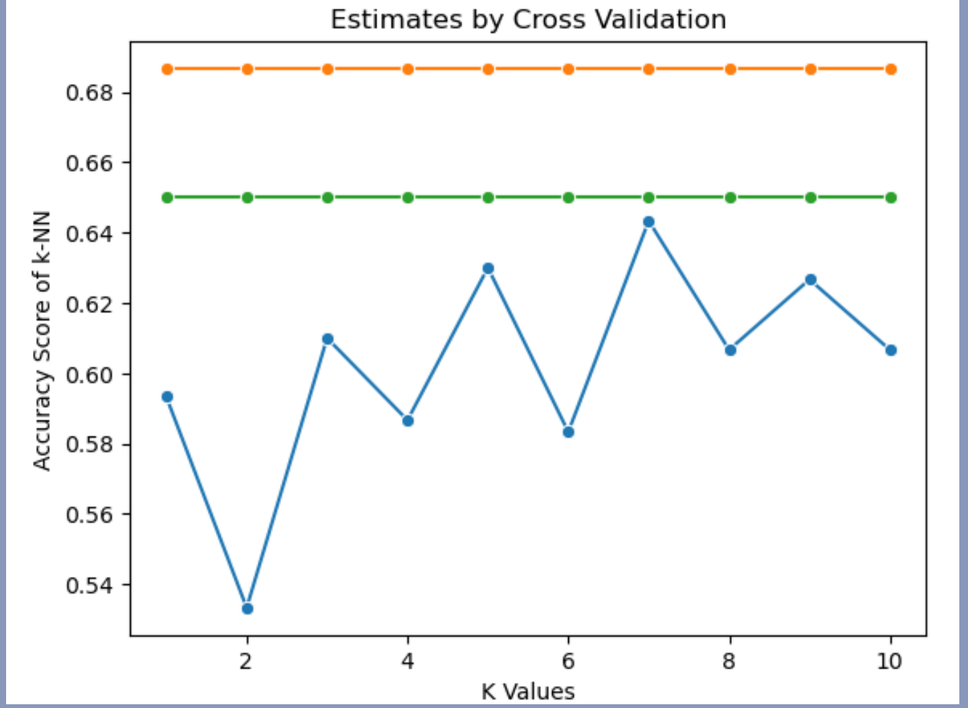
SO THAT THE USER COULD EXPERIENCE THE DIFFERENT MODELS SEPARATELY, HE CAN RUN EACH ONE INDIVIDUALLY.

1, 2 OR 3. WHICH ONE IS ASSOCIATED WITH A DIFFERENT MODEL: CASE BASE, BAGGING (DIFFERENT DISTANCES) AND DIFFERENT FEATURES, RESPECTIVELY.

TO THE REFERENCED DATASET WE OBTAINED THE FOLLOWED RESULT:

COMPARISON OF THE THREE MODELS

```
Best k Caso Base = 7
Best k Bagging = 1
Best k Features = 1
```



NOTE: UNFORTUNATLY WE COULD NOT OBTAIN THE ROC CURVE TO THE THIRD CASE

K-NEAREST NEIGHBOURS - BAGGING

RESULT ANALYSIS

PERFORMANCE METRICS

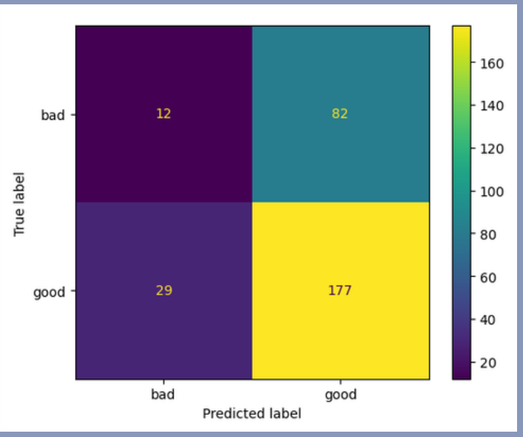
CONFUSION MATRIX

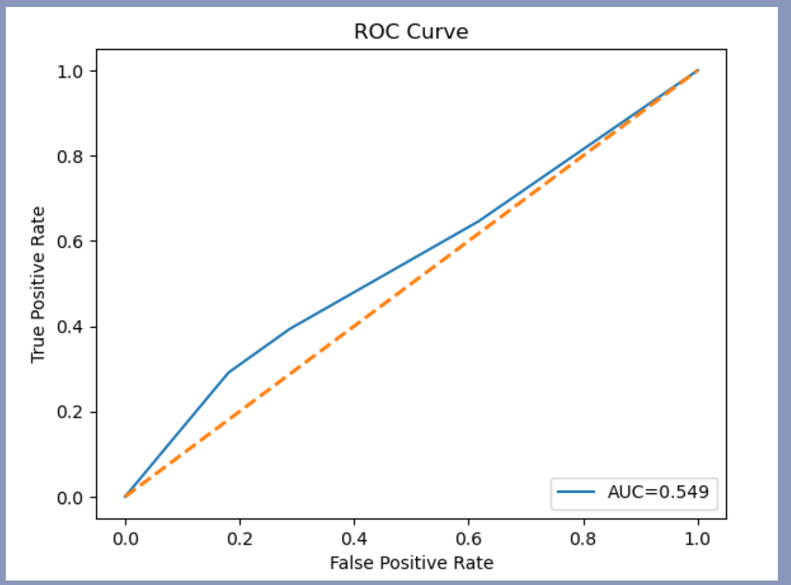
ROC CURVE

1. KNN - BASE CASE

```
dataset = oml.datasets.get_datase
caso = 1 # 1-caso_base 2-bagging
estatisticas(dataset,5,caso) #->
```

```
F1-score: 0.5784563918757467
Accuracy: 0.63
Precision: 0.56097372633958
Recall: 0.63
Error Rate: 0.37
Sensitivity: 0.1276595744680851
Specificity: 0.8592233009708737
```

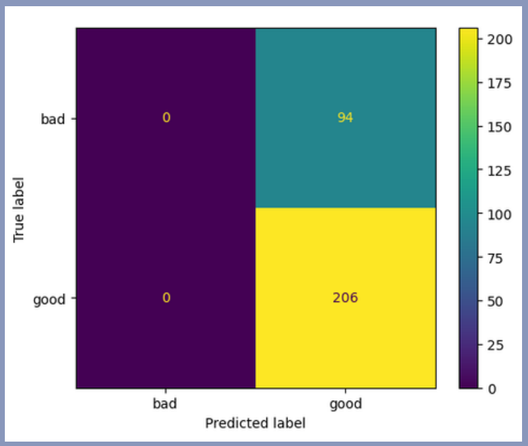


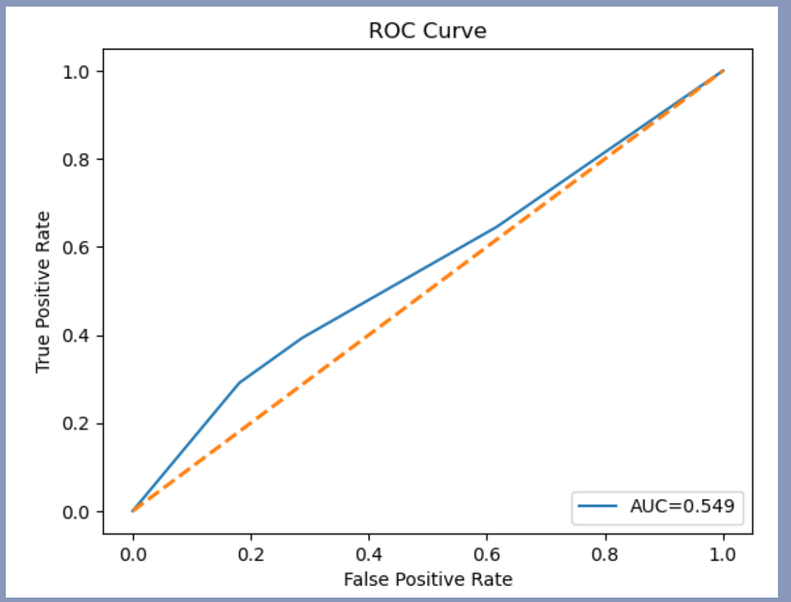


2. BAGGING BY DIFFERENT DISTANCE METRICS

```
dataset = oml.datasets.get_datase
caso = 2 # 1-caso_base 2-bagging
estatisticas(dataset,5,caso) #->
```

```
F1-score: 0.5591040843214757
Accuracy: 0.6866666666666666
Precision: 0.7848444444444443
Recall: 0.6866666666666666
Error Rate: 0.31333333333333335
Sensitivity: 0.0
Specificity: 1.0
```

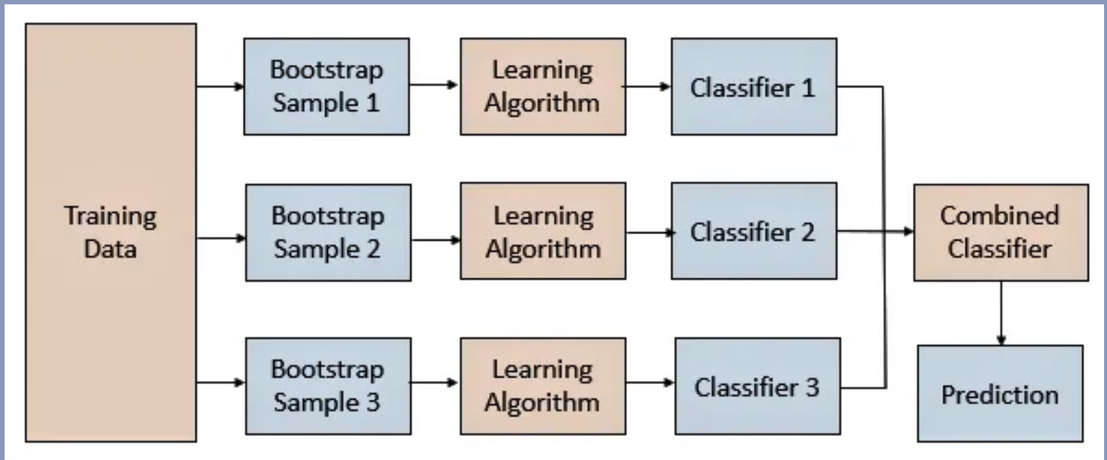




U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

K-NEAREST NEIGHBOURS - BAGGING

OVERFITTING



MITIGATE THE IMPACT OF OVERFITTING

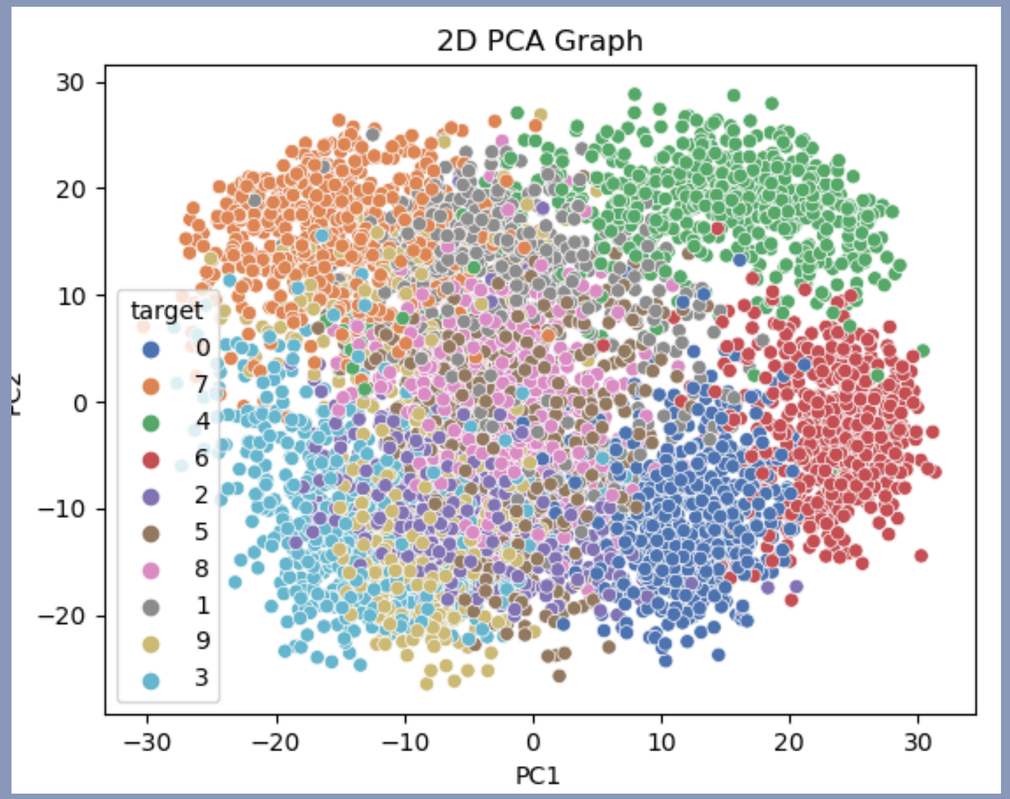
- BAGGING CAN BE APPLIED BY TRAINING MULTIPLE KNN MODELS WITH DIFFERENT SUBSETS OF FEATURES AND THEN AGGREGATING THEIR PREDICTIONS;
- EACH KNN MODEL IS TRAINED USING A SUBSET OF THE AVAILABLE FEATURES;
- THE INDIVIDUAL KNN MODELS FOCUS ON DIFFERENT ASPECTS OR REPRESENTATIONS OF THE DATA.

BY TRAINING MULTIPLE KNN MODELS WITH DIFFERENT FEATURE SUBSETS, BAGGING HELPS TO REDUCE OVERFITTING IN THE FOLLOWING WAYS:

- INCREASED DIVERSITY AMONG MODELS HELPS TO MITIGATE OVERFITTING BY REDUCING THE RELIANCE ON A SPECIFIC SET OF FEATURES;
- CAN MAKE THE MODELS LESS SENSITIVE TO NOISY OR IRRELEVANT FEATURES, IF CERTAIN FEATURES ARE NOISY OR HAVE LITTLE PREDICTIVE POWER, THE INCLUSION OF DIFFERENT SUBSETS OF FEATURES ALLOWS THE MODELS TO LEARN MORE ROBUST PATTERNS;
- AGGREGATING PREDICTIONS FROM MULTIPLE MODELS HELPS TO SMOOTH OUT INDIVIDUAL MODEL ERRORS AND BIASES. IT ALLOWS THE ENSEMBLE TO MAKE DECISIONS BASED ON A MORE ROBUST AND COLLECTIVE PERSPECTIVE, WHICH CAN LEAD TO IMPROVED GENERALIZATION PERFORMANCE.

K-NEAREST NEIGHBOURS - BAGGING

DATASET MORE LIKELY TO HAVE OVERFITTING



CHOICE OF THE DATASET

WE STARTED BY ANALYZING ALL AVAILABLE DATASETS.
AFTER THIS SORTING, THE ONE MOST LIKELY TO HAVE OVERFITTING WAS CHOSEN FOR THE STUDY IN QUESTION WHOSE ID IS 28.

THIS DATASET CONSISTS OF:

- 5620 ROWS * 65 COLUMNS;
- 10 CLASSES ON THE TARGET: THE NUMBERS FROM 0 TO 9.

NOTE: UNFORTUNATLY WE COULD NOT OBTAIN THE ROC CURVE

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

K-NEAREST NEIGHBOURS - BAGGING

RESULT ANALYSIS

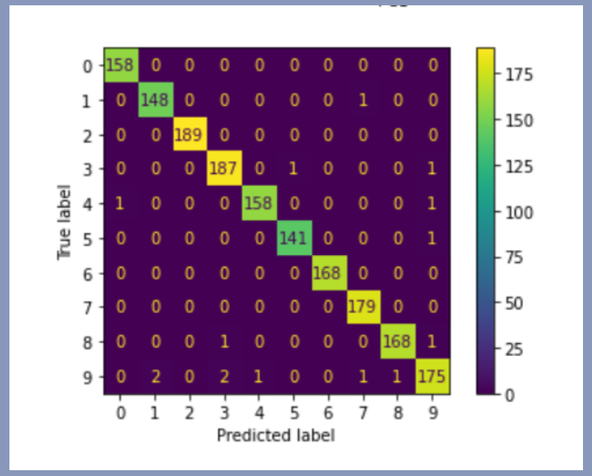
1. KNN - BASE CASE

PERFORMANCE METRICS

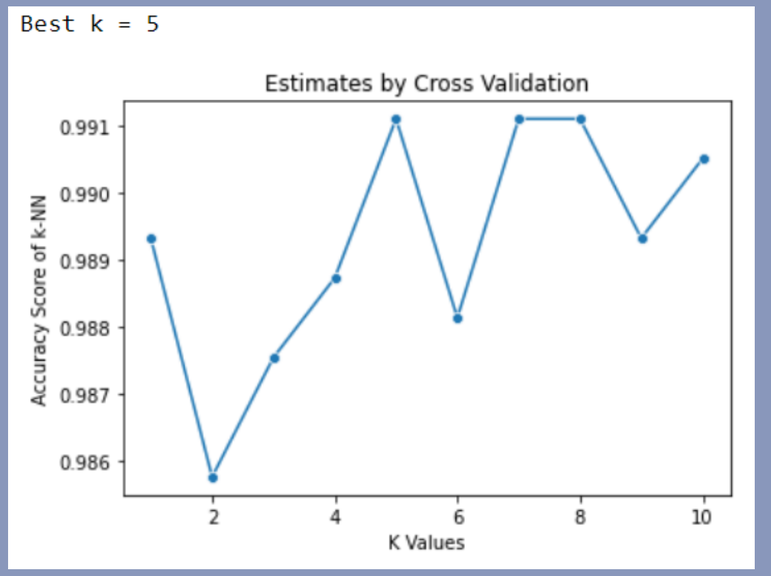
```
dataset = oml.datasets.get_dataset(28)
caso = 1
estatisticas(dataset,5,caso)
```

```
F1-score: 0.9910817986517135
Accuracy: 0.9911032028469751
Precision: 0.991090311974272
Recall: 0.9911032028469751
Error Rate: 0.008896797153024938
Sensitivity: 1.0
Specificity: 1.0
```

CONFUSION MATRIX

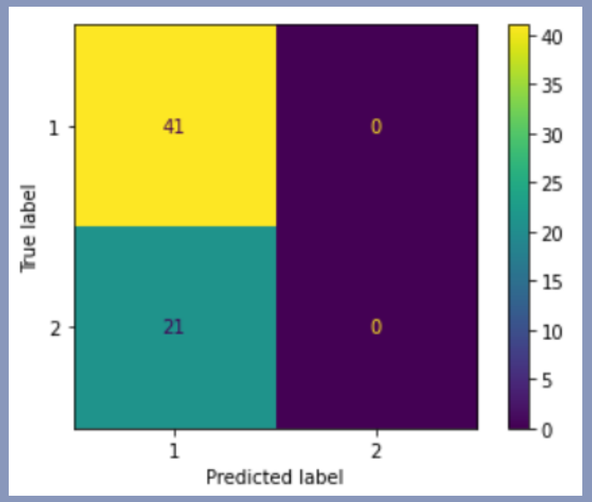


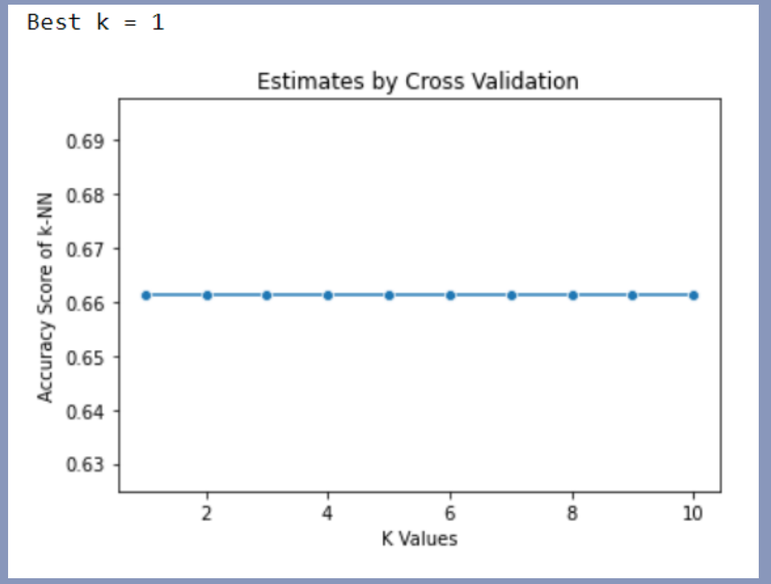
CROSS VALIDATION



3. BAGGING BY DIFFERENT FEATURES

```
F1-score: 0.5264641403069213
Accuracy: 0.6612903225806451
Precision: 0.7760145681581685
Recall: 0.6612903225806451
Error Rate: 0.33870967741935487
Sensitivity: 1.0
Specificity: 0.0
```





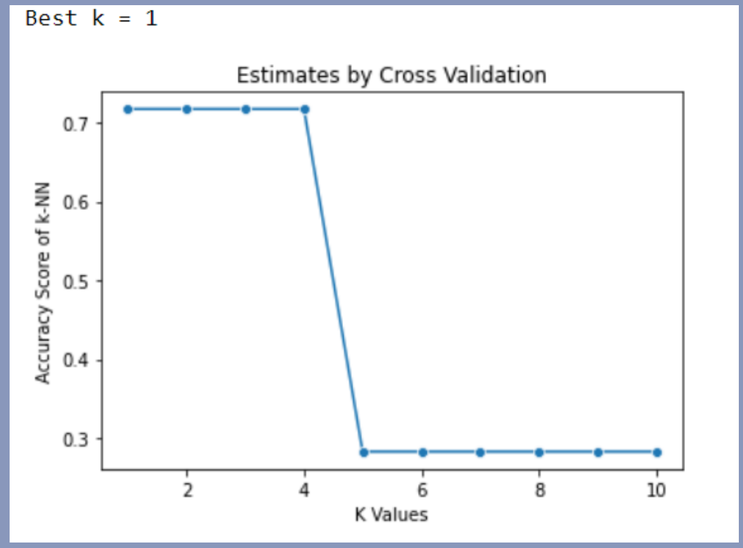
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

K-NEAREST NEIGHBOURS - BAGGING

CONCLUSION AND FUTURE WORK

AFTER ANALYSING THE RESULTS WE REACHED SEVERAL CONCLUSIONS:

- IN GENERAL, A DECREASE IN ACCURACY WAS NOTED. HOWEVER, IN SOME CASES, IT INCREASED SUCH AS IN THE CASE OF THE FIRST DATASET REFERRED HERE.
- CHANGING THE NUMBER OF NEIGHBOURS IN OUR IMPLEMENTATIONS DID NOT CHANGE THE RESULTS IN MOST CASES WITH THE EXCEPTION OF A FEW, SUCH AS THE ONE SHOWN IN THE IMAGE.
- WE ALSO CONCLUDE THAT IN FUTURE WORK, WE SHOULD CHANGE OUR APPROACH TO THE PROBLEM BY CHANGING OUR EVALUATION METHODS.
- ONE OF THE MAIN ISSUES THAT WE ENCOUNTERED WAS THE TIME OF EXECUTION. THE METHODS PROVED TO BE INEFFICIENT.



ID = 28, CASE = 2