Capstone Project

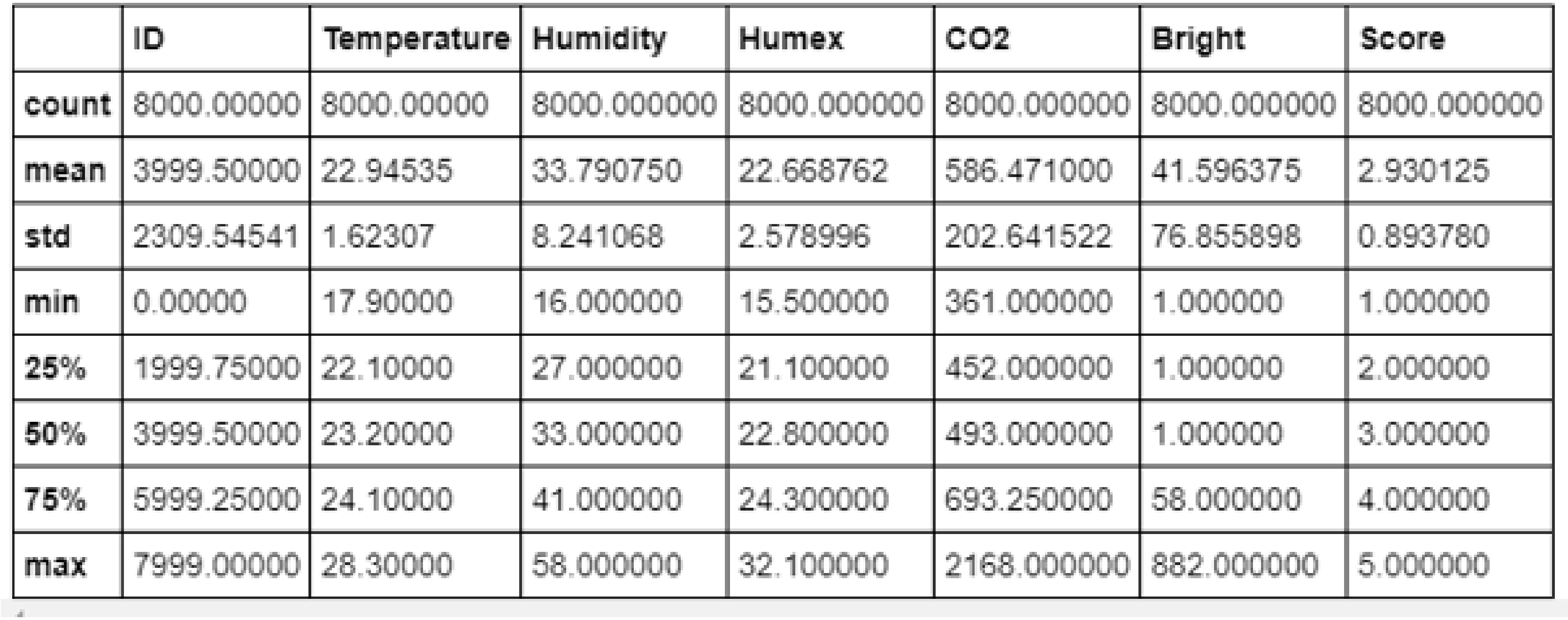
Optimazing well being at work

BERRADI Aymane

Data Description

❑**ID:** integer, uniquely identifies each observation,
❑**String:** defines date under format yyyy-mm-¬d hh:mm:ss,
❑**Temperature:** real number, temperature inside the room,
❑**Humidity:** real number, humidity of ambient air in the room,
❑**Humex:** real number, indicator of air quality in the room,
❑**CO2:** Integer, CO2 level in the room, in ppm (parts per million),
❑**Bright:** Integer, characterizes the brightness of the room,
❑**Score:** The classes are {1,2,3,4,5}, 5 being the optimal comfort and 1 the worst.

Data Description

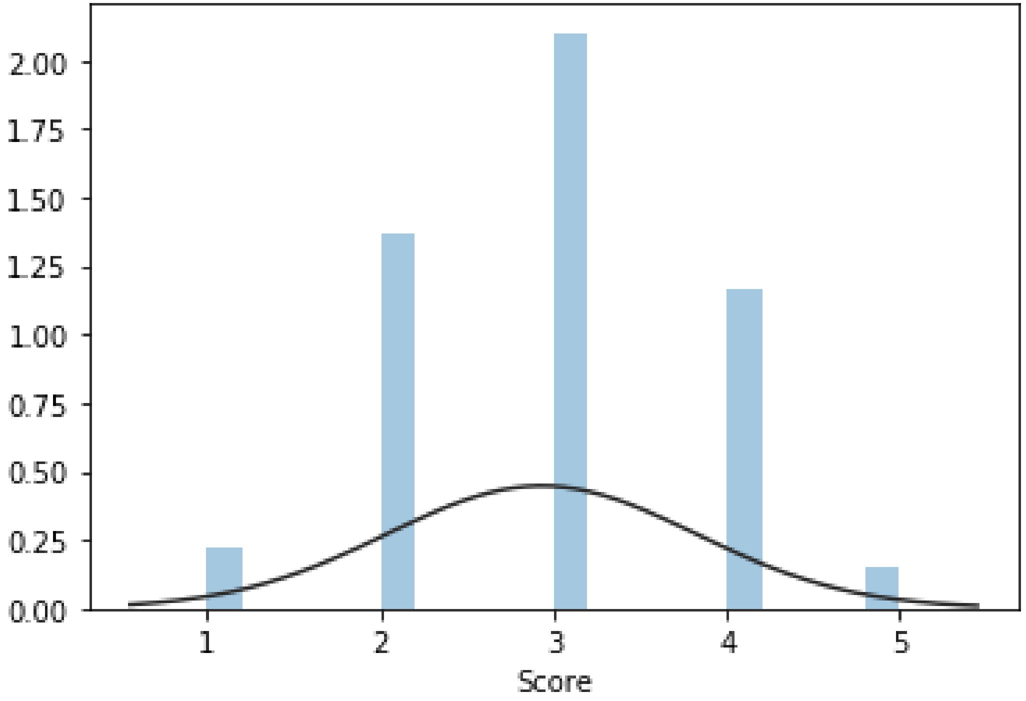


	ID	Temperature	Humidity	Humex	CO2	Bright	Score
count	8000.00000	8000.00000	8000.000000	8000.000000	8000.000000	8000.000000	8000.000000
mean	3999.50000	22.94535	33.790750	22.668762	586.471000	41.596375	2.930125
std	2309.54541	1.62307	8.241068	2.578996	202.641522	76.855898	0.893780
min	0.00000	17.90000	16.000000	15.500000	361.000000	1.000000	1.000000
25%	1999.75000	22.10000	27.000000	21.100000	452.000000	1.000000	2.000000
50%	3999.50000	23.20000	33.000000	22.800000	493.000000	1.000000	3.000000
75%	5999.25000	24.10000	41.000000	24.300000	693.250000	58.000000	4.000000
max	7999.00000	28.30000	58.000000	32.100000	2168.000000	882.000000	5.000000

- ❑ No ouliers for all variables
- ❑ 2168 ppm (maximun value for CO2)which is considered as dangerous level.

Data Visualisation

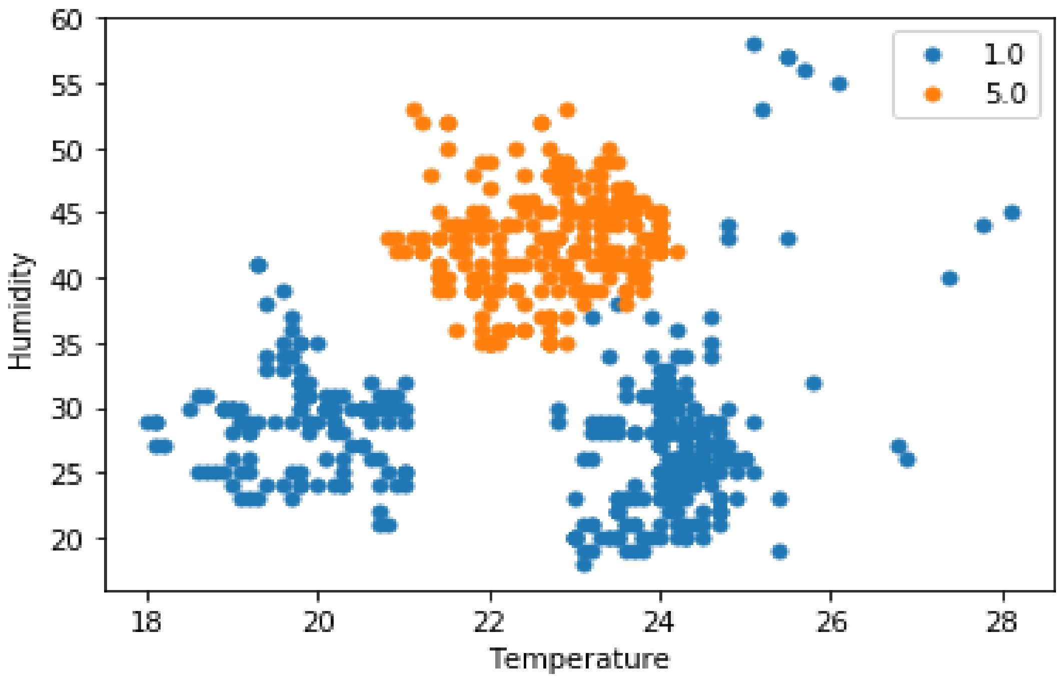
- Distrubution of the target variable



The Score variable has a **Normal Distribution**.

Data Visualisation

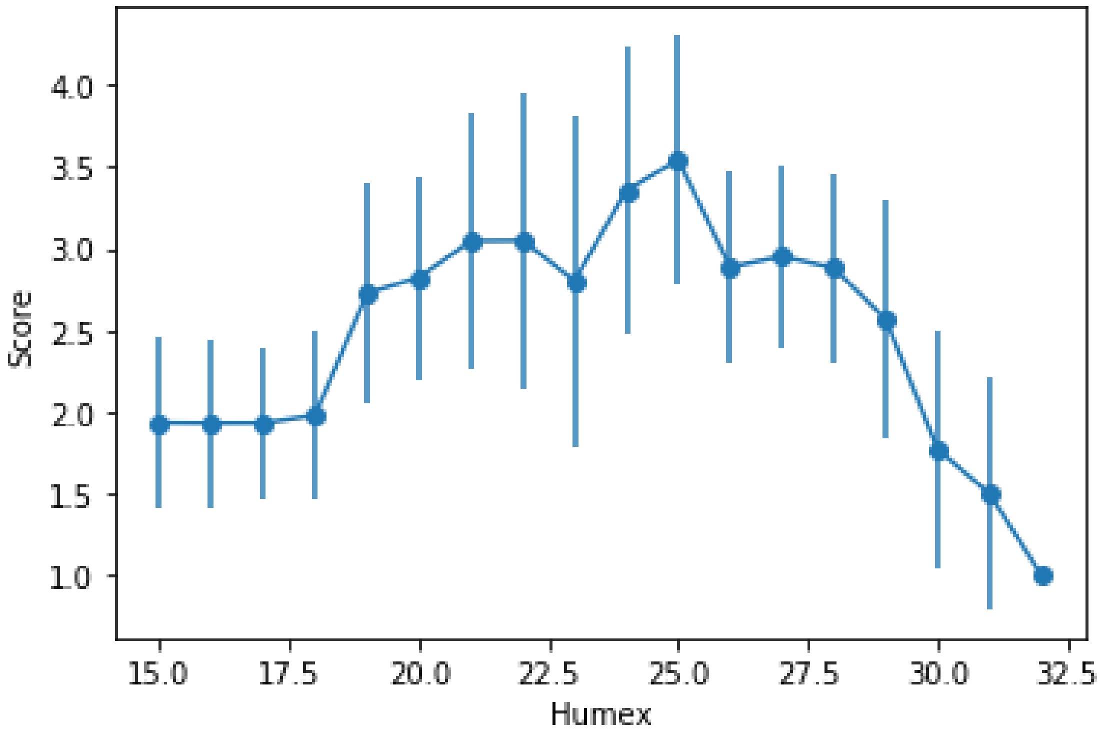
- Score Plot in function of physical variables



We can see that the best case (orange color) corresponds to a range of **(20.8 24.2)** of Temperature, and **(35 53)** of Humidity,

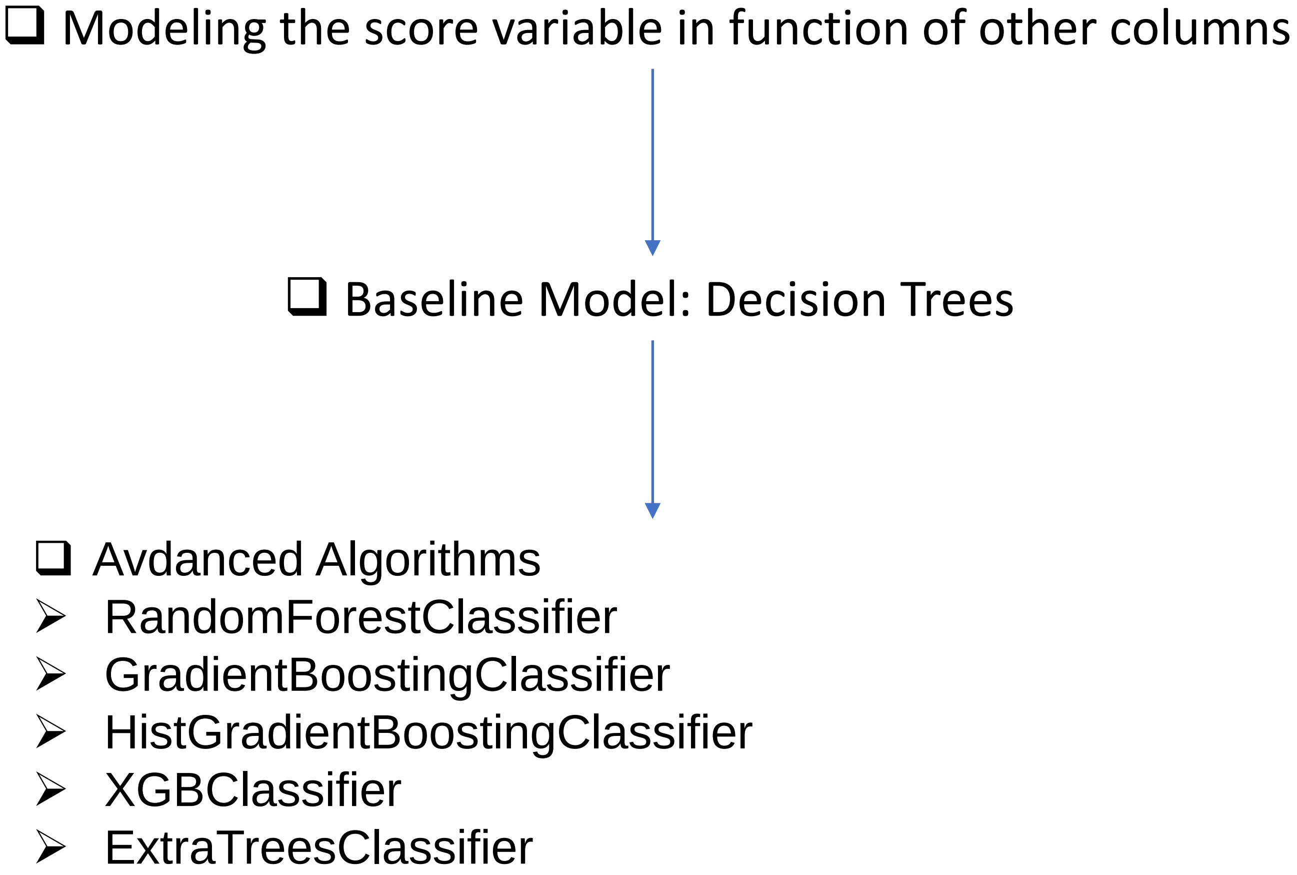
Data Visualisation

➢ Score Plot in function of Humex



❑ We can pick the optimum values that leads to the best values of Score.

Models



□ Modeling the score variable in function of other columns

□ Baseline Model: Decision Trees

□ Avdanced Algorithms

➢ RandomForestClassifier

➢ GradientBoostingClassifier

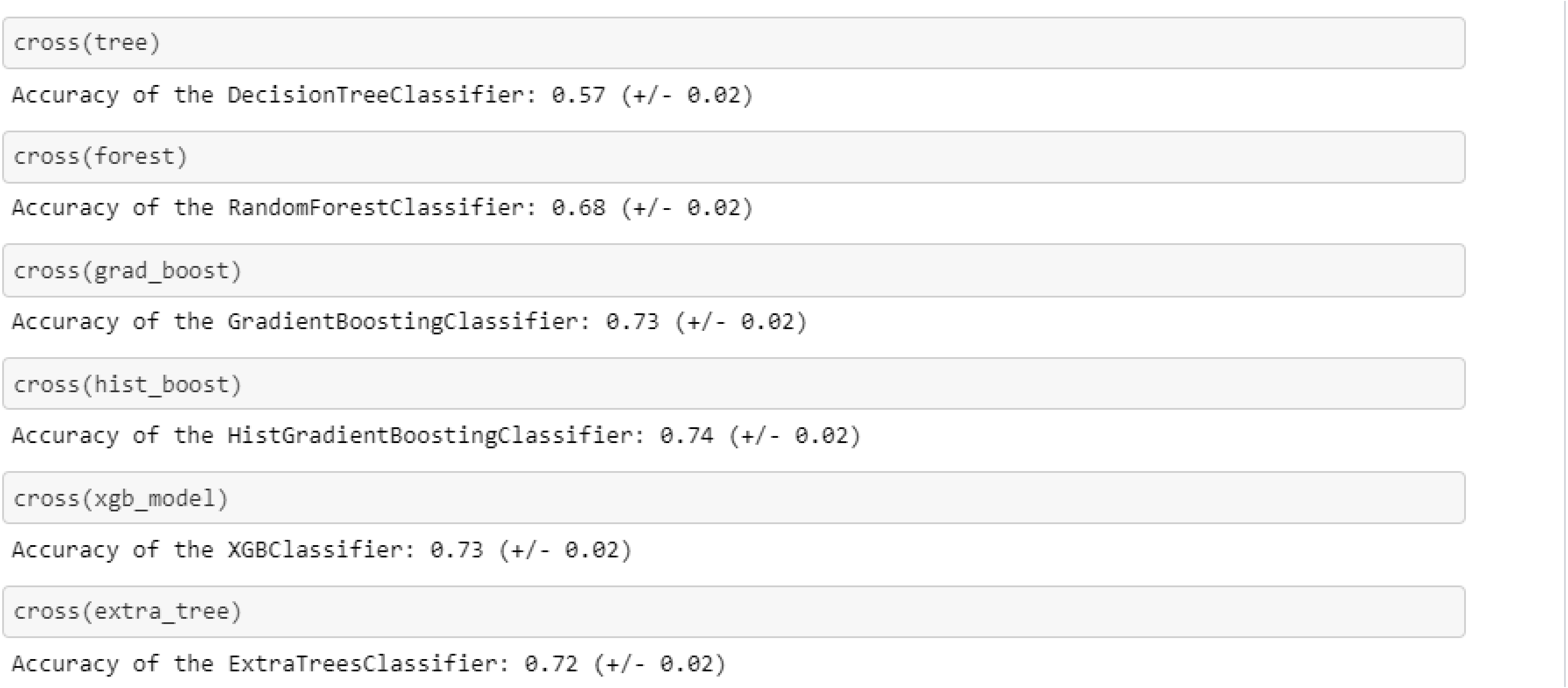
➢ HistGradientBoostingClassifier

➢ XGBClassifier

➢ ExtraTreesClassifier

Models: Results

Grid Search+Cross validation



```
cross(tree)
```

```
Accuracy of the DecisionTreeClassifier: 0.57 (+/- 0.02)
```

```
cross(forest)
```

```
Accuracy of the RandomForestClassifier: 0.68 (+/- 0.02)
```

```
cross(grad_boost)
```

```
Accuracy of the GradientBoostingClassifier: 0.73 (+/- 0.02)
```

```
cross(hist_boost)
```

```
Accuracy of the HistGradientBoostingClassifier: 0.74 (+/- 0.02)
```

```
cross(xgb_model)
```

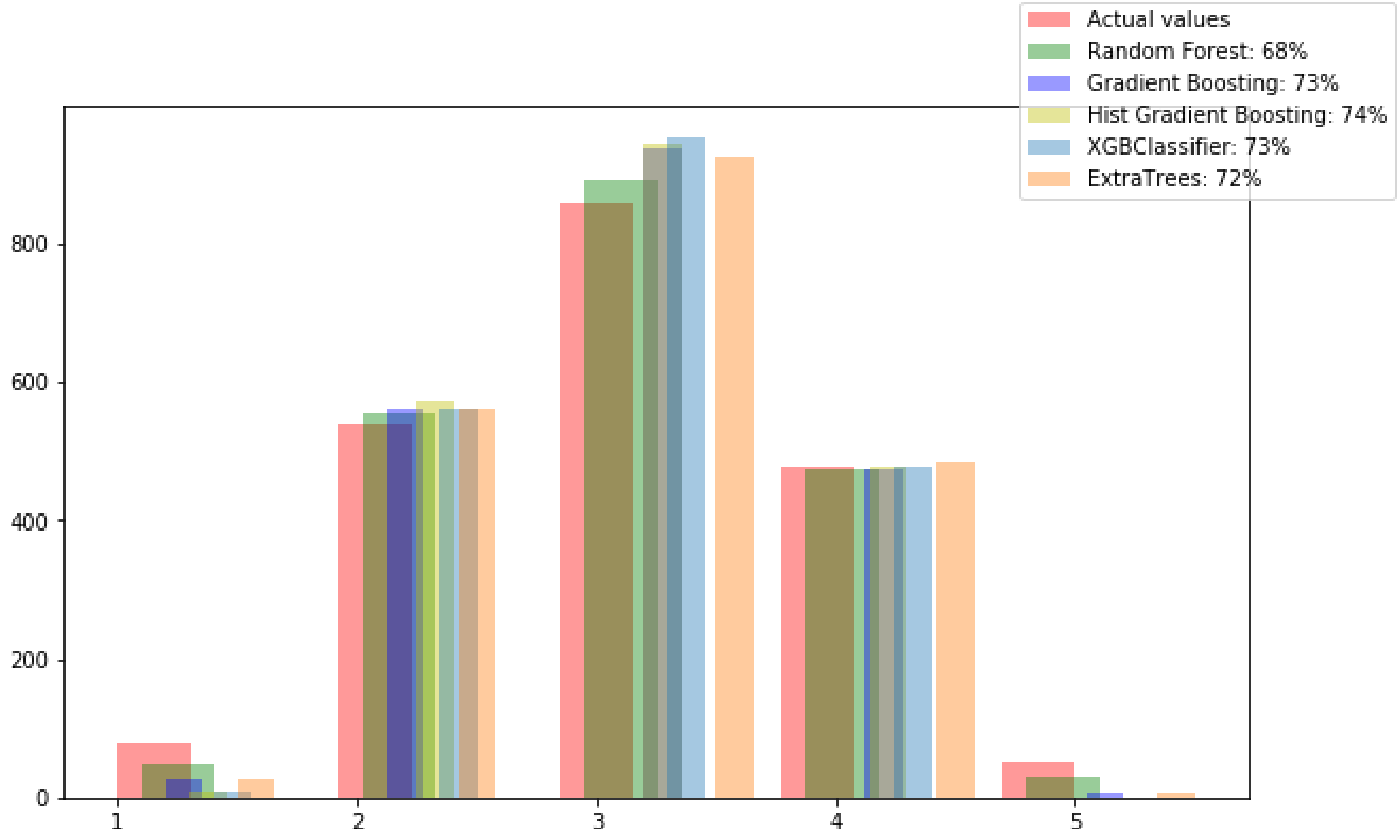
```
Accuracy of the XGBClassifier: 0.73 (+/- 0.02)
```

```
cross(extra_tree)
```

```
Accuracy of the ExtraTreesClassifier: 0.72 (+/- 0.02)
```

- ❑ The results of the accuracy show that **HistGradientBoostingClassifier** performs better than other models

Models: Visualisation



❑ we can observe that all classifiers **don't perfom well on predicting the classes 1 and 5**.

Futur Work:

❑ Improve the model by using the Stacking or the Voting method of all models