**LOOKING FOR A CLASSIFIER THAT IS ACCTUALLY DOING SOMETHING.**

|  |
| --- |
| SVM (all kernels tested with same result) |
| Classifier type: SVM  ----- NOT NORMALIZED DATA -----  Confusion Matrix:  [[1598 0]  [ 91 0]]  ----- NORMALIZED DATA -----  Confusion Matrix:  [[1589 0]  [ 100 0]] |

Conclusions:

This classifier does not a good job, because both with normalized and not normalized data confusion matrix shows us that classifier always says that it is no raining/snowing.

|  |
| --- |
| MLP |
| Classifier type: MLP  ----- NOT NORMALIZED DATA -----  Confusion Matrix:  [[1601 0]  [ 88 0]]  ----- NORMALIZED DATA -----  Confusion Matrix:  [[1609 0]  [ 80 0]] |

Conclusions:

This classifier does not a good job, because both with normalized and not normalized data confusion matrix shows us that classifier always says that it is no raining/snowing.

|  |
| --- |
| K Neighbors |
| Classifier type: K Neighbors  ----- NOT NORMALIZED DATA -----  Confusion Matrix:  [[1586 7]  [ 18 78]]  ----- NORMALIZED DATA -----  Confusion Matrix:  [[1602 0]  [ 87 0]] |

Conclusions:

This classifier is working for not normalized data.

|  |
| --- |
| Decision Tree |
| Classifier type: Decision Tree  ----- NOT NORMALIZED DATA -----  Confusion Matrix:  [[1596 10]  [ 33 50]]  ----- NORMALIZED DATA -----  Confusion Matrix:  [[1363 252]  [ 49 25]] |

Conclusions:

This classifier is working for normalized & not normalized data.

**OK, NOW WE KNOW THAT KNEIGHBORS FOR NOT NORMALIZED DATA & DECISION TREE FOR NORMALIZED & NOT NORMALIZED IS DOING SOMETHING. LET’S COMPARE OUTPUTS WHEN WE DON’T HAVE RENTED BIKE COUNT FEATURA AND WHEN WE DO.**

|  |
| --- |
| KNEIGHBORS WITHOUT RENTED BIKE COUNT |
| Classifier type: K Neighbors  ----- NOT NORMALIZED DATA -----  Accuracies(train/test/validation):  0.984375  0.9804618117229129  0.98698224852071  Confusion Matrix:  [[1594 7]  [ 26 62]]  Cross validation:  CV: 5  Accuracy: 0.9732488291875824  F1: 0.7063745667260253  Recall: 0.6372093023255814  Precision: 0.8001807995925643 |

|  |
| --- |
| KNEIGHBORS WITH RENTED BIKE COUNT |
| Classifier type: K Neighbors  ----- NOT NORMALIZED DATA -----  Accuracies(train/test/validation):  0.986032196969697  0.9863824748371818  0.9857988165680474  Confusion Matrix:  [[1592 7]  [ 16 74]]  Cross validation:  CV: 5  Accuracy: 0.9775100255195042  F1: 0.7499174386554307  Recall: 0.7110975609756097  Precision: 0.7990655064269975 |

|  |
| --- |
| DECISION TREE WITHOUT RENTED BIKE COUNT |
| Classifier type: Decision Tree  ----- NOT NORMALIZED DATA -----  Accuracies(train/test/validation):  0.9637784090909091  0.9597394908229722  0.9664694280078896  Confusion Matrix:  [[1598 3]  [ 65 23]]  Cross validation:  CV: 5  Accuracy: 0.9585689447264365  F1: 0.4340604990247848  Recall: 0.3560975609756098  Precision: 0.6542201957163594  ----- NORMALIZED DATA -----  Accuracies(train/test/validation):  0.9649621212121212  0.8052101835405565  0.7972386587771203  Confusion Matrix:  [[1340 270]  [ 59 20]]  Cross validation:  CV: 5  Accuracy: 0.9580964132477074  F1: 0.4042007568426355  Recall: 0.3282051282051282  Precision: 0.6342224339128364 |

|  |
| --- |
| DECISION TREE WITH RENTED BIKE COUNT |
| Classifier type: Decision Tree  ----- NOT NORMALIZED DATA -----  Accuracies(train/test/validation):  0.9675662878787878  0.9698046181172292  0.9660749506903353  Confusion Matrix:  [[1580 30]  [ 21 58]]  Cross validation:  CV: 5  Accuracy: 0.9559670770352505  F1: 0.5322635222807695  Recall: 0.49523809523809526  Precision: 0.6048039118849504  ----- NORMALIZED DATA -----  Accuracies(train/test/validation):  0.9730113636363636  0.7986974541148608  0.8047337278106509  Confusion Matrix:  [[1326 283]  [ 57 23]]  Cross validation:  CV: 5  Accuracy: 0.9704085925011918  F1: 0.5830388031790557  Recall: 0.4780938833570412  Precision: 0.77587841191067 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ***WITHOUT RENTED BIKE COUNT*** | | | ***WITH RENTED BIKE COUNT*** | | |
|  | **DECISION TREE** | | **K NEIGHBOR** | **DECISION TREE** | | **K NEIGHBOR** |
| **mesure** | **normalized** | **not-normalized** | **not - normalized** | **normalized** | **not-normalized** | **not - normalized** |
| ***ACCURACIES*** | | | | | | |
| **training accuracy** | 0,96 | 0,96 | 0,98 | 0,97 | 0,97 | 0,99 |
| **testing accuracy** | 0,81 | 0,96 | 0,98 | 0,8 | 0,97 | 0,99 |
| **validation accuracy** | 0,8 | 0,97 | 0,99 | 0,8 | 0,97 | 0,99 |
| ***CROSS VALIDATION*** | | | | | | |
| **Accuracy** | 0,96 | 0,96 | 0,97 | 0,97 | 0,96 | 0,98 |
| **F1** | 0,4 | 0,43 | 0,71 | 0,58 | 0,53 | 0,75 |
| **Recall** | **0,33** | 0,36 | 0,64 | **0,48** | 0,5 | 0,71 |
| **Precision** | **0,63** | 0,65 | 0,8 | **0,78** | 0,6 | 0,8 |
| ***CONFUSION MATRIX*** | | | | | | |
| **TP** | 20 | 23 | 62 | 23 | 58 | 74 |
| **FP** | 270 | 3 | 7 | 283 | 30 | 7 |
| **TN** | 1340 | 1598 | 1594 | 1326 | 1580 | 1592 |
| **FN** | 59 | 65 | 26 | 57 | 21 | 16 |

**NOW LET’S PUT ALL THIS DATA IN ONE TABLE FOR BETTER VISIBILITY**

***Measures that had improved after adding Rented Bike Count are highlighted.***

As we can see accuracy (which we decided to test in our hypothesis) in every example is quite high. This is caused by data distribution in our dataset. We have a lot of data samples when it was no Precipitation and relatively small amount of data with any Rainfall or Snowfall.

**TU WSTAWIĆ HISTOGRAM PRECIPITATION**

At this point we know that our hypothesis (to achieve 20% increase in accuracy after adding Rented Bike Cunt) is impossible to accomplish. Although we can try to compare different measures to see how adding Rented Bike Count influenced our models.

Looking at the table above we can see that adding Rented Bike Cunt have indeed improved Recall in every tested model. Let us have a closer look on that and compare TPR (Recall) , TNR and Precision before and after adding this extra feature. (Now we calculate this values from confusion matrix, not cross validation).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ***WITHOUT RENTED BIKE COUNT*** | | | ***WITH RENTED BIKE COUNT*** | | |
|  | **DECISION TREE** | | **K NEIGHBOR** | **DECISION TREE** | | **K NEIGHBOR** |
| **mesure** | **normalized** | **not-normalized** | **not - normalized** | **normalized** | **not-normalized** | **not - normalized** |
| TPR | **0,25** | **0,26** | **0,70** | **0,29** | **0,73** | **0,82** |
| TNR | 0,83 | 1,00 | 1,00 | 0,82 | 0,98 | 1,00 |
| Precission | 0,07 | 0,88 | 0,90 | 0,08 | 0,66 | 0,91 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ***WITHOUT RENTED BIKE COUNT*** | | | ***WITH RENTED BIKE COUNT*** | | |
|  | **DECISION TREE** | | **K NEIGHBOR** | **DECISION TREE** | | **K NEIGHBOR** |
| **mesure** | **normalized** | **not-normalized** | **not - normalized** | **normalized** | **not-normalized** | **not - normalized** |
| TPR (mean of 100) | **0,45** | **0,41** | **0,77** | **0,50** | **0,54** | **0,77** |

We see that adding Rented Bike Count seems to have improved outcomes of every model. Now let’s test statistically this measure. We calculate TPR one hundred times and calculate mean.

Now we see that for decision tree model (simple classifier) we indeed see some improvement (by 0,05 and 0,13 for normalized and not normalized data), but for K Neighbour (more complexed classifier) we don’t see any improvement.

CONCLUSION:

* Our hypothesis is not true.
* Adding Rented Bike Count seems to help a bit for decision tree classifier, but it is not that big improvement to draw strong conclusions.

WHAT COULD BE DONE NEXT:

* Test this hypothesis on different data set. (Maybe from different city where rainy and not rainy days distribution is more equal.)
* Change hypothesis to predict different, more represented weather feature (e.g. wind) and check the results.