1D CNN FOR WEATHER PREDICTION USING OPTUNA

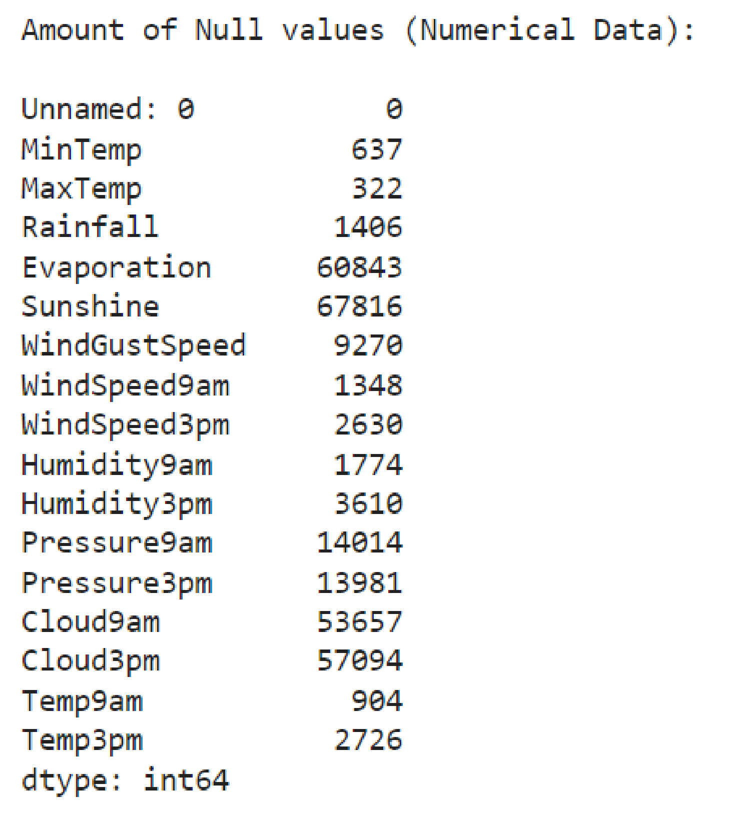
(AND COMPARISON WITH ML MODELS)

LIBRARIES

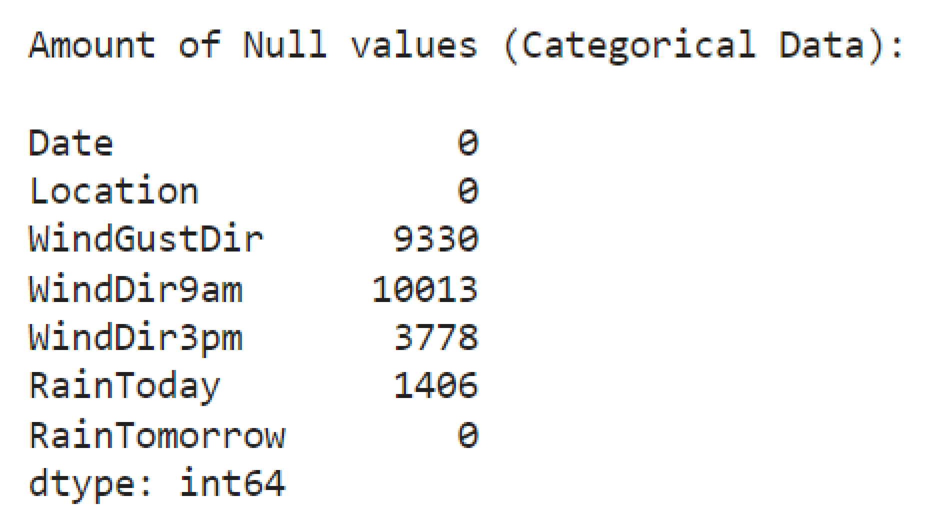
- Tensorflow → Used for neural network implementation
- Scikit-learn → Used for machine learning algorithms implementation
- Pandas → Used for data handling
- Numpy → Used for data handling
- Optuna → Used for deciding best hyperparameters
- Seaborn → Used for visualization
- Matplotlib → Used for visualization
- Scikit-plot → Used for visualization
- Joblib → Used for saving machine learning model

DATA PREPROCESSING

1. Checking for the categorical and numerical data and Null values



```
Amount of Null values (Numerical Data):

Unnamed: 0          0
MinTemp           637
MaxTemp           322
Rainfall         1406
Evaporation     60843
Sunshine        67816
WindGustSpeed    9270
WindSpeed9am     1348
WindSpeed3pm     2630
Humidity9am      1774
Humidity3pm      3610
Pressure9am     14014
Pressure3pm     13981
Cloud9am        53657
Cloud3pm        57094
Temp9am           904
Temp3pm          2726
dtype: int64
```



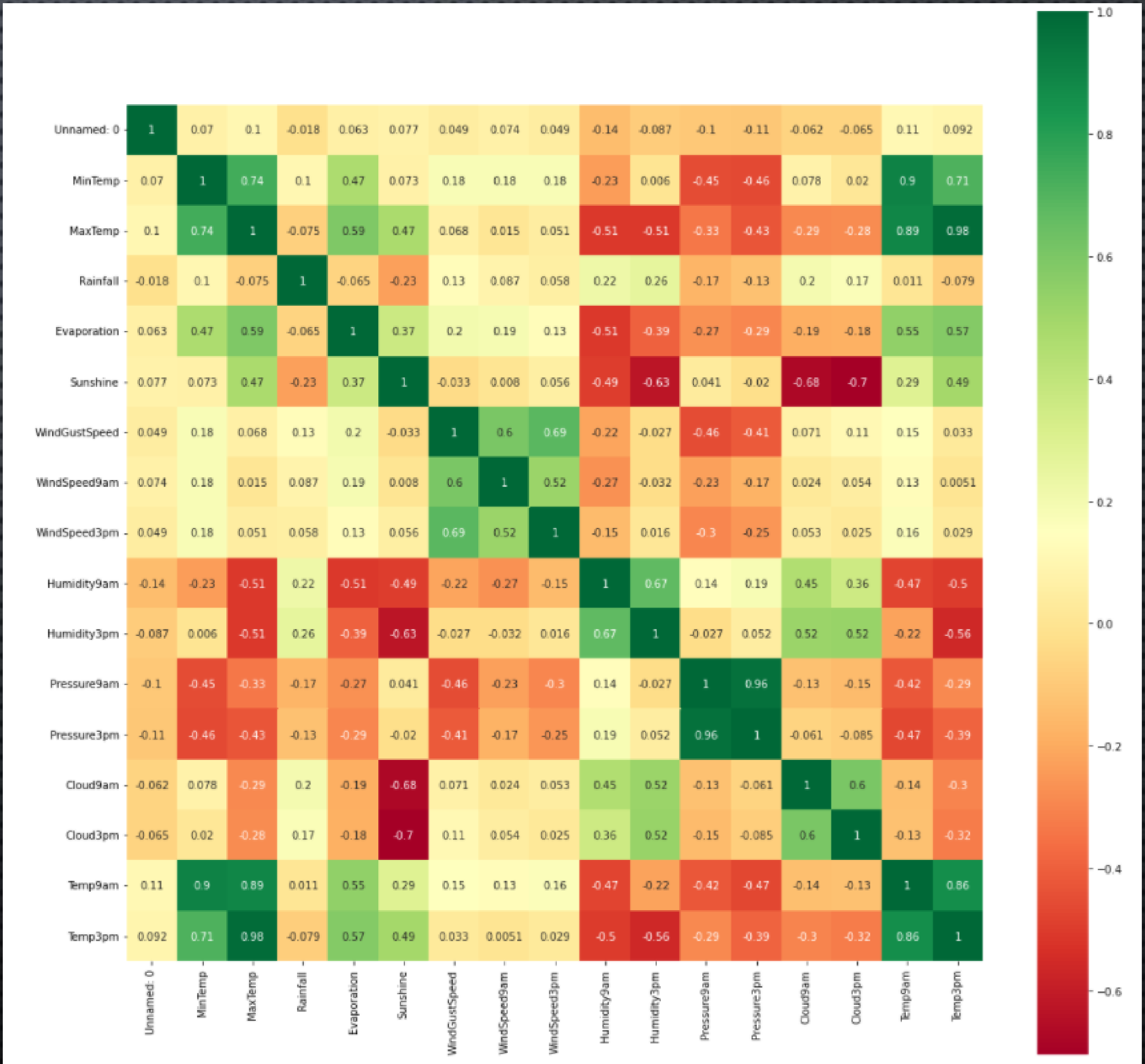
```
Amount of Null values (Categorical Data):

Date                0
Location            0
WindGustDir      9330
WindDir9am      10013
WindDir3pm       3778
RainToday        1406
RainTomorrow        0
dtype: int64
```

“Evaporation”, “Sunshine”, “Cloud9am”, “Cloud3pm” Columns have a high volume of missing data, They will be checked in the next step to decide for deletion of these columns.

2. **Checking the correlation among numeric data and delete unnecessary columns**

“Evaporation”, “Sunshine”, “Cloud9am”, “Cloud3pm” Columns were deleted due to the high volume of missing data and not have a strong correlation among the other numerical variables.

“Unnamed: 0” , “Date” and “Location” columns which are non-meaningful were also deleted.



	Unnamed: 0	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
Unnamed: 0	1	0.07	0.1	-0.018	0.063	0.077	0.049	0.074	0.049	-0.14	-0.087	-0.1	-0.11	-0.062	-0.065	0.11	0.092
MinTemp	0.07	1	0.74	0.1	0.47	0.073	0.18	0.18	0.18	-0.23	0.006	-0.45	-0.46	0.078	0.02	0.9	0.71
MaxTemp	0.1	0.74	1	-0.075	0.59	0.47	0.068	0.015	0.051	-0.51	-0.51	-0.33	-0.43	-0.29	-0.28	0.89	0.98
Rainfall	-0.018	0.1	-0.075	1	-0.065	-0.23	0.13	0.087	0.058	0.22	0.26	-0.17	-0.13	0.2	0.17	0.011	-0.079
Evaporation	0.063	0.47	0.59	-0.065	1	0.37	0.2	0.19	0.13	-0.51	-0.39	-0.27	-0.29	-0.19	-0.18	0.55	0.57
Sunshine	0.077	0.073	0.47	-0.23	0.37	1	-0.033	0.008	0.056	-0.49	-0.63	0.041	-0.02	-0.68	-0.7	0.29	0.49
WindGustSpeed	0.049	0.18	0.068	0.13	0.2	-0.033	1	0.6	0.69	-0.22	-0.027	-0.46	-0.41	0.071	0.11	0.15	0.033
WindSpeed9am	0.074	0.18	0.015	0.087	0.19	0.008	0.6	1	0.52	-0.27	-0.032	-0.23	-0.17	0.024	0.054	0.13	0.0051
WindSpeed3pm	0.049	0.18	0.051	0.058	0.13	0.056	0.69	0.52	1	-0.15	0.016	-0.3	-0.25	0.053	0.025	0.16	0.029
Humidity9am	-0.14	-0.23	-0.51	0.22	-0.51	-0.49	-0.22	-0.27	-0.15	1	0.67	0.14	0.19	0.45	0.36	-0.47	-0.5
Humidity3pm	-0.087	0.006	-0.51	0.26	-0.39	-0.63	-0.027	-0.032	0.016	0.67	1	-0.027	0.052	0.52	0.52	-0.22	-0.56
Pressure9am	-0.1	-0.45	-0.33	-0.17	-0.27	0.041	-0.46	-0.23	-0.3	0.14	-0.027	1	0.96	-0.13	-0.15	-0.42	-0.29
Pressure3pm	-0.11	-0.46	-0.43	-0.13	-0.29	-0.02	-0.41	-0.17	-0.25	0.19	0.052	0.96	1	-0.061	-0.085	-0.47	-0.39
Cloud9am	-0.062	0.078	-0.29	0.2	-0.19	-0.68	0.071	0.024	0.053	0.45	0.52	-0.13	-0.061	1	0.6	-0.14	-0.3
Cloud3pm	-0.065	0.02	-0.28	0.17	-0.18	-0.7	0.11	0.054	0.025	0.36	0.52	-0.15	-0.085	0.6	1	-0.13	-0.32
Temp9am	0.11	0.9	0.89	0.011	0.55	0.29	0.15	0.13	0.16	-0.47	-0.22	-0.42	-0.47	-0.14	-0.13	1	0.86
Temp3pm	0.092	0.71	0.98	-0.079	0.57	0.49	0.033	0.0051	0.029	-0.5	-0.56	-0.29	-0.39	-0.3	-0.32	0.86	1

3. Fill Null values with proper values

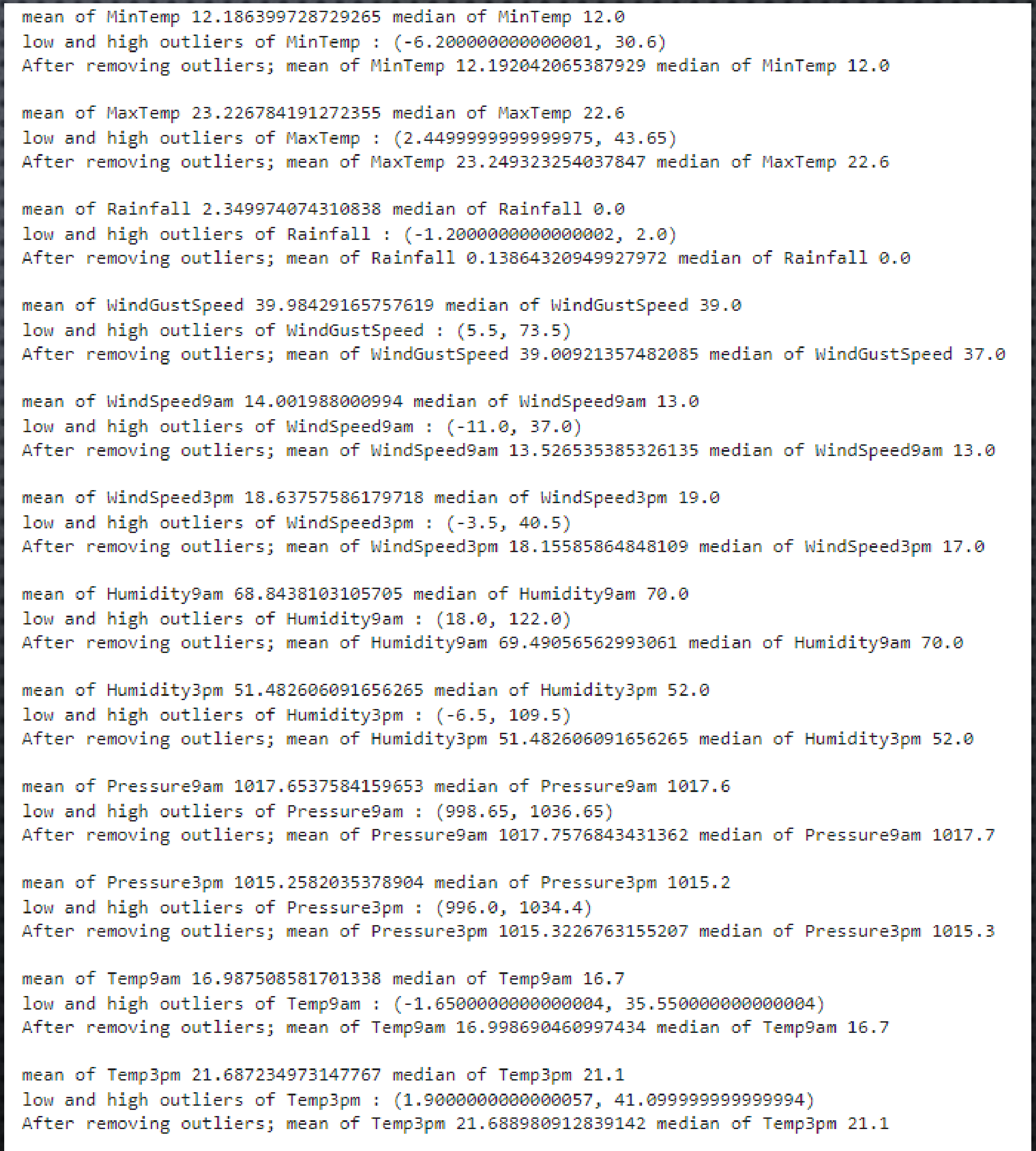
3.1. Numerical data

For each numeric columns, the average of the columns was calculated without considering outliers. Then Null values were replaced with the averages. To find outliers, Tukey method was used.

Tukey method:

The method finds the outliers based on quartiles of the data. The first quartile **Q1 >= ¼ of the data**, the second quartile **Q2 >= ½ of the data** or it is called median value. The third quartile **Q3 >= ¾ of the data**. The interquartile range IQR, is **Q3 − Q1.** According to the Tukey method, the outliers are values that are more than 1.5 times the interquartile range from the quartiles.

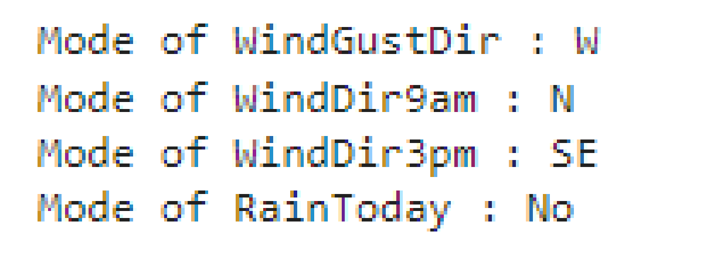
For above (high outliers): **Q3 + 1.5IQR**
For below (low outliers): **Q1 − 1.5IQR**



```
mean of MinTemp 12.186399728729265 median of MinTemp 12.0
low and high outliers of MinTemp : (-6.200000000000001, 30.6)
After removing outliers; mean of MinTemp 12.192042065387929 median of MinTemp 12.0

mean of MaxTemp 23.226784191272355 median of MaxTemp 22.6
low and high outliers of MaxTemp : (2.4499999999999975, 43.65)
After removing outliers; mean of MaxTemp 23.249323254037847 median of MaxTemp 22.6

mean of Rainfall 2.349974074310838 median of Rainfall 0.0
low and high outliers of Rainfall : (-1.2000000000000002, 2.0)
After removing outliers; mean of Rainfall 0.13864320949927972 median of Rainfall 0.0

mean of WindGustSpeed 39.98429165757619 median of WindGustSpeed 39.0
low and high outliers of WindGustSpeed : (5.5, 73.5)
After removing outliers; mean of WindGustSpeed 39.00921357482085 median of WindGustSpeed 37.0

mean of WindSpeed9am 14.001988000994 median of WindSpeed9am 13.0
low and high outliers of WindSpeed9am : (-11.0, 37.0)
After removing outliers; mean of WindSpeed9am 13.526535385326135 median of WindSpeed9am 13.0

mean of WindSpeed3pm 18.63757586179718 median of WindSpeed3pm 19.0
low and high outliers of WindSpeed3pm : (-3.5, 40.5)
After removing outliers; mean of WindSpeed3pm 18.15585864848109 median of WindSpeed3pm 17.0

mean of Humidity9am 68.8438103105705 median of Humidity9am 70.0
low and high outliers of Humidity9am : (18.0, 122.0)
After removing outliers; mean of Humidity9am 69.49056562993061 median of Humidity9am 70.0

mean of Humidity3pm 51.482606091656265 median of Humidity3pm 52.0
low and high outliers of Humidity3pm : (-6.5, 109.5)
After removing outliers; mean of Humidity3pm 51.482606091656265 median of Humidity3pm 52.0

mean of Pressure9am 1017.6537584159653 median of Pressure9am 1017.6
low and high outliers of Pressure9am : (998.65, 1036.65)
After removing outliers; mean of Pressure9am 1017.7576843431362 median of Pressure9am 1017.7

mean of Pressure3pm 1015.2582035378904 median of Pressure3pm 1015.2
low and high outliers of Pressure3pm : (996.0, 1034.4)
After removing outliers; mean of Pressure3pm 1015.3226763155207 median of Pressure3pm 1015.3

mean of Temp9am 16.987508581701338 median of Temp9am 16.7
low and high outliers of Temp9am : (-1.6500000000000004, 35.550000000000004)
After removing outliers; mean of Temp9am 16.998690460997434 median of Temp9am 16.7

mean of Temp3pm 21.687234973147767 median of Temp3pm 21.1
low and high outliers of Temp3pm : (1.9000000000000057, 41.099999999999994)
After removing outliers; mean of Temp3pm 21.688980912839142 median of Temp3pm 21.1
```

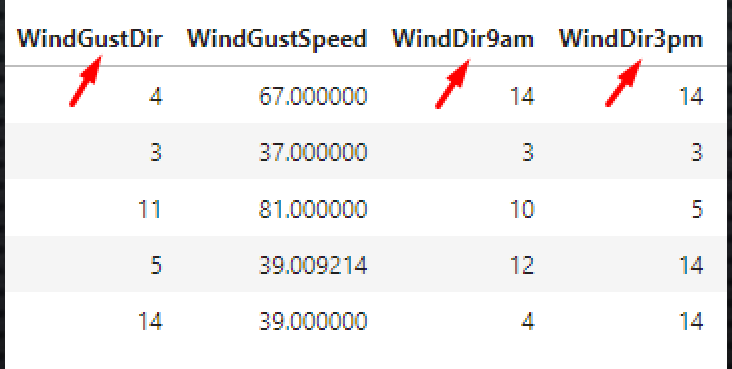
3.2. Categorical Data

For each categoric column, mode values were calculated and then Null values were replaced with mode values.

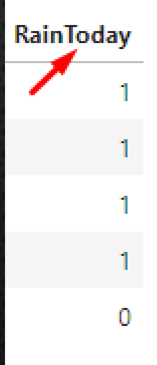


```
Mode of WindGustDir : W
Mode of WindDir9am : N
Mode of WindDir3pm : SE
Mode of RainToday : No
```

3.3. Converting categorical data to numerical data



WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm
4	67.000000	14	14
3	37.000000	3	3
11	81.000000	10	5
5	39.009214	12	14
14	39.000000	4	14

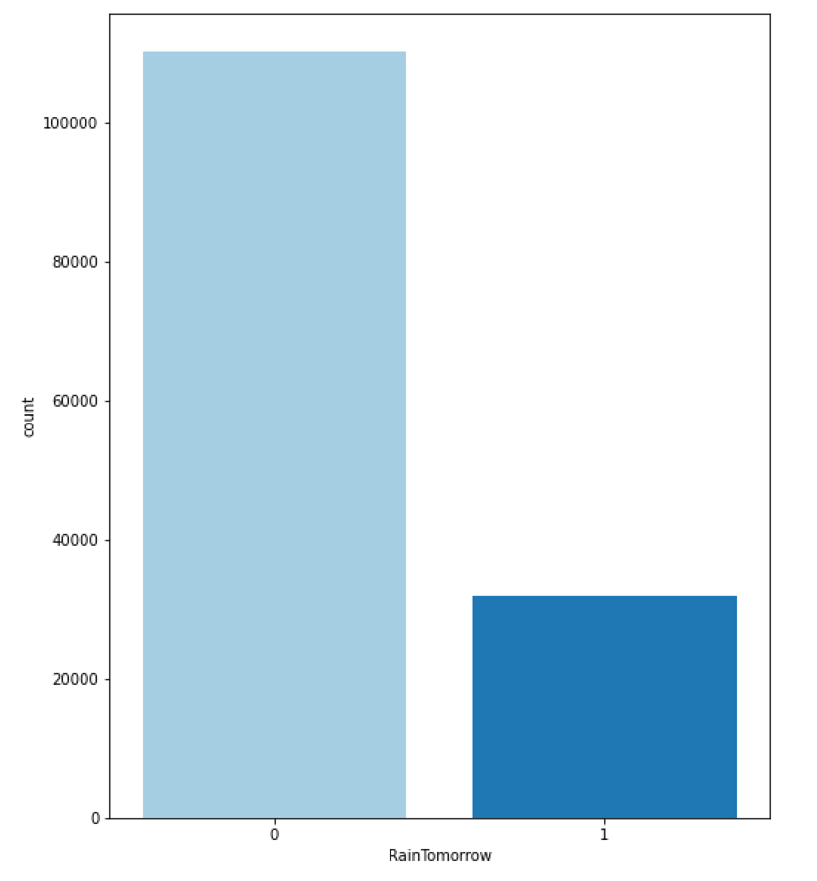


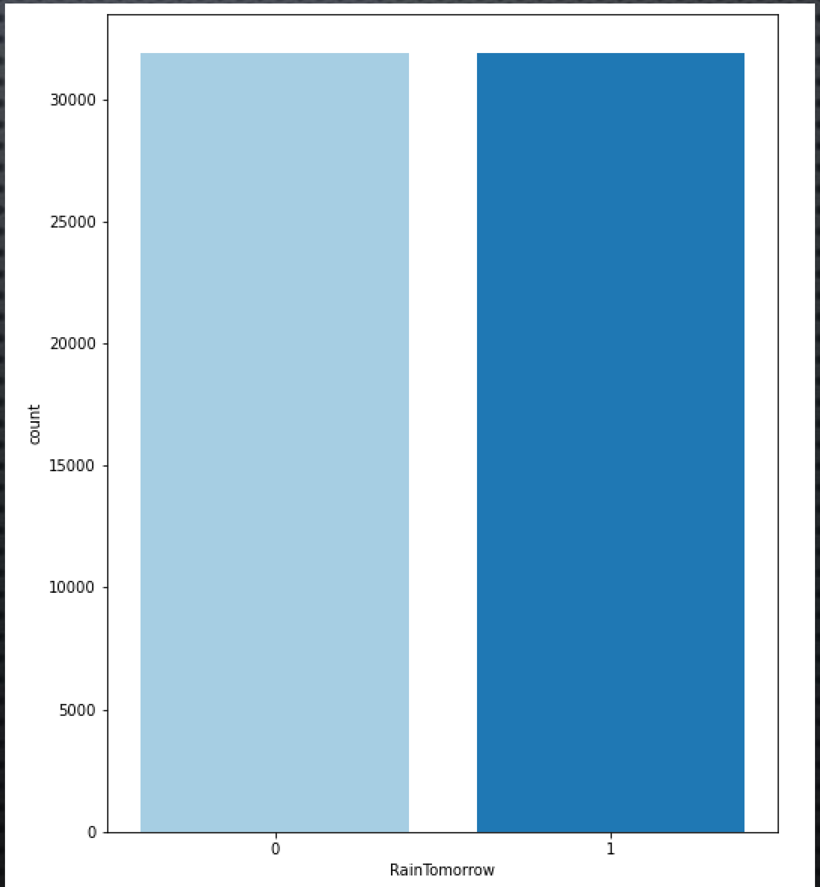
RainToday
1
1
1
1
0

4. **Balance Dataset and Visualization for RainTomorrow column**

The figure from the right side depicts that amount of “RainTomorrow” = No is very larger than the “RainTomorror” = Yes.
To get a balanced dataset, some of the rows which are “RainTomorrow” = No were deleted randomly. Because the classifier should learn information equally for both labels.

Note: 0 means No, 1 means Yes

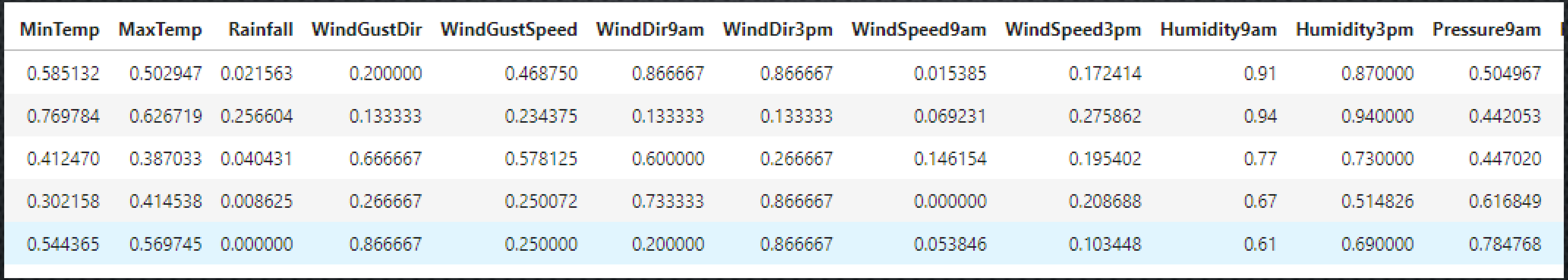




5. Feature Scaling

Min Max Scaler was used for normalization. Because the data has values with different ranges.

Mostly machine learning classifications/algorithms use Euclidean distance and some of them use weighted sum of input variables (neural networks etc.), so it is necessary to normalize the data. Because the data can be in different units and scales. If one data is larger than the other and if normalization is not done, that data will affect the result more. However, each data should contribute equally to the result, so normalization should be done.

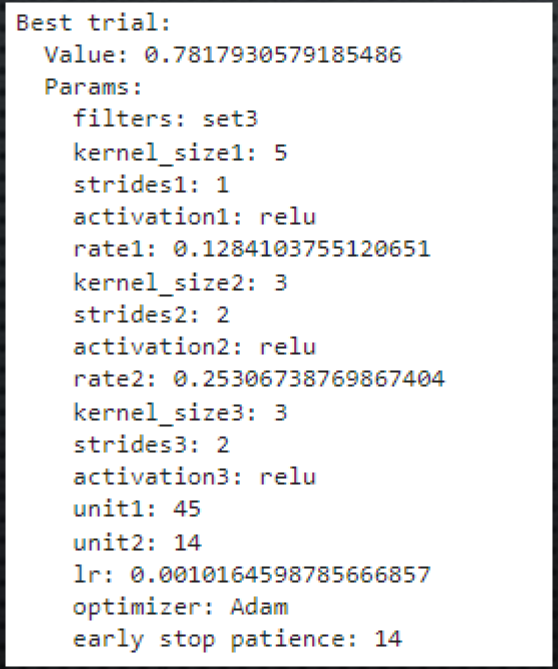


MinTemp	MaxTemp	Rainfall	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am
0.585132	0.502947	0.021563	0.200000	0.468750	0.866667	0.866667	0.015385	0.172414	0.91	0.870000	0.504967
0.769784	0.626719	0.256604	0.133333	0.234375	0.133333	0.133333	0.069231	0.275862	0.94	0.940000	0.442053
0.412470	0.387033	0.040431	0.666667	0.578125	0.600000	0.266667	0.146154	0.195402	0.77	0.730000	0.447020
0.302158	0.414538	0.008625	0.266667	0.250072	0.733333	0.866667	0.000000	0.208688	0.67	0.514826	0.616849
0.544365	0.569745	0.000000	0.866667	0.250000	0.200000	0.866667	0.053846	0.103448	0.61	0.690000	0.784768

1D CNN TRAINING WITH OPTUNA

1 Dimensional Convolutional Neural Network were used for the binary classification task because CNN has filters to extract information from the raw data for learning. In this way, training is done with smaller size information which is called features obtained from the data.

Optuna which is a auto machine learning tool, was used for optimization the hyperparameters. Different amount of layers, different amount of filters (nodes) and filter sizes of 1D CNN were tried with Optuna then the best model was saved.



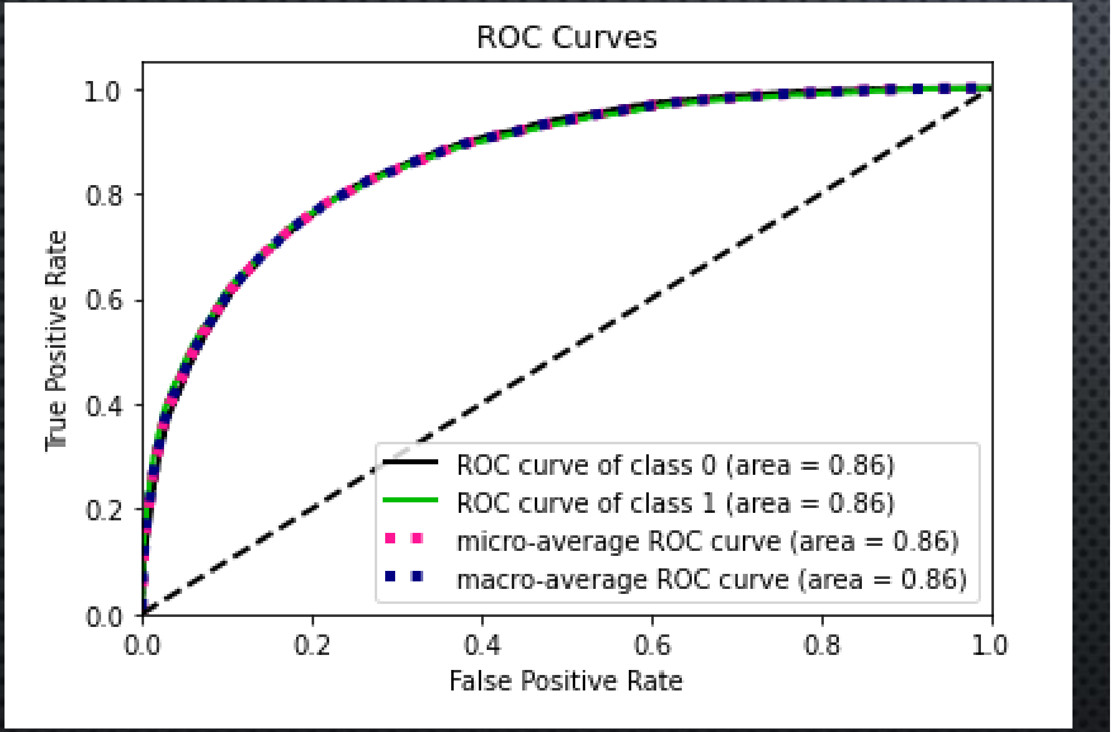
```
Best trial:
  Value: 0.7817930579185486
  Params:
    filters: set3
    kernel_size1: 5
    strides1: 1
    activation1: relu
    rate1: 0.1284103755120651
    kernel_size2: 3
    strides2: 2
    activation2: relu
    rate2: 0.25306738769867404
    kernel_size3: 3
    strides3: 2
    activation3: relu
    unit1: 45
    unit2: 14
    lr: 0.0010164598785666857
    optimizer: Adam
    early stop patience: 14
```

Optuna found the best trial which gave the 0.78 accuracy on test dataset (%25 of the data).

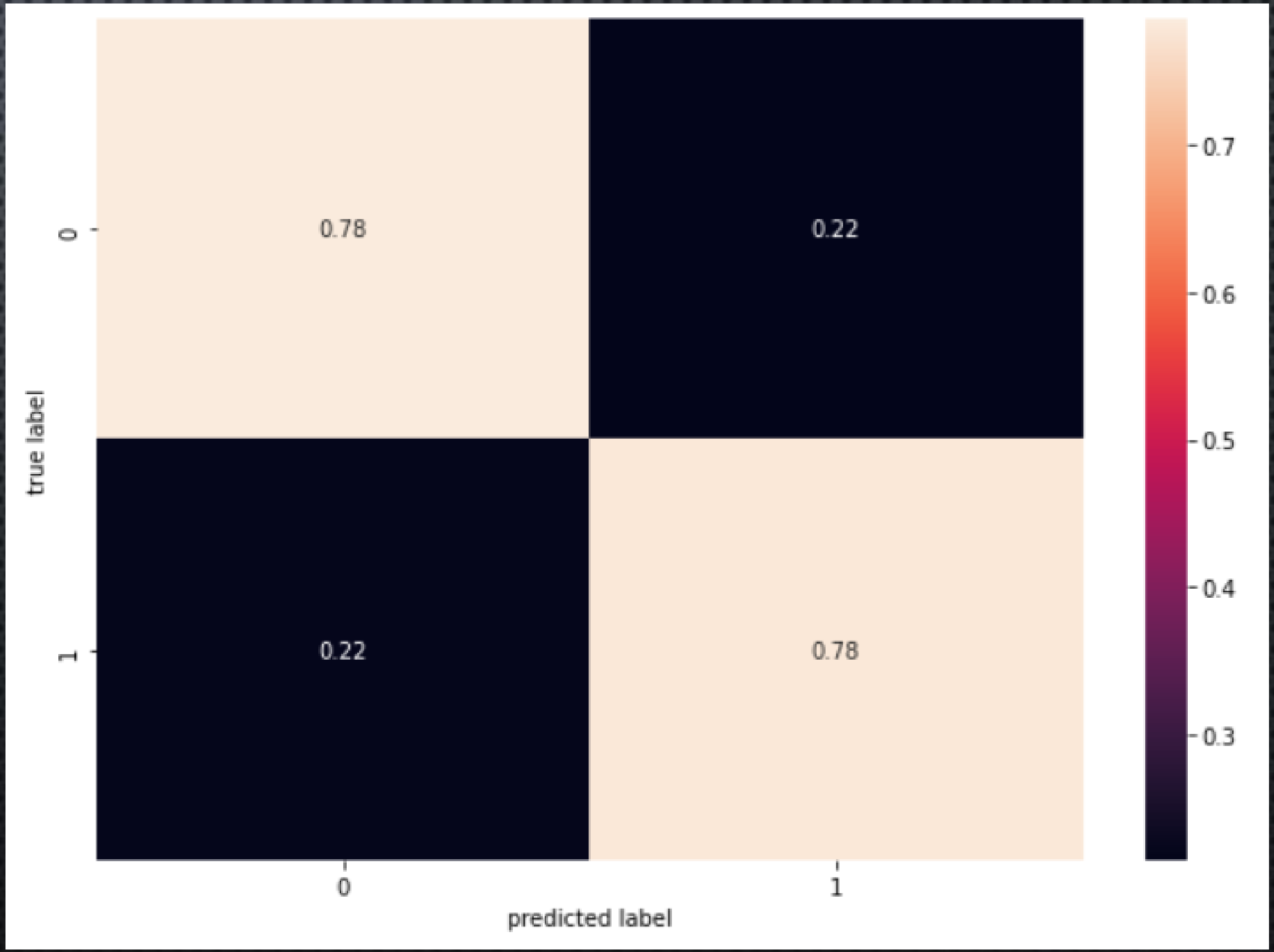
1. Results of 1D CNN Training

0.78 accuracy on test dataset

Receiver operating characteristic



Confusion Matrix



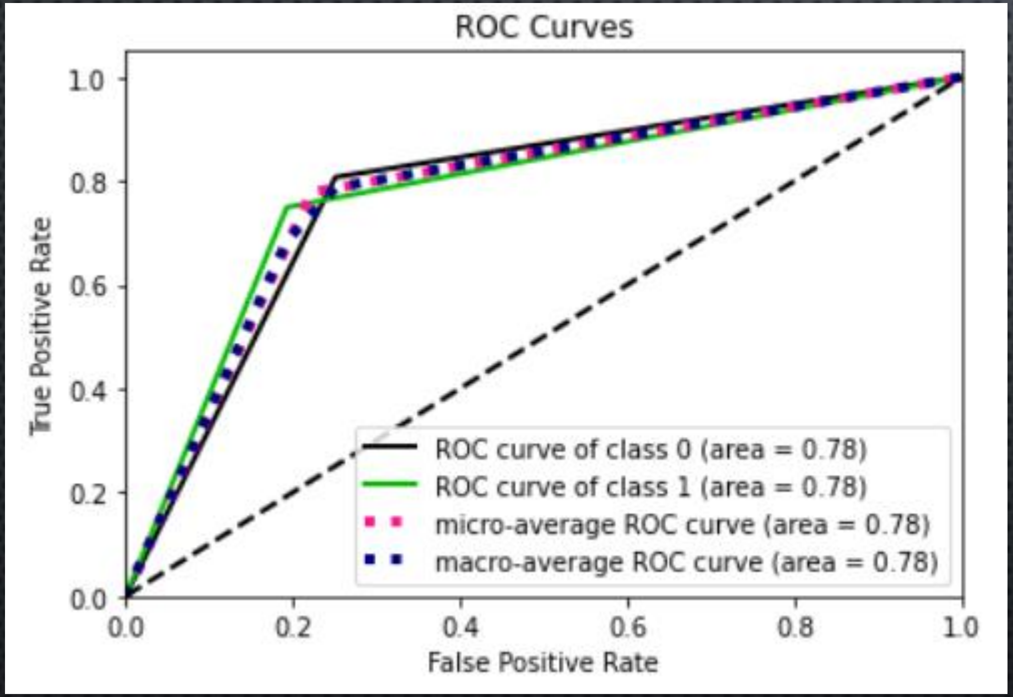
MACHINE LEARNING ALGORITHMS TRAINING WITH OPTUNA

For the binary classification task, Logistic Regression and SVM were chosen, and using Optuna, the best classifier with the best parameters has been chosen. After searching for the best results, the model is saved. This part is done for comparison with 1D CNN results. The result is %78 accuracy on the test dataset with SVM (‘C': 80.30161564418209, 'kernel': 'rbf', 'gamma': 0.6791700860838945).

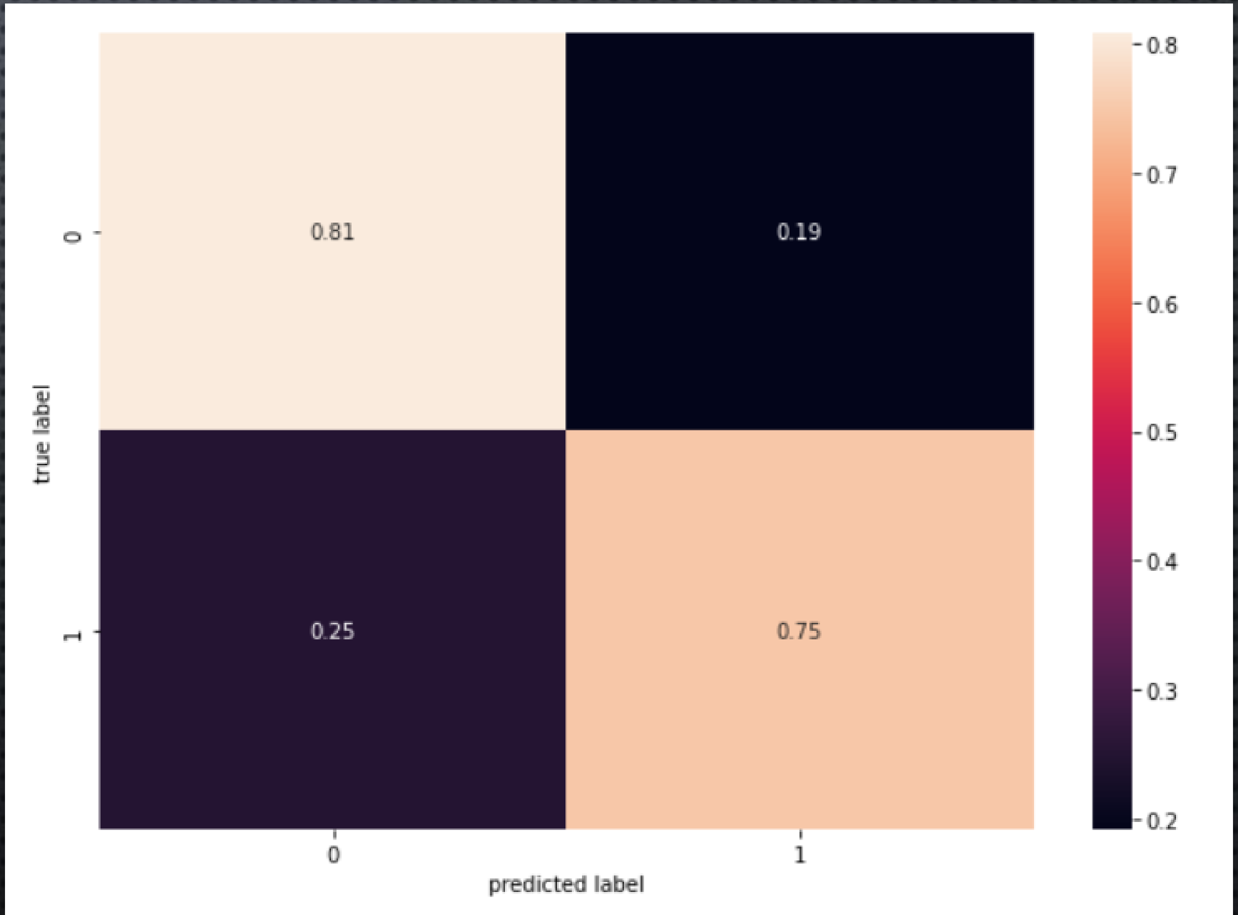
1. Results of SVM Training

%78 accuracy on test dataset

Receiver operating characteristic



Confusion Matrix



COMPARISON OF THE RESULTS

If the classifier has a larger area under the curve (ROC curve), that means the classier is better. In these 2 curves above the previous slides, It can be seen that 1D CNN reached better results.

Both classifiers have the same accuracy on the test dataset but If we compare the confusion matrix of these two classifiers, It can be seen that the error rate is the same in both classes on 1D CNN results. The other classifier has different error rates on the labels which lead to predicting “YES” with lower accuracy than CNN. That means 1D CNN results are more balanced and reached better results.

CONCLUSION

As a result, 1D CNN slightly outperformed machine learning algorithms for this task. For future work, depending on which column has more outliers, some rows can be deleted, and the models can be retrained to look at the results. More data can be collected and training results can be improved, the last dataset with newly added data should contain equal amounts of both labels.