Machine Learning Fundamental – Final Project

Predictive Forecast Weather for Singapore
(Kind of rain)

Professor: Ali El-Sharif

TA: Nail Senbas

Student: Mehrad Tavanamehr

CONTEMPORARY TECHNOLOGY UNIVERSITY

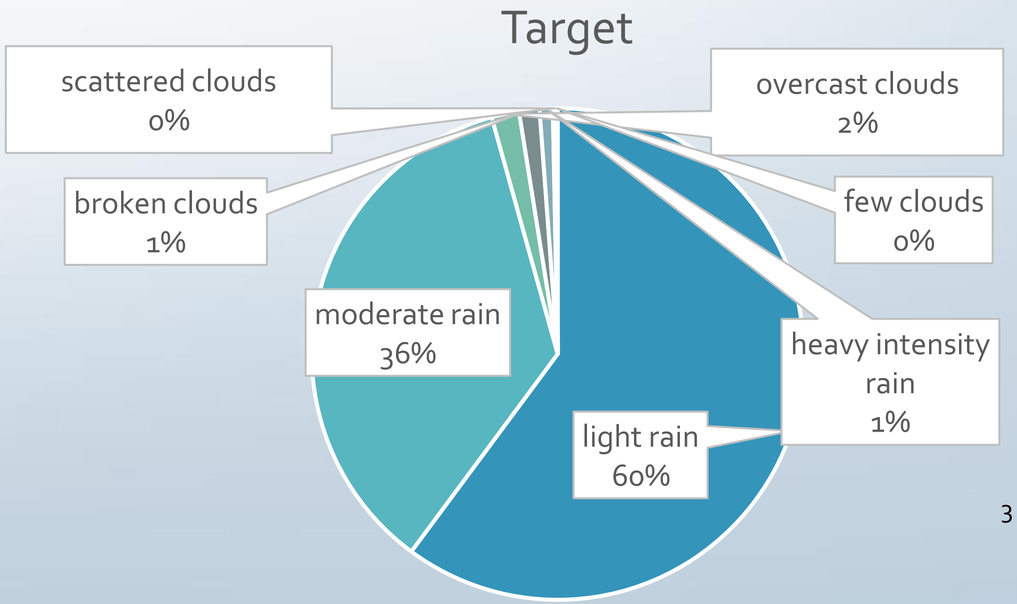
Introduction

One of the main issues for people living in south-east Asia is the rainy weather, and sometimes the intensity of rain is bothering people, according to the dataset I use for this project, I have access to different measurements to predict a variety of specific situations of rainy and cloudy weather.

Introduction

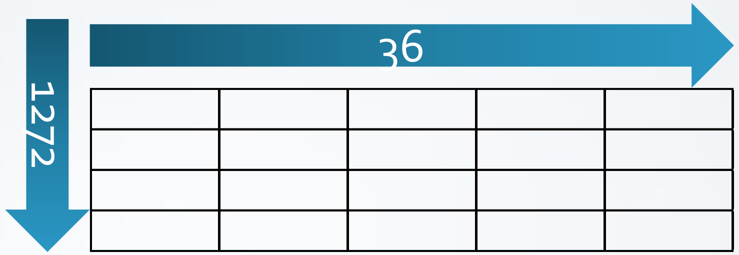
The target value I face is unbalanced so I change the target to the prediction of the intensity of rain and focus on the light rain from other kinds of rain, to make the output balance.

By predicting the intensity of rain, I can help people in that area be ready for the intensity of rain, increase public transportation safety, and control traffic.



Categorical

- Timezone
- weather_main
- weather_description
- weather_icon
- location



Numerical

lat lon	dt Sunrise Sunset Moonrise Moonset moon_phase	Pressure Humidity dew_point	wind_speed wind_deg wind_gust	Pop Rain Uvi	temp_day temp_min temp_max temp_night temp_eve temp_morn	feels_like_day feels_like_night feels_like_eve feels_like_morn

Data Cleaning

- Based on dataset we discussed, there are not a lot to do, just drop feature with unique values

```python
[7] #drop feature with unique values
    unique_val_cols = []
    for cols in dataset.columns:
      if dataset[cols].nunique() == 1:
        unique_val_cols.append(cols)
    dataset.drop(columns = unique_val_cols, inplace = True)
```

Data manipulation

- Fillna of rain feature with the constant value
- Creating two new features instead of ‘sunrise’, ‘sunset’, ’moonrise’, & ‘moonset’

Features contain Null values

- rain

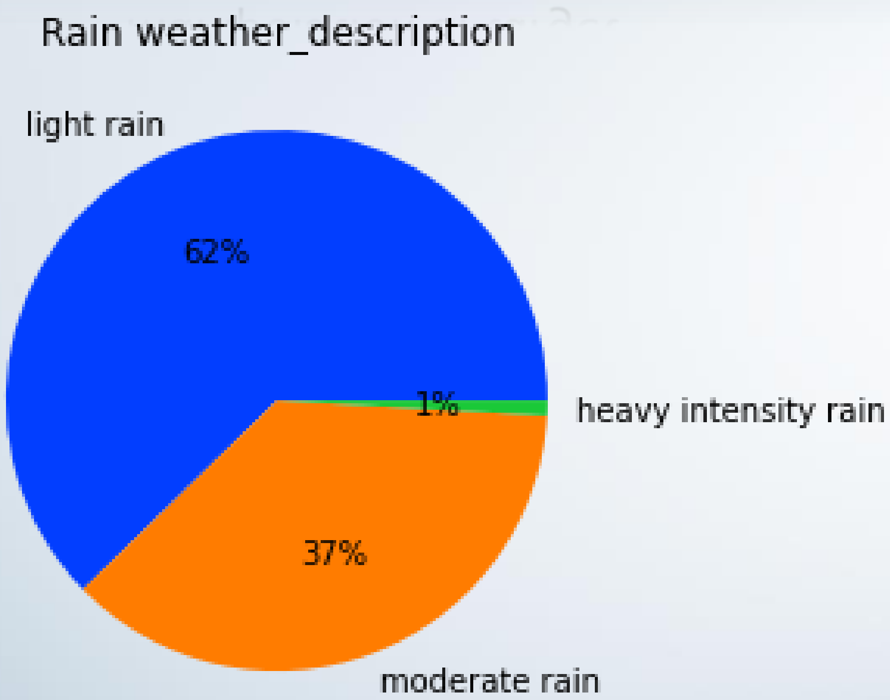
Creating two new features based on correct one

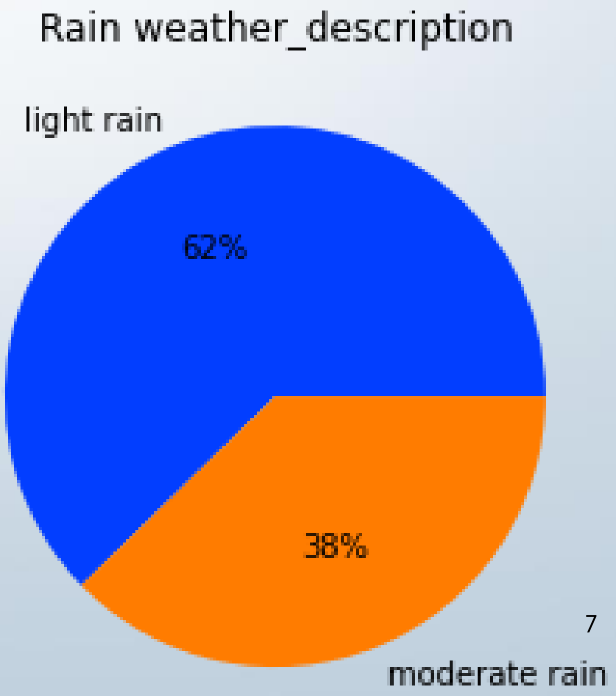
- dataset['day_duration'] = dataset.apply(lambda x: dataset.sunset-dataset.sunrise).mean(axis=1)
- dataset['night_duration'] = dataset.apply(lambda x: dataset.moonrise-dataset.moonset).mean(axis=1)

Project Steps

Data manipulation

- Manipulation of target



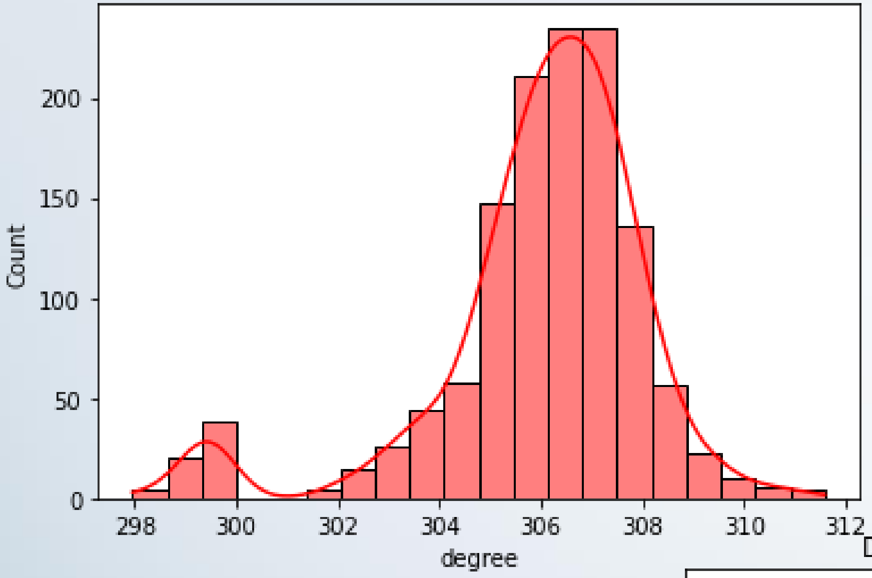



```
weather_description
light rain        500.0
moderate rain     501.0
Name: weather_id, dtype: float64
```

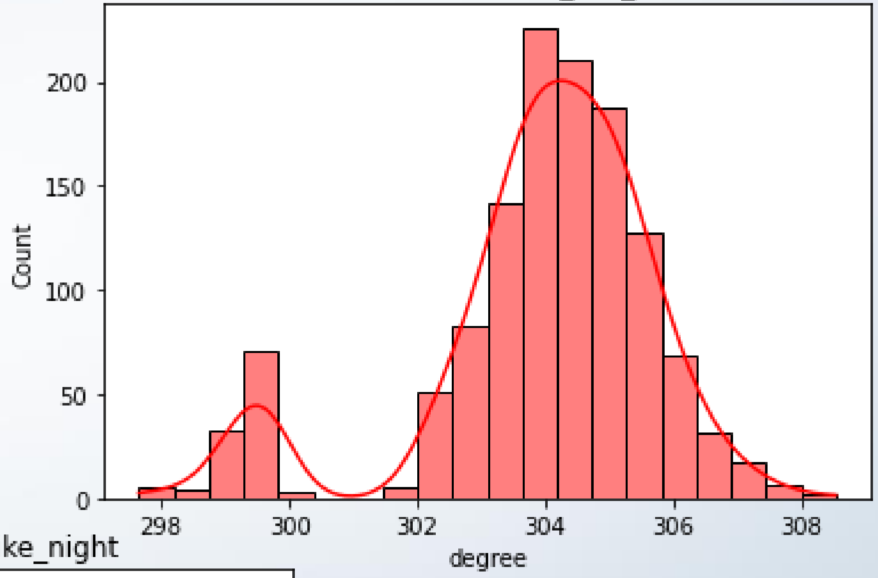
Data Analyze

- Check the distribution of (feels_like_*) features

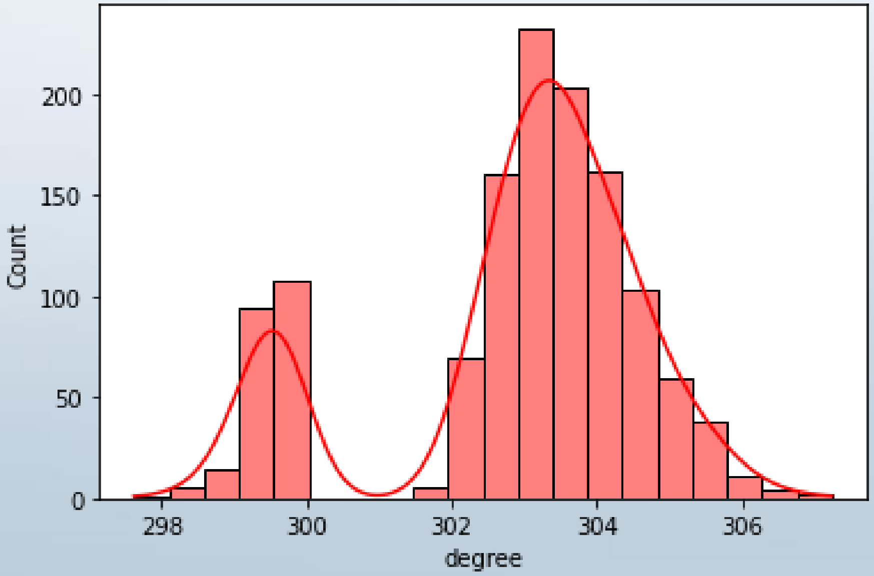
Distribution of feels_like_day



Distribution of feels_like_eve



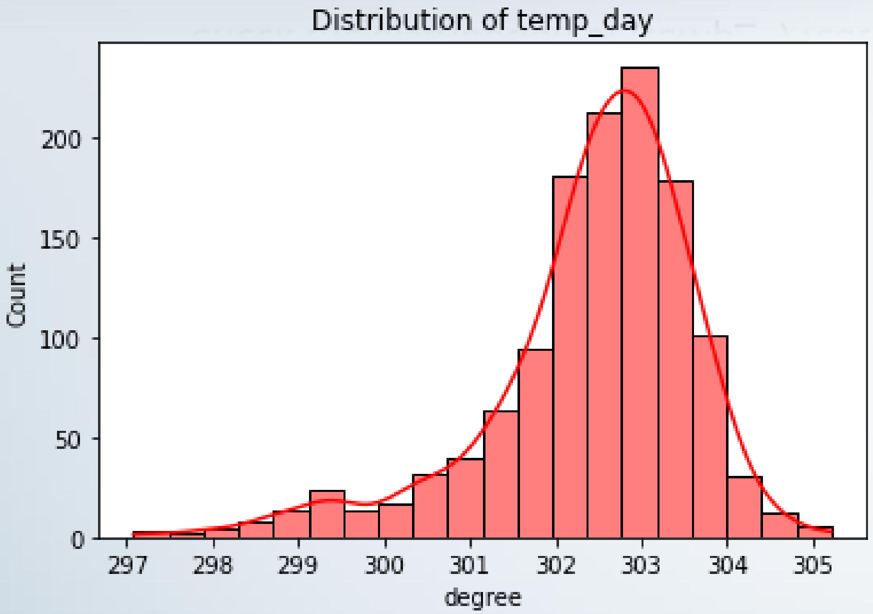
Distribution of feels_like_night

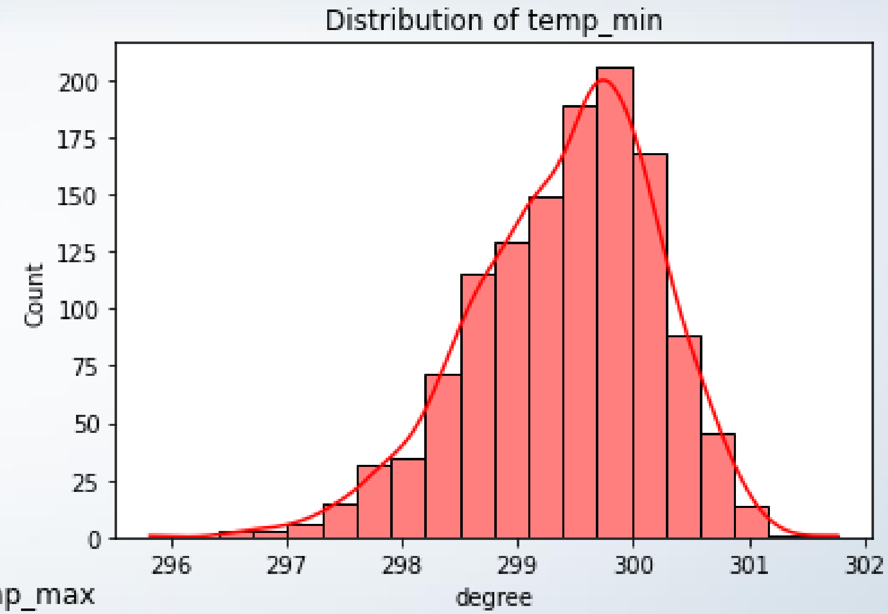


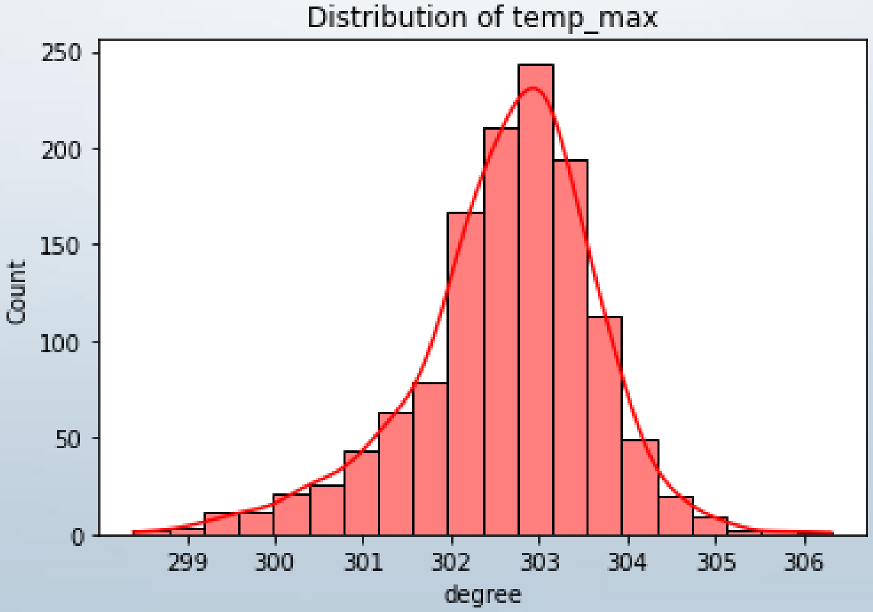
Project Steps

Data Analyze

- Check the distribution of (temp_*) features – P1



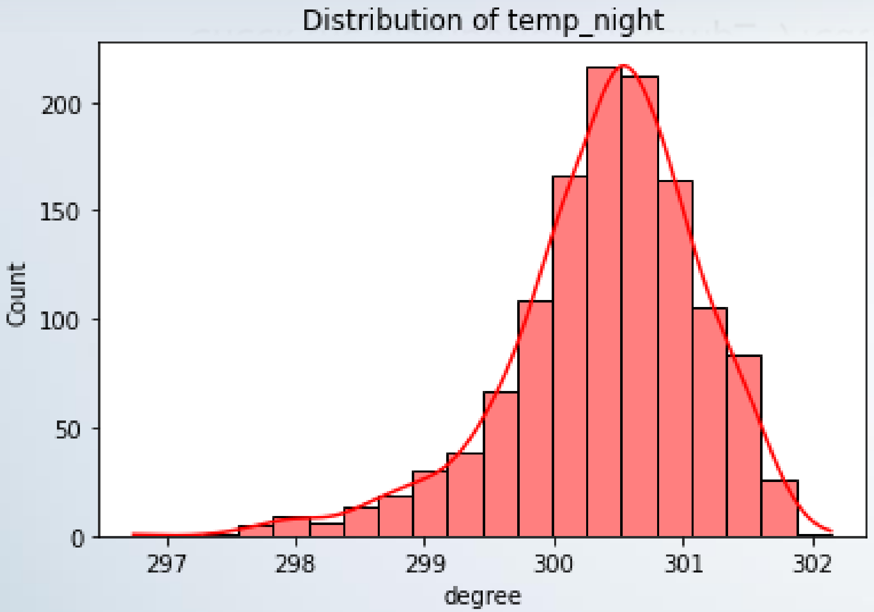


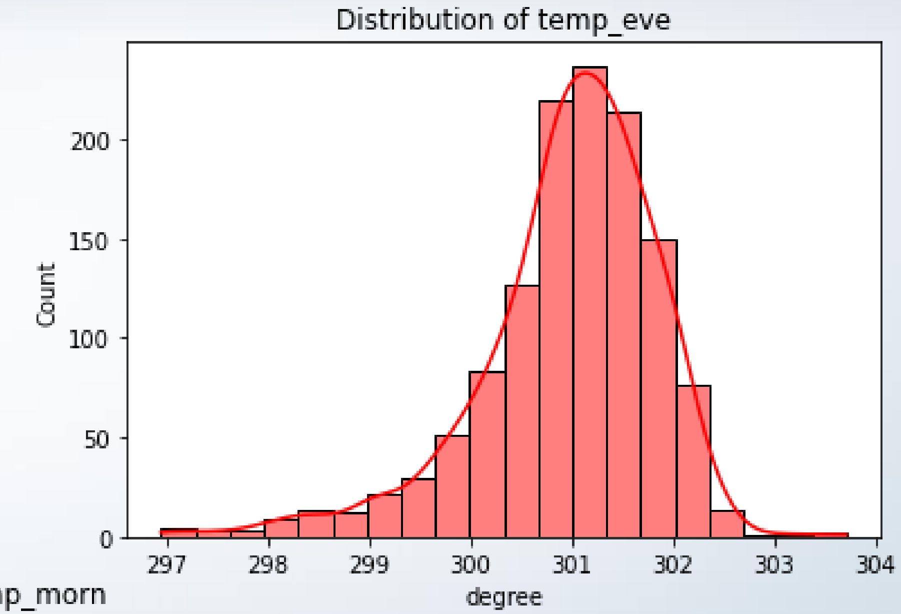


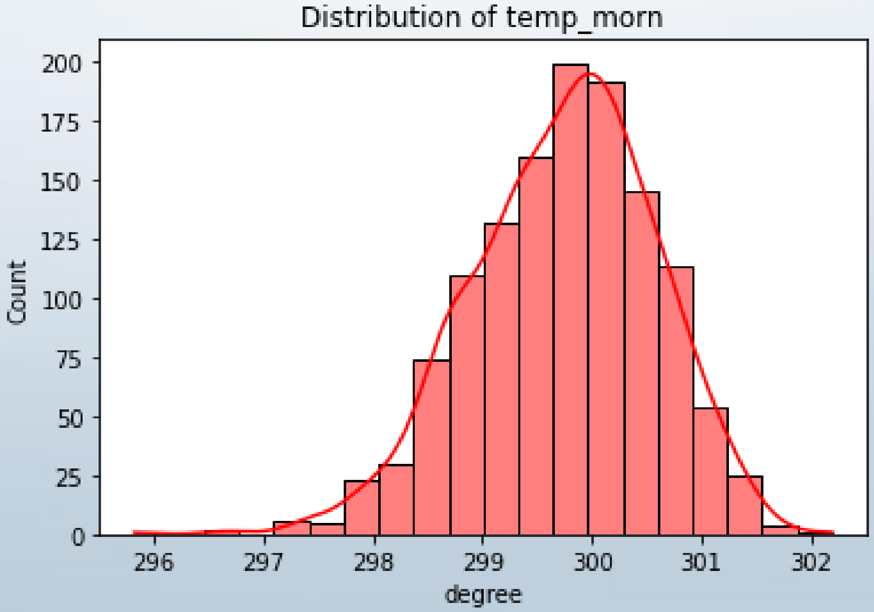
Project Steps

Data Analyze

- Check the distribution of (temp_*) features P2



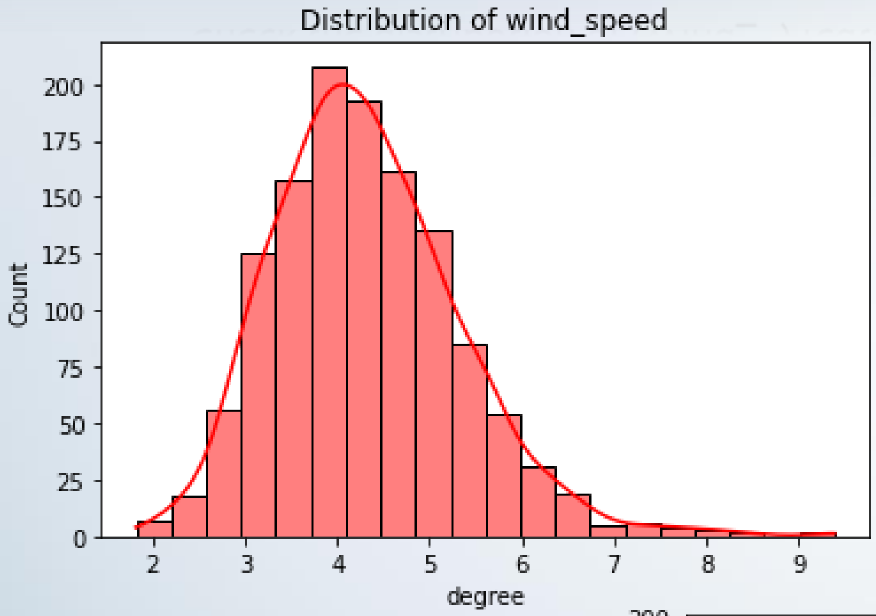


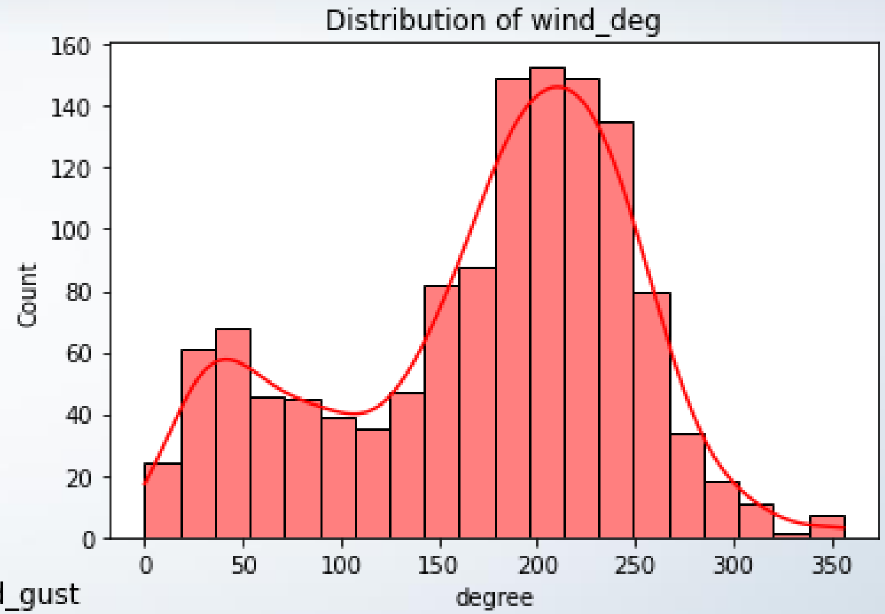


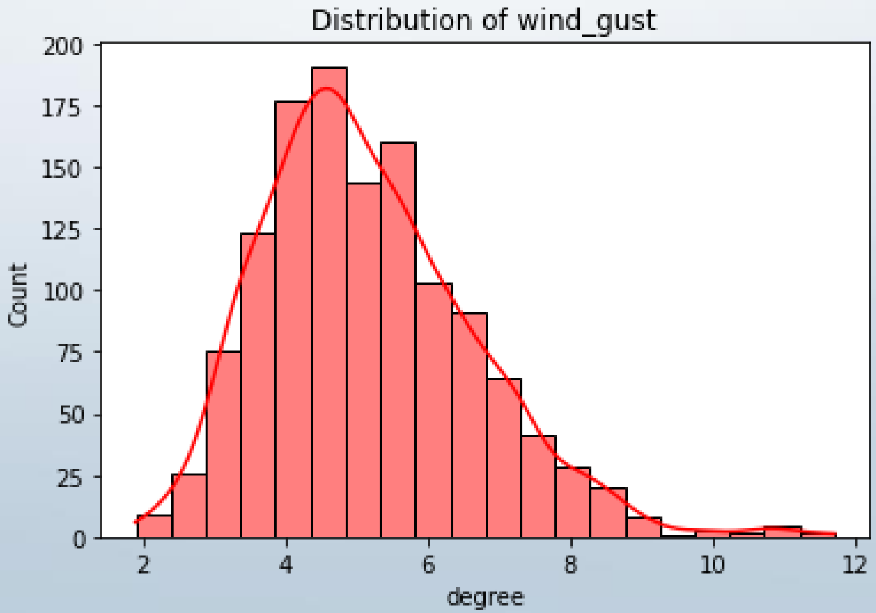
Project Steps

Data Analyze

- Check the distribution of (Wind_*) features

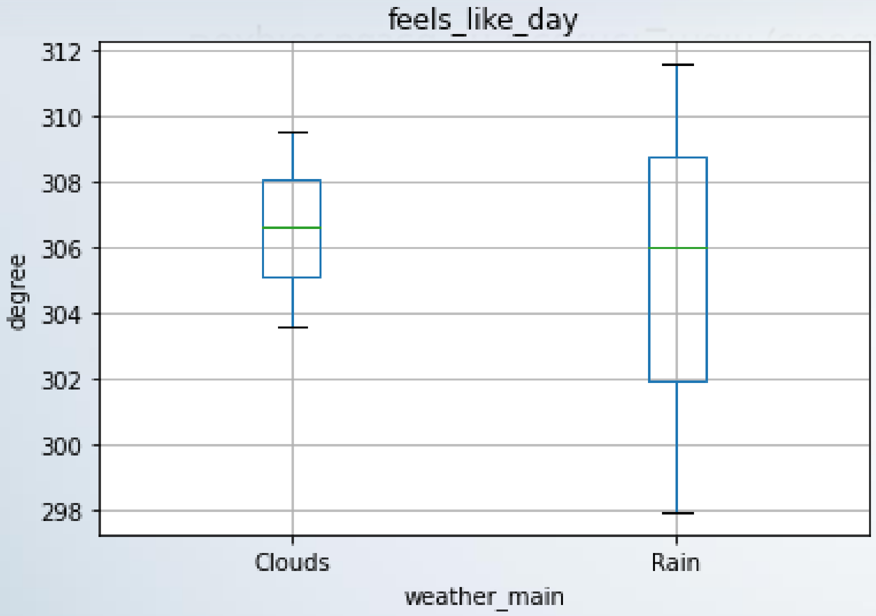


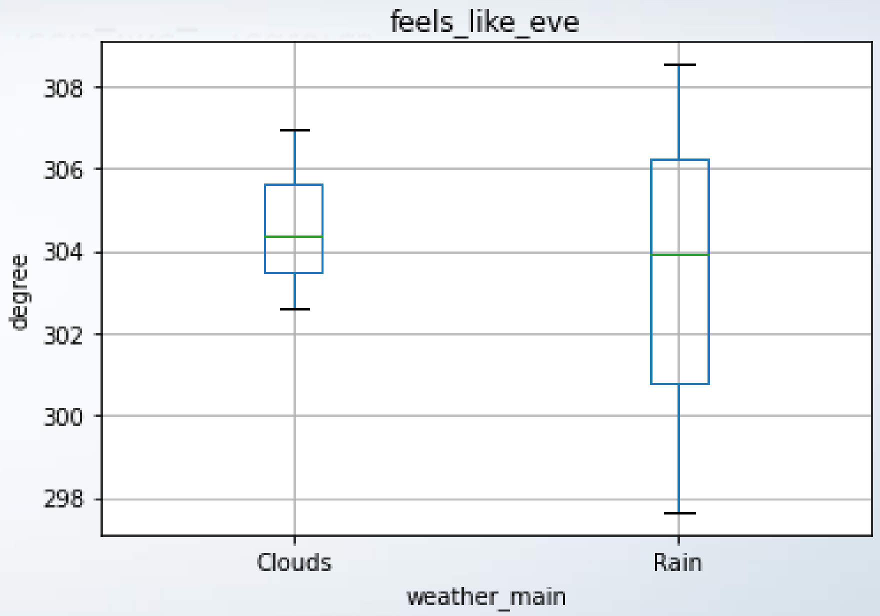


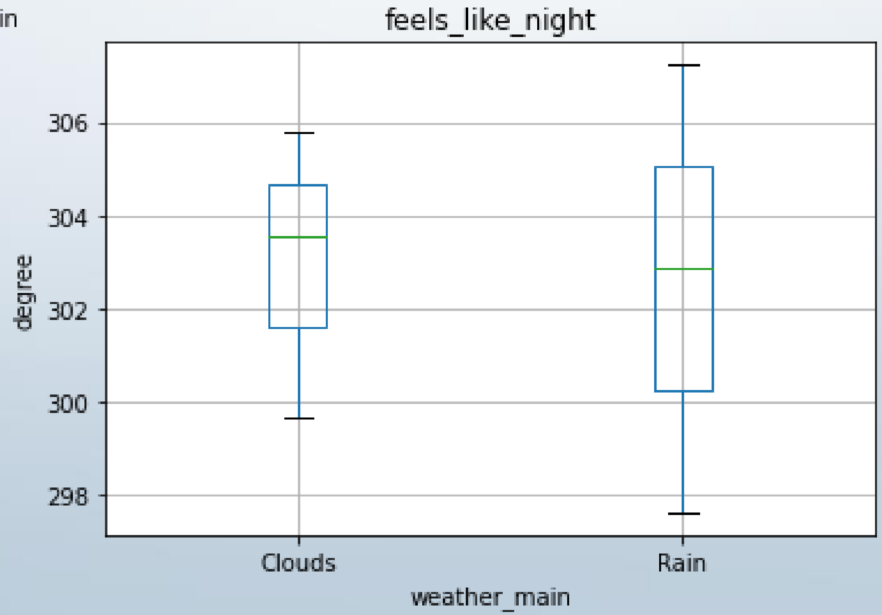


Data Analyze

- Boxplot based on weather_main (clouds, & rain) field for feels_like_* features



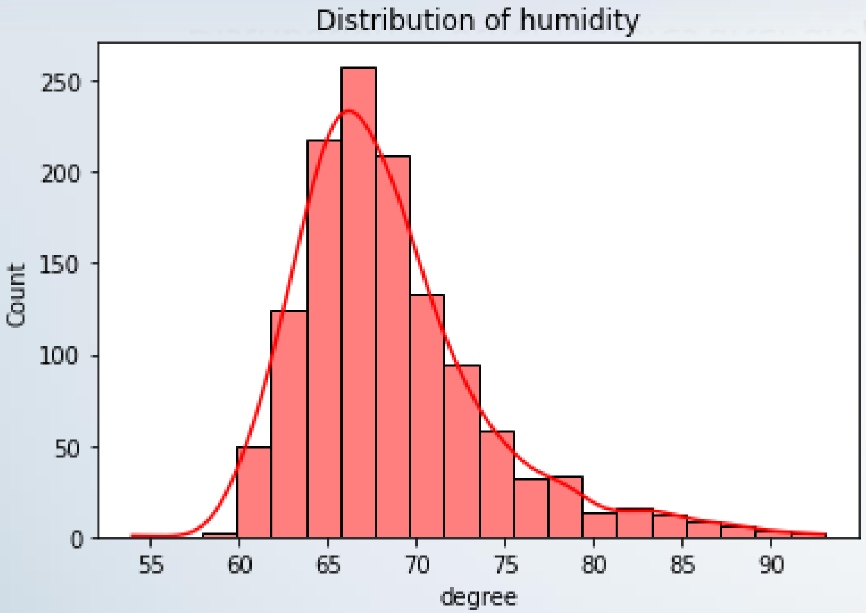


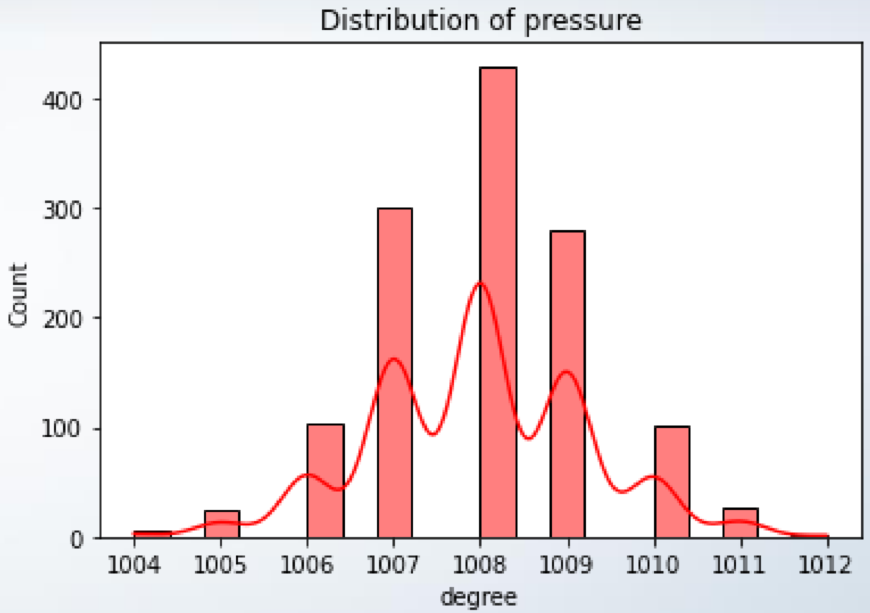


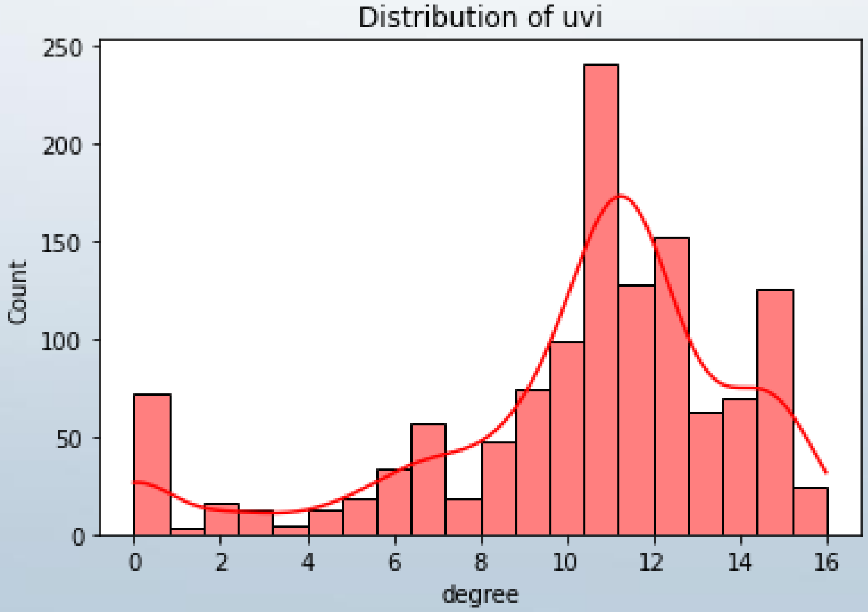
Project Steps

Data Analyze

- Distribution of other features after drop clouds samples

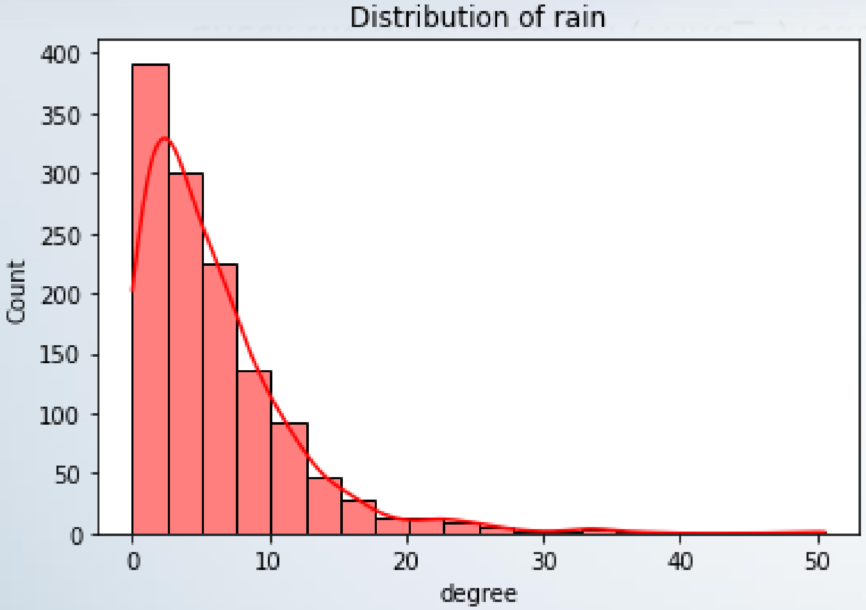


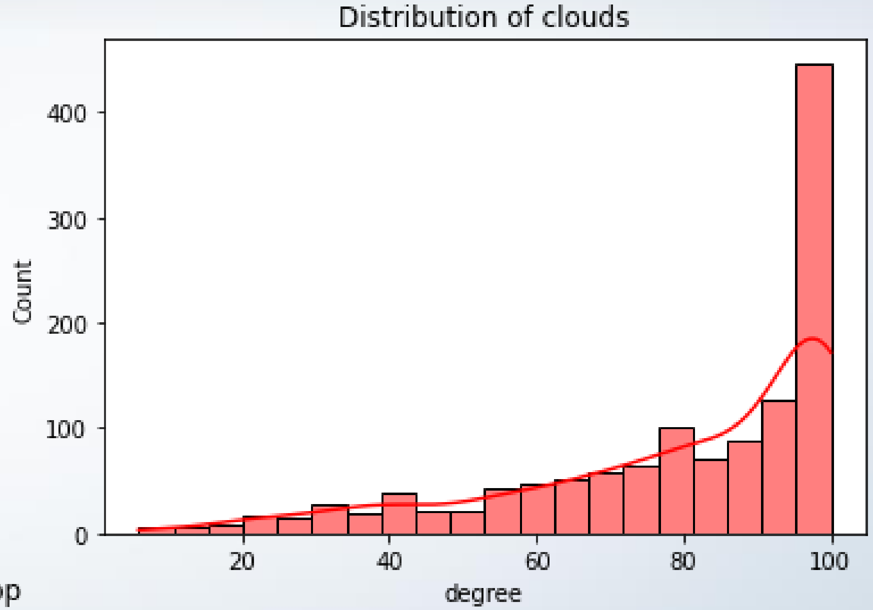


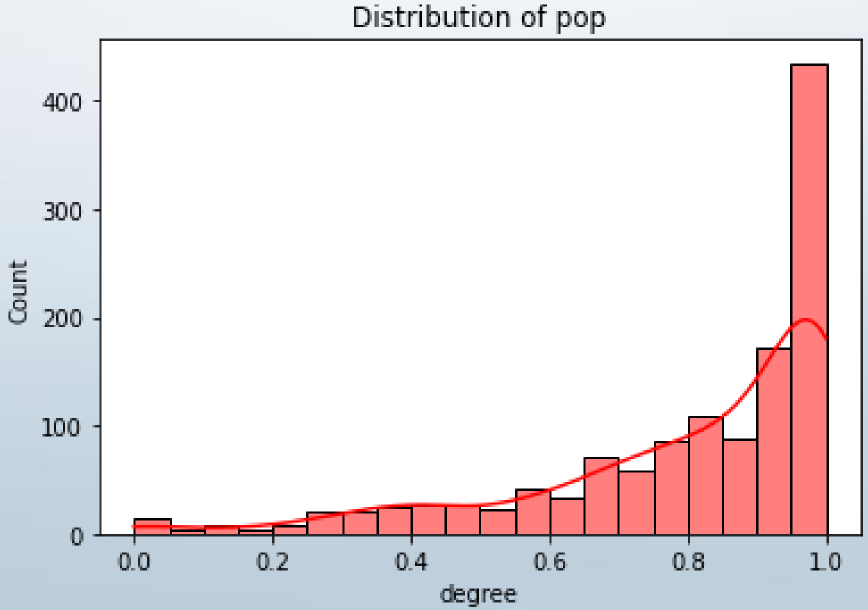


Data Analyze

- Check the distribution of (Wind_*) features





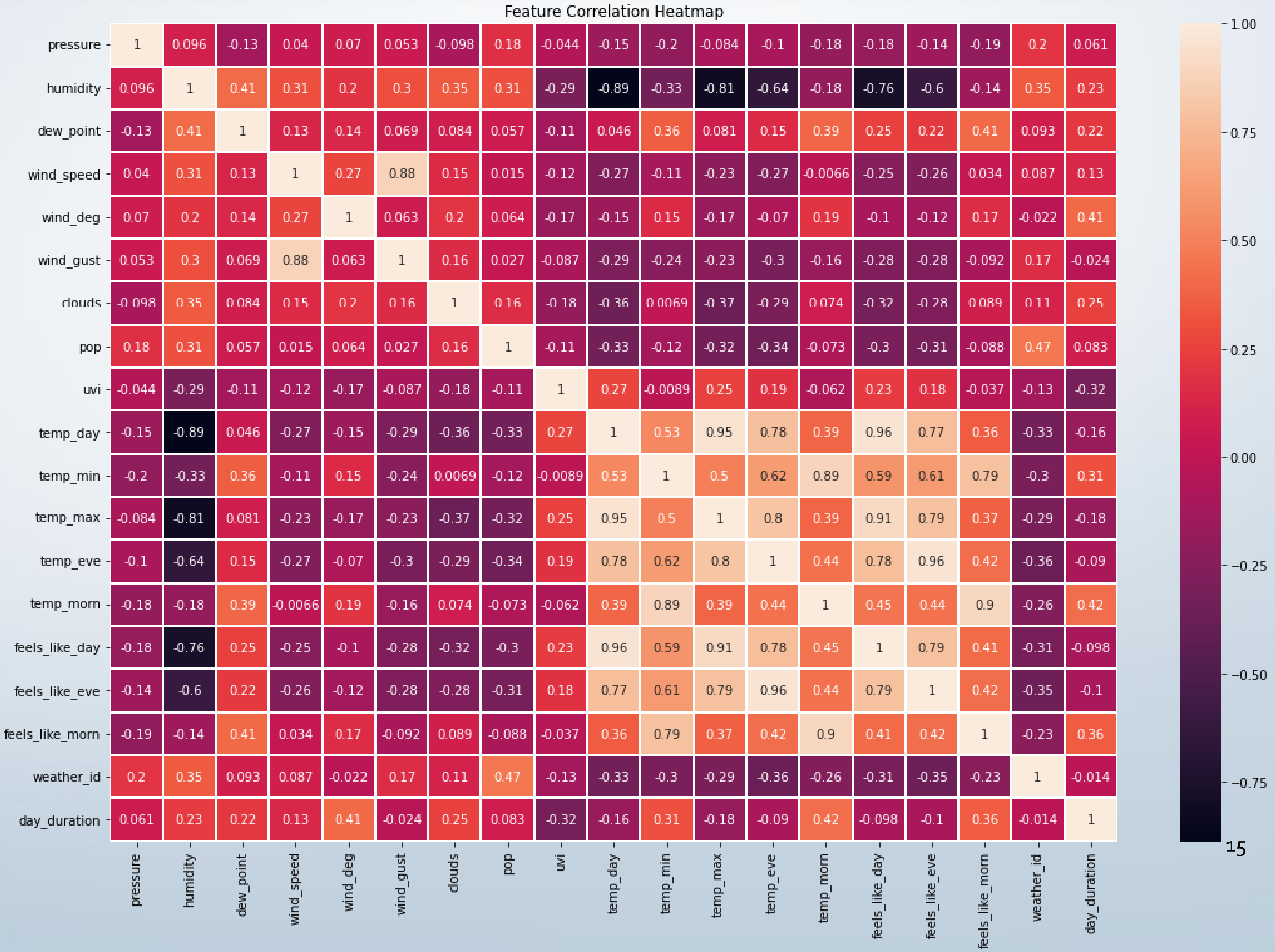


Project Steps

Data Analyze

- Heatmap graph

The amount of rain is related to after occurring so it is dropped. My assumption was the situation happened during the day so all information related to night are dropped.



Project Steps

Data Analyze

- Use Hypothesis test for checking the normal distribution

Features_name	p_value	Hypothesis_Description
pressure	0.442198	Normal Distribution
humidity	0.0	We can't reject Hypothesis test
dew_point	0.000001	We can't reject Hypothesis test
wind_speed	0.0	We can't reject Hypothesis test
wind_deg	0.0	We can't reject Hypothesis test
wind_gust	0.0	We can't reject Hypothesis test
clouds	0.0	We can't reject Hypothesis test
pop	0.0	We can't reject Hypothesis test
uvi	0.0	We can't reject Hypothesis test
temp_day	0.0	We can't reject Hypothesis test
temp_min	0.0	We can't reject Hypothesis test
temp_max	0.0	We can't reject Hypothesis test
temp_eve	0.0	We can't reject Hypothesis test
temp_morn	0.000001	We can't reject Hypothesis test
feels_like_day	0.0	We can't reject Hypothesis test
feels_like_eve	0.0	We can't reject Hypothesis test
feels_like_morn	0.0	We can't reject Hypothesis test
weather_id	0.0	We can't reject Hypothesis test
day_duration	0.0	We can't reject Hypothesis test

Project Steps

Feature selection

- Dependency with target

Hypothesis test for linear dependency

Features_name	p_value	linear_dependency_Description
pressure	0.0	Probably dependent
humidity	0.0	Probably dependent
dew_point	0.001095	Probably dependent
wind_speed	0.002269	Probably dependent
wind_deg	0.431777	Probably independent
wind_gust	0.0	Probably dependent
clouds	0.000085	Probably dependent
pop	0.0	Probably dependent
uvi	0.000008	Probably dependent
temp_day	0.0	Probably dependent
temp_min	0.0	Probably dependent
temp_max	0.0	Probably dependent
temp_eve	0.0	Probably dependent
temp_morn	0.0	Probably dependent
feels_like_day	0.0	Probably dependent
feels_like_eve	0.0	Probably dependent
feels_like_morn	0.0	Probably dependent
weather_id	0.0	Probably dependent
day_duration	0.613735	Probably independent

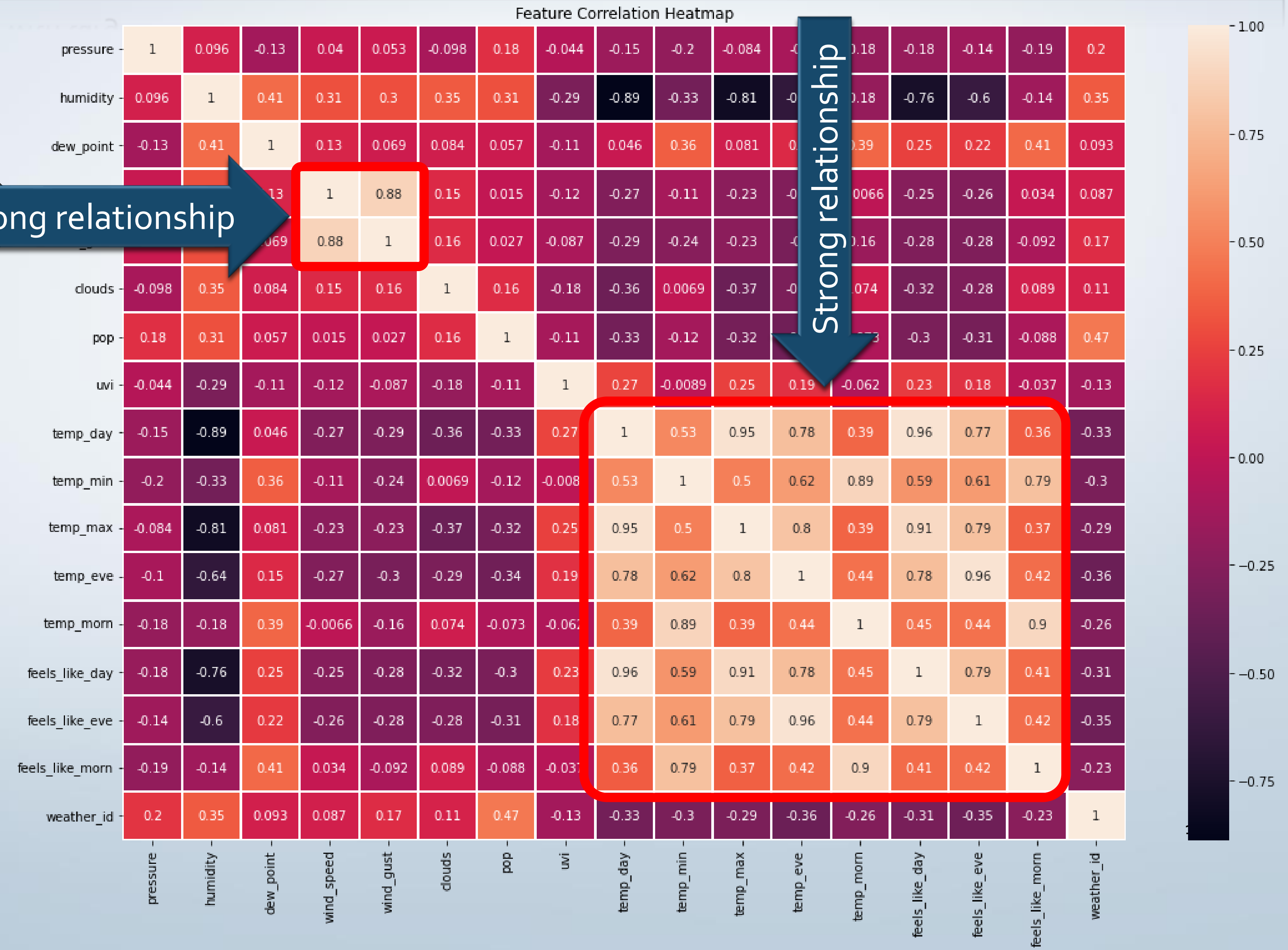
Hypothesis test for monotonic dependency

Features_name	p_value	monotonic_dependancy_Description
pressure	0.0	Probably dependent
humidity	0.0	Probably dependent
dew_point	0.000778	Probably dependent
wind_speed	0.075576	Probably independent
wind_deg	0.946618	Probably independent
wind_gust	0.0	Probably dependent
clouds	0.000002	Probably dependent
pop	0.0	Probably dependent
uvi	0.00007	Probably dependent
temp_day	0.0	Probably dependent
temp_min	0.0	Probably dependent
temp_max	0.0	Probably dependent
temp_eve	0.0	Probably dependent
temp_morn	0.0	Probably dependent
feels_like_day	0.0	Probably dependent
feels_like_eve	0.0	Probably dependent
feels_like_morn	0.0	Probably dependent
weather_id	0.0	Probably dependent
day_duration	0.817769	Probably independent

Project Steps

Feature selection

- Dependency with target



- four features: 'humidity','feels_like_day','temp_max','temp_day'
- two features: 'temp_morn','feels_like_morn'
- two features: 'wind_speed','wind_gust'

Has strong relation to each other and can bais learning algorithm

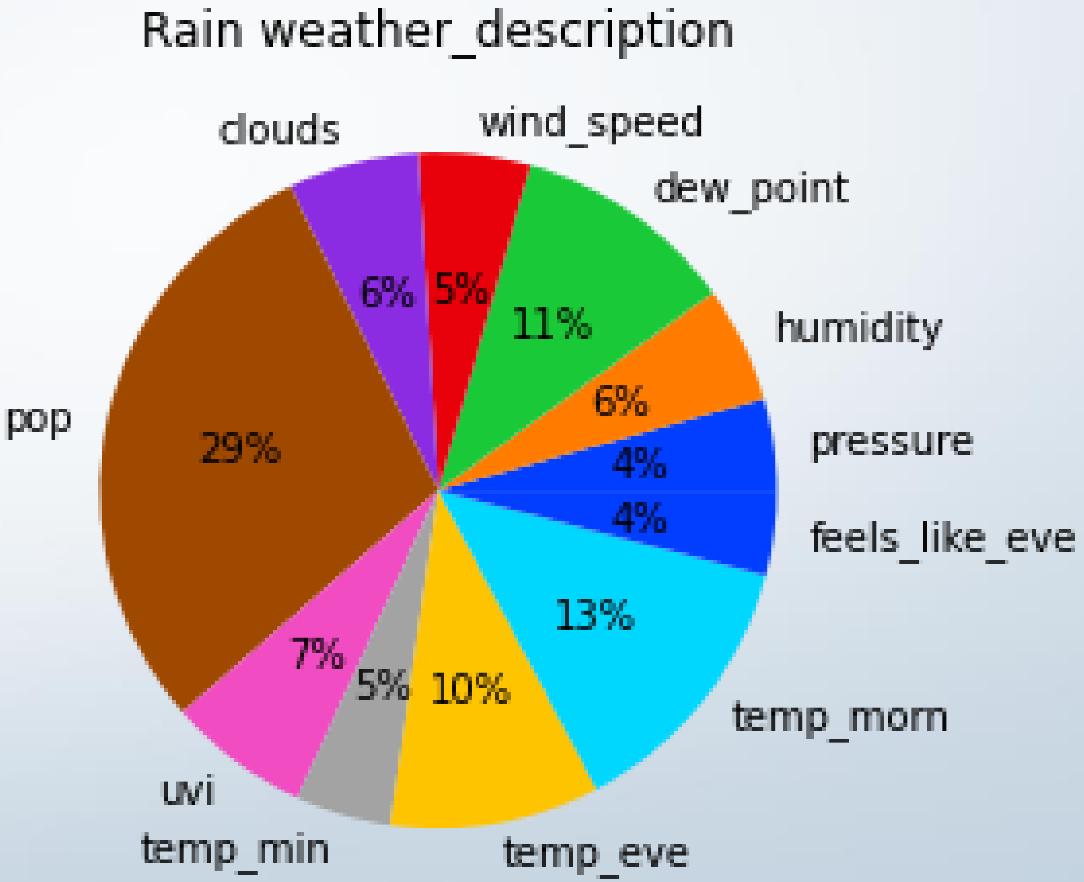
Feature selection

- Feature Importance with DecisionTree Algorithms

Important Features Selection without scaler

```
DecisionTreeClassifier()
```

	feature	importance
5	pop	0.296312
9	temp_morn	0.129936
2	dew_point	0.113511
8	temp_eve	0.106330
4	clouds	0.066606
7	temp_min	0.056014
6	uvi	0.052007
3	wind_speed	0.050781
10	feels_like_eve	0.049048
1	humidity	0.046357
0	pressure	0.033098



Project Steps

Feature selection

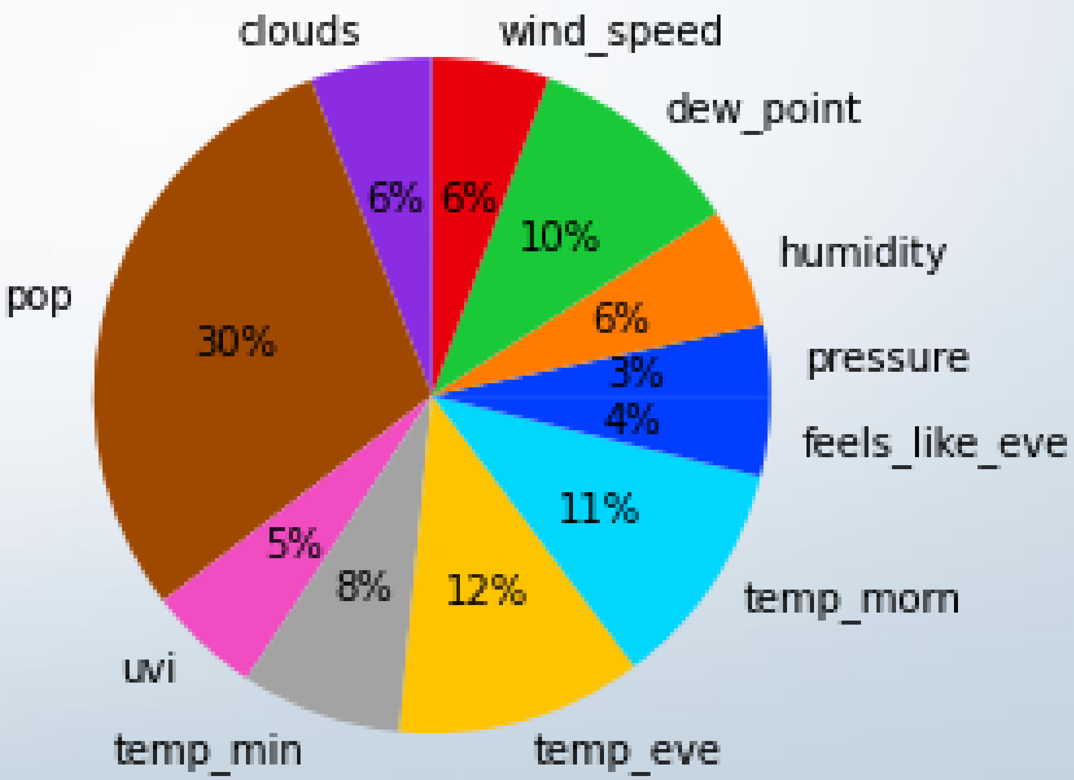
- Feature Importance with DecisionTree Algorithms

Important Features Selection with scaler

```
DecisionTreeClassifier()
```

	feature	importance
5	pop	0.302230
9	temp_morn	0.112852
2	dew_point	0.098716
8	temp_eve	0.098160
7	temp_min	0.074012
6	uvi	0.068048
1	humidity	0.060233
3	wind_speed	0.055582
4	clouds	0.051992
10	feels_like_eve	0.039181
0	pressure	0.038995

Rain weather_description



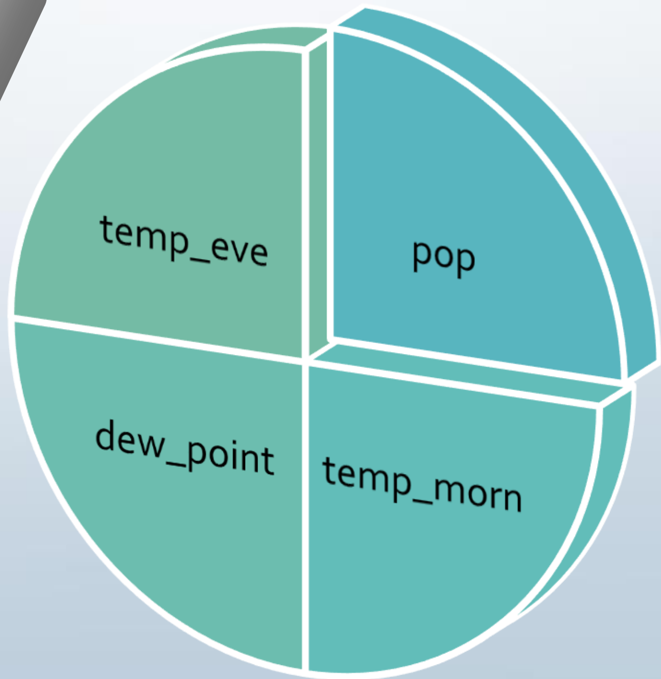
Project Steps

Feature selection

- Final Dataset for Machine Learning Algorithms

	pressure	humidity	dew_point	wind_speed	clouds	pop	uvi	temp_min	temp_eve	temp_morn	feels_like_eve	weather_id
0	1009	73	296.42	5.13	100	0.98	13.39	297.92	300.72	298.03	303.70	501
1	1008	60	294.75	5.13	67	0.41	14.43	297.90	301.17	297.90	304.27	500
2	1007	60	294.06	4.51	77	0.38	15.20	297.40	300.91	297.40	303.73	500
3	1007	66	295.27	4.27	99	0.30	14.38	297.77	300.94	297.79	304.05	500
4	1007	69	296.01	3.06	86	0.97	14.31	298.87	300.56	299.19	303.71	500

4 Important Features



Project Steps

Modeling and Hyperparameter Optimization

- **Default & Tunned Result with whole data**

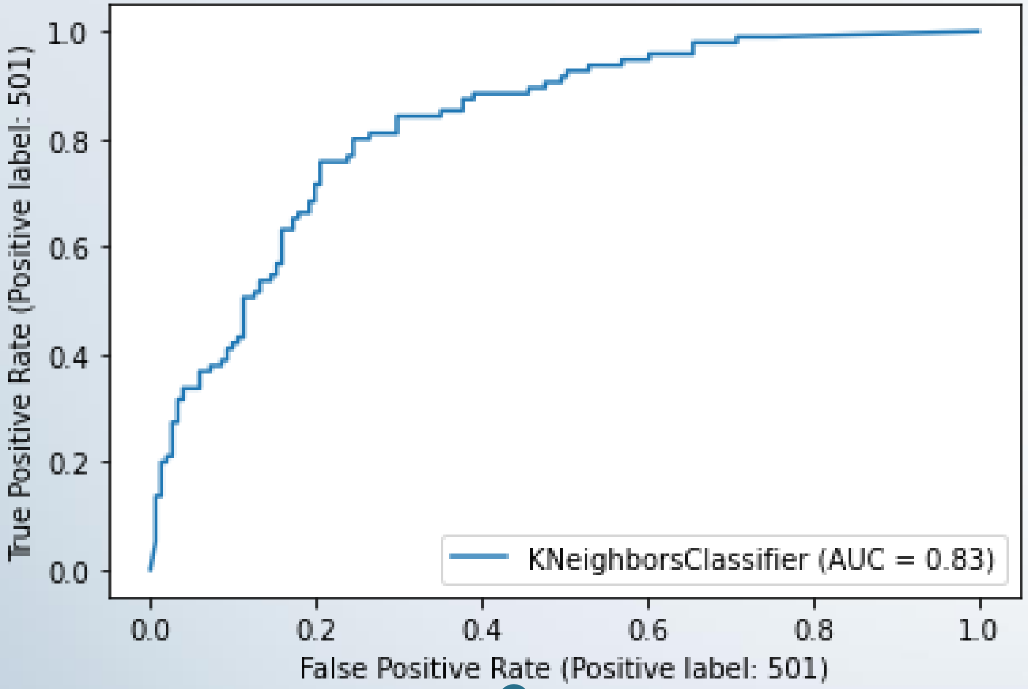
Default Model	Training_time	test_Accuracy	preci_score_500	recall_score_500	preci_score_501	recall_score_501	f1_score
DecisionTree	0.012645	0.711382	0.768212	0.763158	0.621053	0.627660	0.711669
KNN	0.006194	0.752033	0.847682	0.771084	0.600000	0.712500	0.756793
LogesticRegression	0.038268	0.760163	0.841060	0.783951	0.631579	0.714286	0.763318
RandomForest	0.265652	0.743902	0.841060	0.765060	0.589474	0.700000	0.748819
SVM	0.036104	0.739837	0.841060	0.760479	0.578947	0.696203	0.745254

Tunned Model	Training_time	test_Accuracy	preci_score_500	recall_score_500	preci_score_501	recall_score_501	f1_score
DecisionTree	0.009862	0.731707	0.754967	0.797203	0.694737	0.640777	0.729938
KNN	0.005131	0.735772	0.867550	0.744318	0.526316	0.714286	0.745689
LogisticRegression	0.039320	0.739837	0.814570	0.773585	0.621053	0.678161	0.742198
RandomForest	0.134123	0.784553	0.854305	0.806250	0.673684	0.744186	0.786792
SVC	0.041927	0.764228	0.834437	0.792453	0.652632	0.712644	0.766367

Project Steps

Modeling and Hyperparameter Optimization

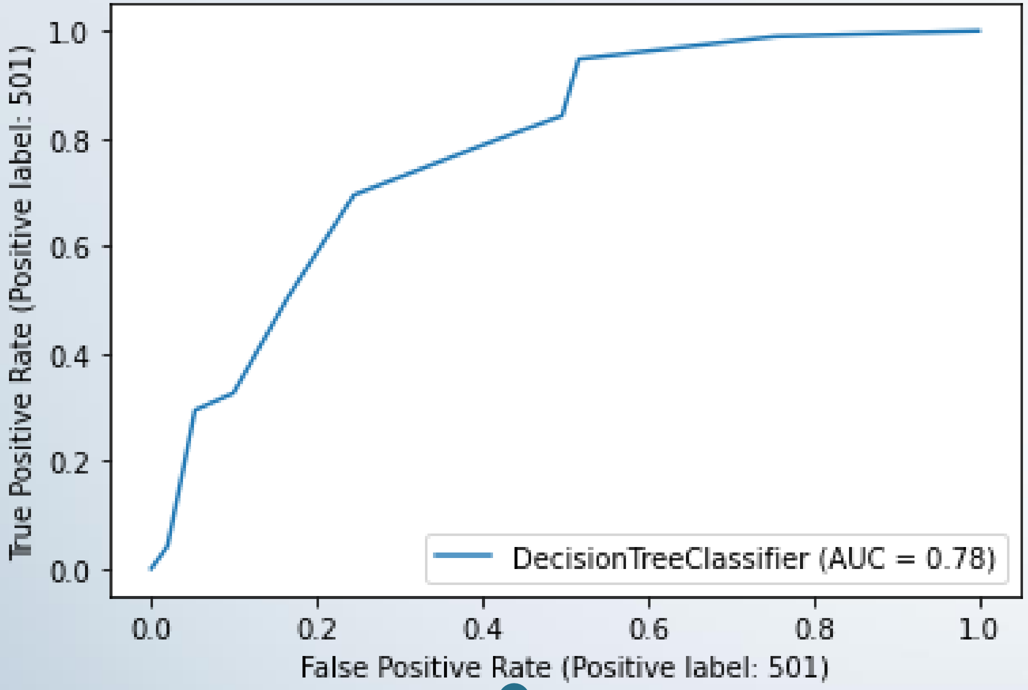
- KNN evaluation Details




```
Classification report:
              precision    recall  f1-score   support

         500       0.74      0.87      0.80       151
         501       0.71      0.53      0.61        95

    accuracy                           0.74       246
   macro avg       0.73      0.70      0.70       246
weighted avg       0.73      0.74      0.73       246

Confusion matrix (Rows actual, Columns predicted):
     0   1
0  131  20
1   45  50
```

Modeling and Hyperparameter Optimization

- **DecisionTree evaluation Details**

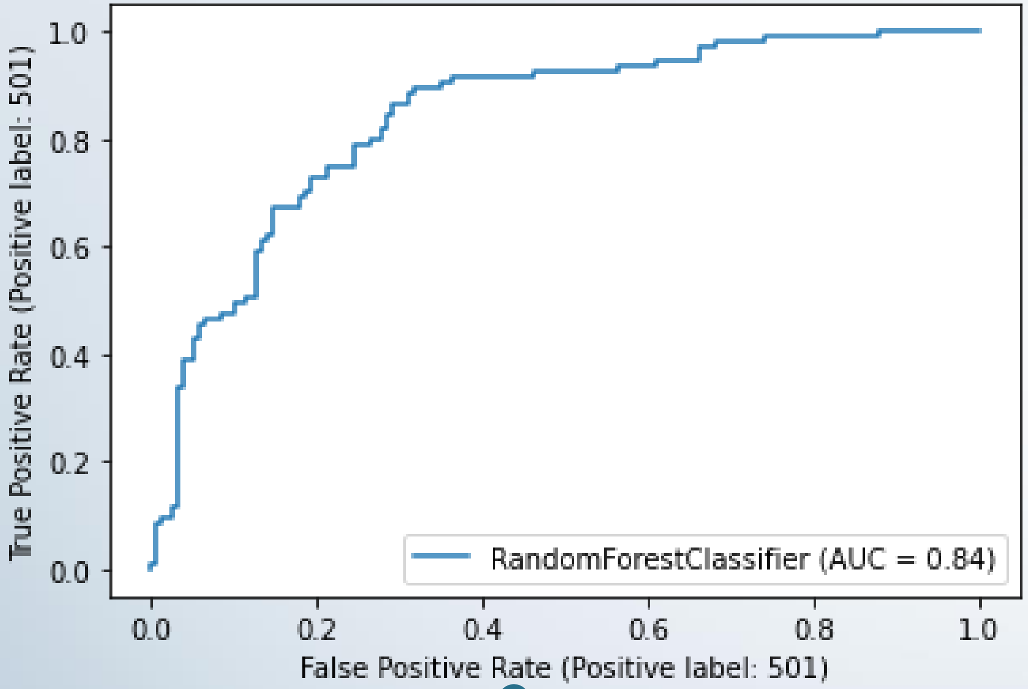



```
Classification report:
              precision    recall  f1-score   support

         500       0.80      0.75      0.78       151
         501       0.64      0.69      0.67        95

    accuracy                           0.73       246
   macro avg       0.72      0.72      0.72       246
weighted avg       0.74      0.73      0.73       246

Confusion matrix (Rows actual, Columns predicted):
     0   1
0  114  37
1   29  66
```

Project Steps

Modeling and Hyperparameter Optimization

- **RandomForest evaluation Details**

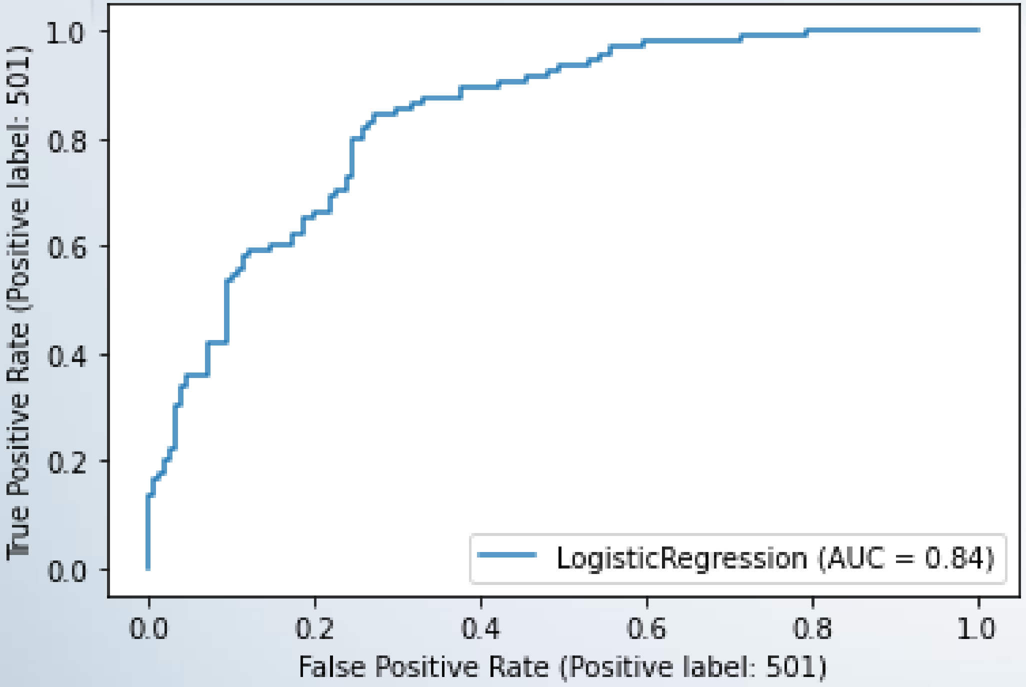



```
Classification report:
              precision    recall  f1-score   support

         500       0.81      0.85      0.83       151
         501       0.74      0.67      0.71        95

    accuracy                           0.78       246
   macro avg       0.78      0.76      0.77       246
weighted avg       0.78      0.78      0.78       246

Confusion matrix (Rows actual, Columns predicted):
     0   1
0  129  22
1   31  64
```

Project Steps

Modeling and Hyperparameter Optimization

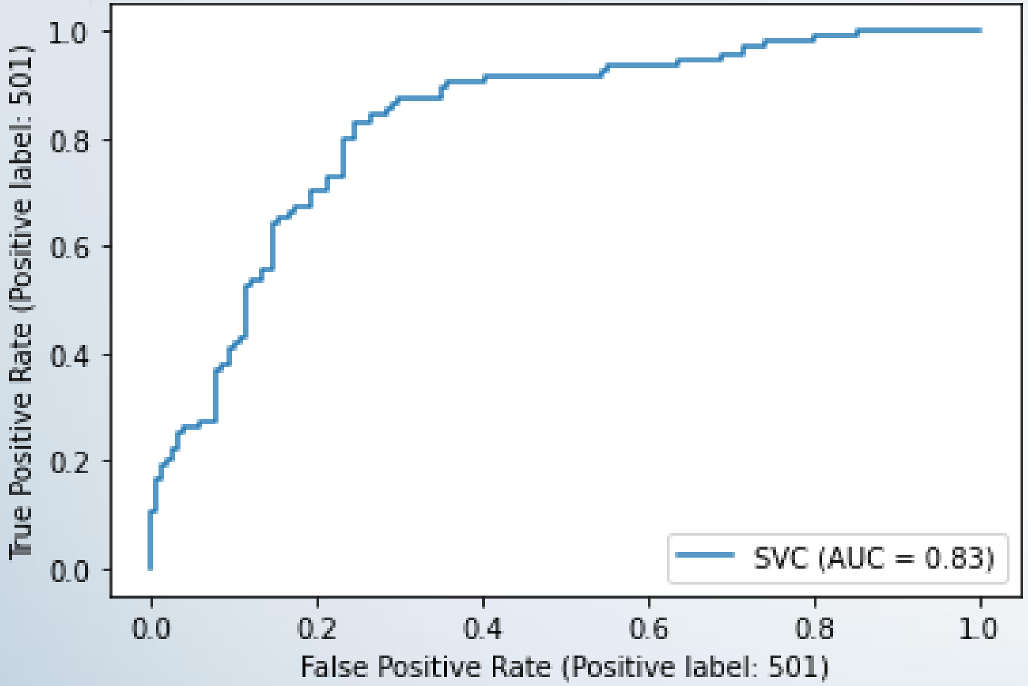
- LogesticRegression evaluation Details




```
Classification report:
              precision    recall  f1-score   support

         500       0.77      0.81      0.79       151
         501       0.68      0.62      0.65        95

    accuracy                           0.74       246
   macro avg       0.73      0.72      0.72       246
weighted avg       0.74      0.74      0.74       246

Confusion matrix (Rows actual, Columns predicted):
     0   1
0  123  28
1   36  59
```

Modeling and Hyperparameter Optimization

- SVM evaluation Details




```
Classification report:
              precision    recall  f1-score   support

         500       0.79      0.83      0.81       151
         501       0.71      0.65      0.68        95

    accuracy                           0.76       246
   macro avg       0.75      0.74      0.75       246
weighted avg       0.76      0.76      0.76       246

Confusion matrix (Rows actual, Columns predicted):
     0   1
0  126  25
1   33  62
```

Project Steps

Modeling and Hyperparameter Optimization

- Default & Tunned Result with 4 important features

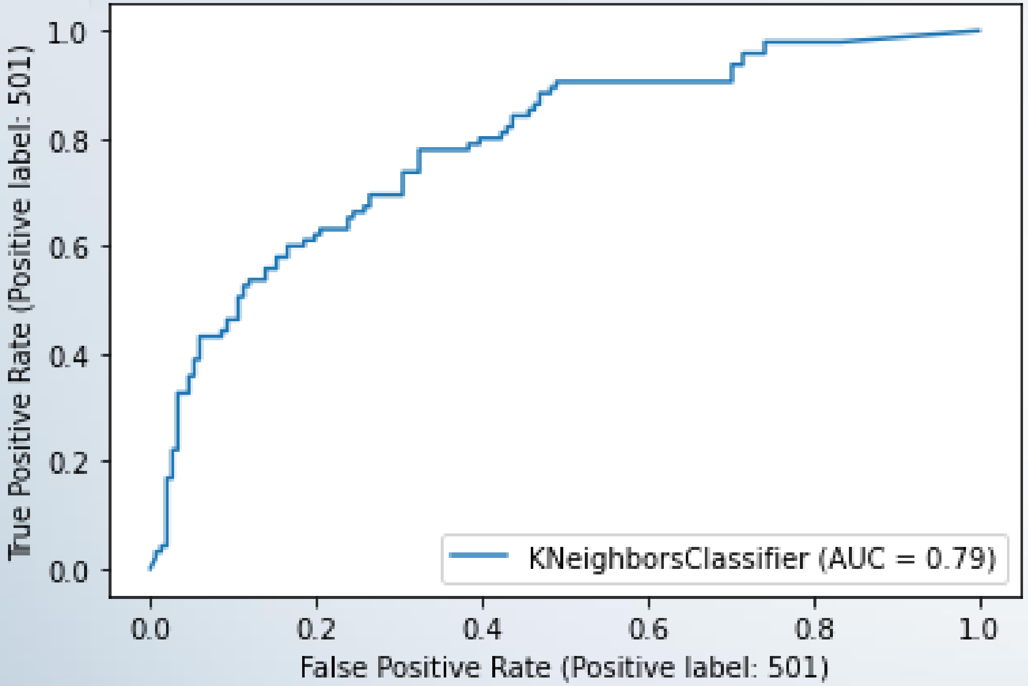
Default Model	Training_time	test_Accuracy	preci_score_500	recall_score_500	preci_score_501	recall_score_501	f1_score
DecisionTree	0.008922	0.727642	0.768212	0.783784	0.663158	0.642857	0.726892
KNN	0.006016	0.727642	0.807947	0.762500	0.600000	0.662791	0.730473
LogisticRegression	0.029366	0.747967	0.801325	0.790850	0.663158	0.677419	0.748479
RandomForest	0.243966	0.743902	0.788079	0.793333	0.673684	0.666667	0.743657
SVC	0.039934	0.613821	1.000000	0.613821	0.000000	0.000000	0.760705

Tunned Model	Training_time	test_Accuracy	preci_score_500	recall_score_500	preci_score_501	recall_score_501	f1_score
DecisionTree	0.009302	0.735772	0.761589	0.798611	0.694737	0.647059	0.734213
KNN	0.007191	0.743902	0.834437	0.768293	0.600000	0.695122	0.748023
LogisticRegression	0.018005	0.711382	0.834437	0.732558	0.515789	0.662162	0.719932
RandomForest	0.118569	0.772358	0.827815	0.806452	0.684211	0.714286	0.773318
SVC	0.042829	0.727642	0.788079	0.772727	0.631579	0.652174	0.728487

Project Steps

Modeling and Hyperparameter Optimization

- **KNN evaluation Details**

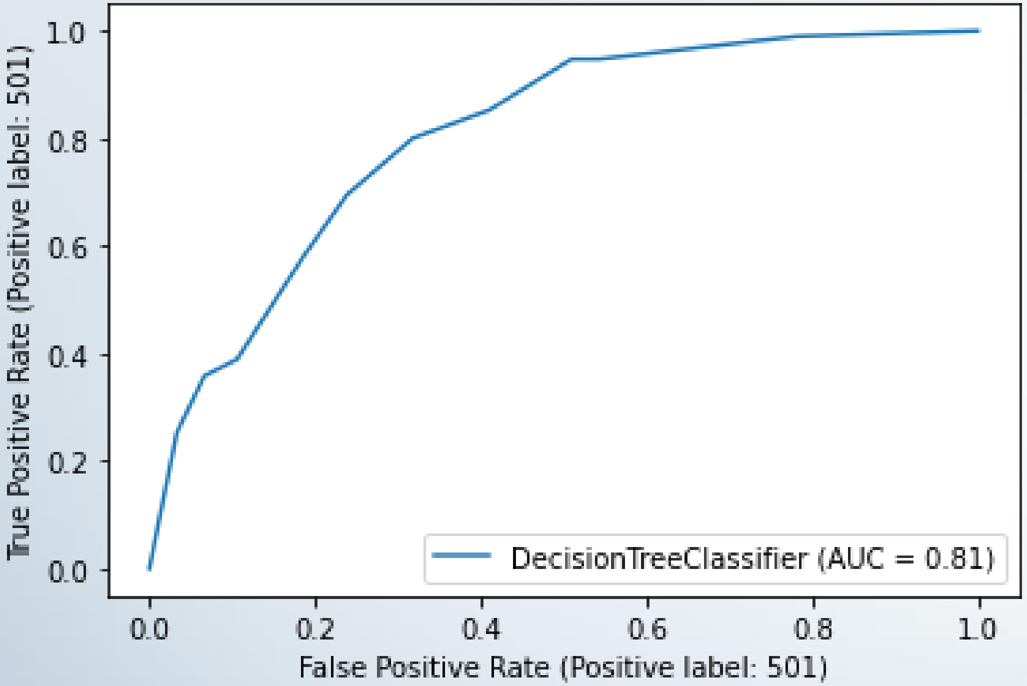



```
Classification report:
              precision    recall  f1-score   support

         500       0.77      0.83      0.80       151
         501       0.70      0.60      0.64        95

    accuracy                           0.74       246
   macro avg       0.73      0.72      0.72       246
weighted avg       0.74      0.74      0.74       246

Confusion matrix (Rows actual, Columns predicted):
     0   1
0  126  25
1   38  57
```

Project Steps

Modeling and Hyperparameter Optimization

- **DecisionTree evaluation Details**

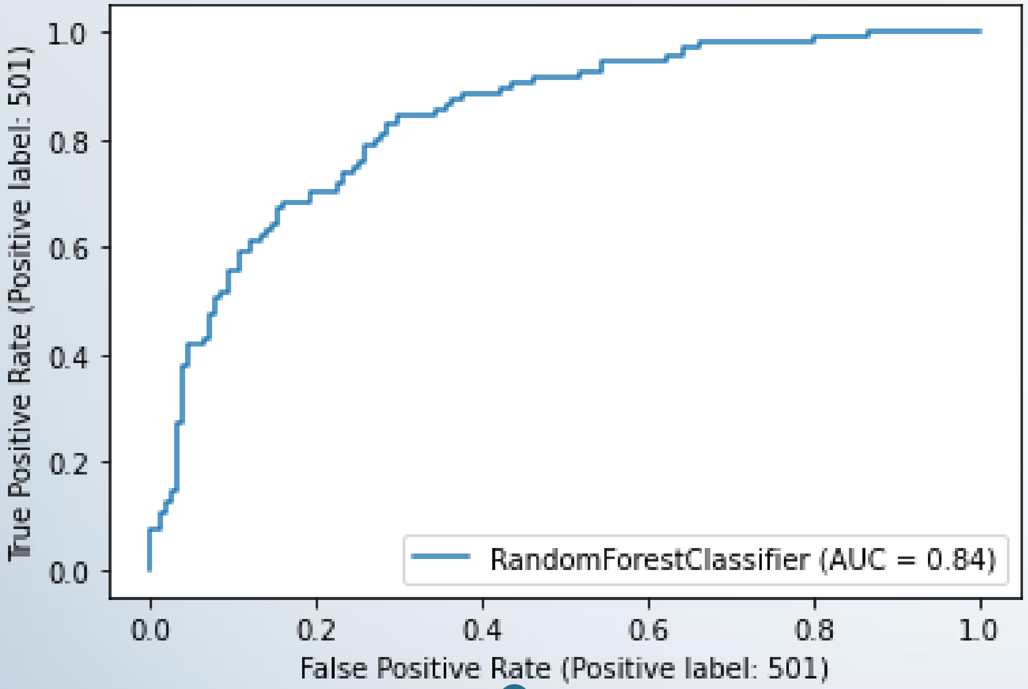



```
Classification report:
              precision    recall  f1-score   support

         500       0.80      0.76      0.78       151
         501       0.65      0.69      0.67        95

    accuracy                           0.74       246
   macro avg       0.72      0.73      0.72       246
weighted avg       0.74      0.74      0.74       246

Confusion matrix (Rows actual, Columns predicted):
     0    1
0  115   36
1   29   66
```

Project Steps

Modeling and Hyperparameter Optimization

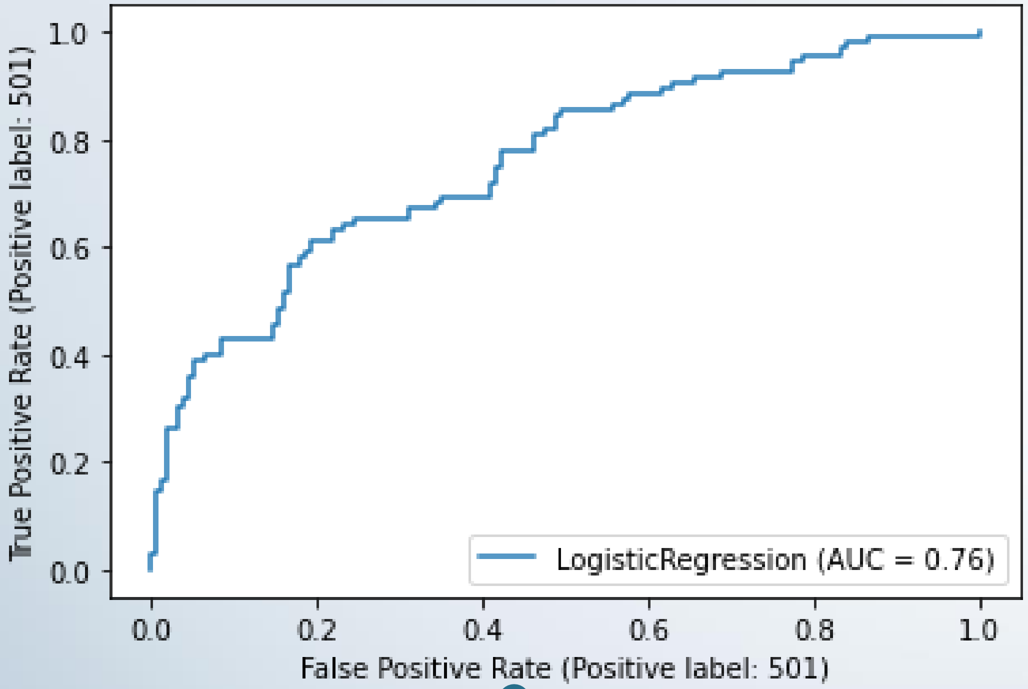
- **RandomForest evaluation Details**




```
Classification report:
              precision    recall  f1-score   support

         500       0.81      0.83      0.82       151
         501       0.71      0.68      0.70        95

    accuracy                           0.77       246
   macro avg       0.76      0.76      0.76       246
weighted avg       0.77      0.77      0.77       246

Confusion matrix (Rows actual, Columns predicted):
     0   1
0  125  26
1   30  65
```

Modeling and Hyperparameter Optimization

- LogesticRegression evaluation Details



Classification report:

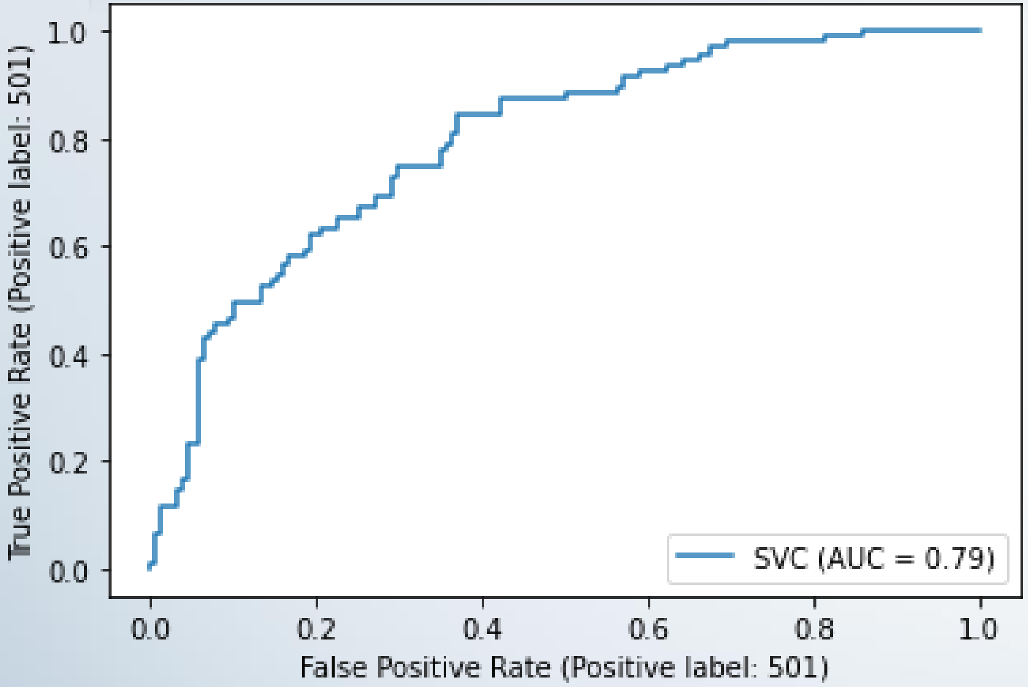
	precision	recall	f1-score	support
500	0.73	0.83	0.78	151
501	0.66	0.52	0.58	95
accuracy			0.71	246
macro avg	0.70	0.68	0.68	246
weighted avg	0.71	0.71	0.70	246

Confusion matrix (Rows actual, Columns predicted):

	0	1
0	126	25
1	46	49

Modeling and Hyperparameter Optimization

- **SVM evaluation Details**




```
Classification report:
              precision    recall  f1-score   support

         500       0.77      0.79      0.78       151
         501       0.65      0.63      0.64        95

    accuracy                           0.73       246
   macro avg       0.71      0.71      0.71       246
weighted avg       0.73      0.73      0.73       246

Confusion matrix (Rows actual, Columns predicted):
     0   1
0  119  32
1   35  60
```