Stroke Prediction Dataset

Classification Project

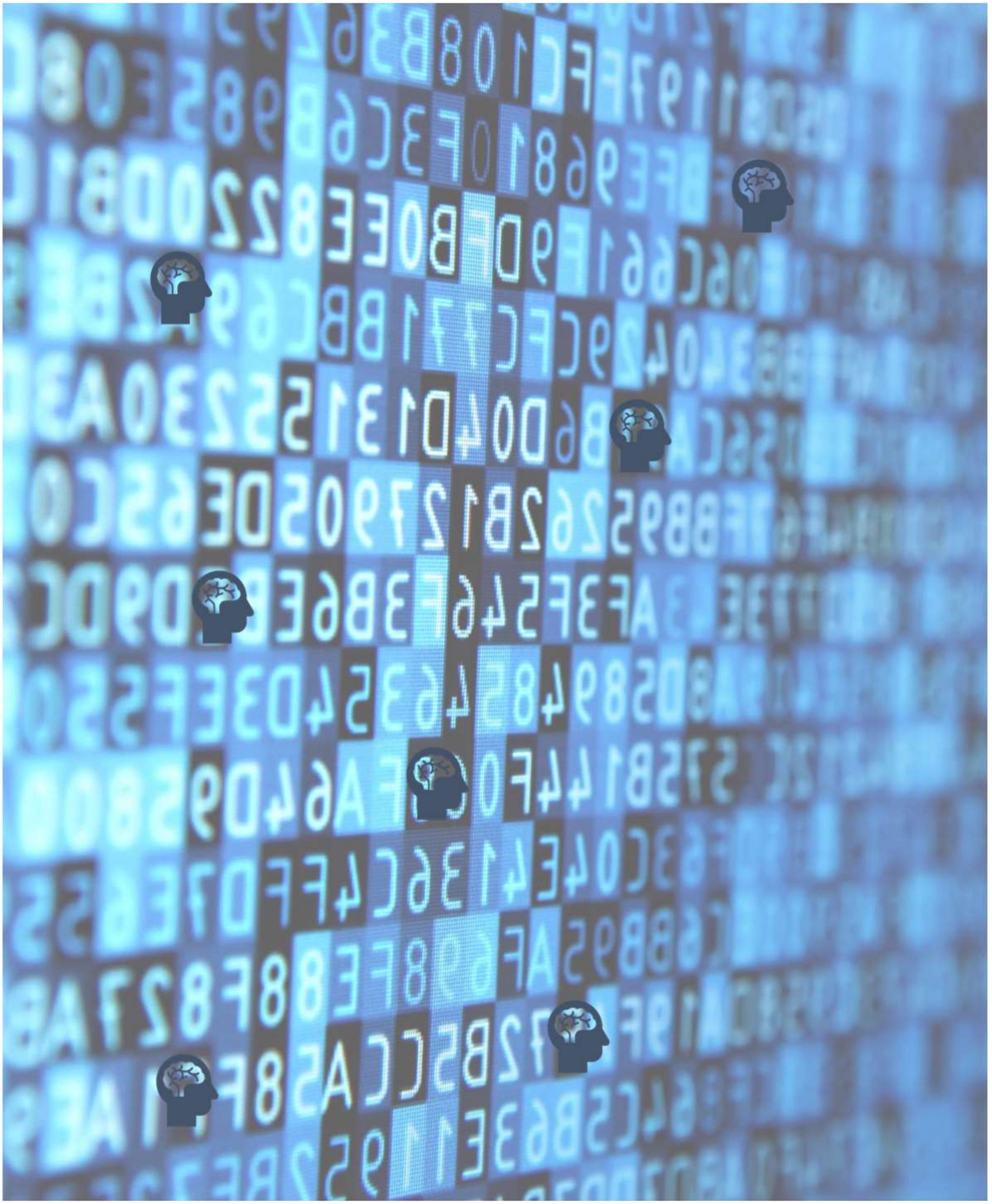
Tomer Badug. Shirli Miller. Judi Eliya

project goal

Prediction Stroke

what is a stroke?

Stroke is a medical emergency. A stroke occurs when blood flow to a part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die within minutes



Data content

Feature:

1. id
2. gender
3. age
4. hypertension
5. heart_disease
6. ever_married
7. work_type
8. Residence_type
9. avg_glucose_level
10. bmi
11. smoking_status

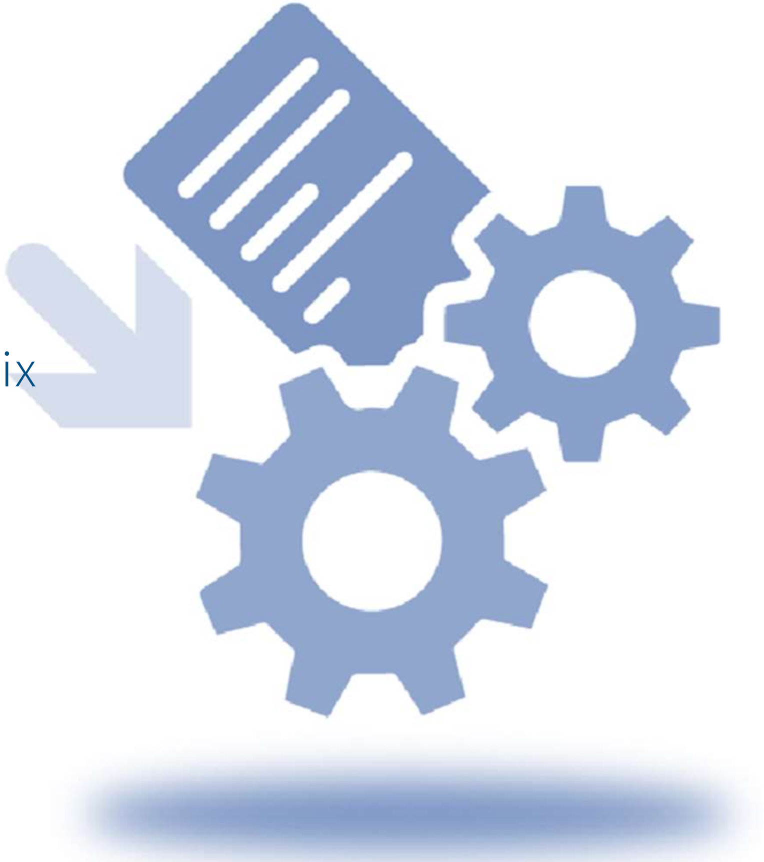
12 columns

5110 row

Target variable – stroke

Exploratory data analysis -EDA

- ➢Missing Values Handling
- ➢Remove duplicate and outlier
- ➢Feature selection
- ➢Compute pairwise correlation of columns – matrix
- ➢Encoding
- ➢Understanding the variables



EDA

1. ID

Feature selection

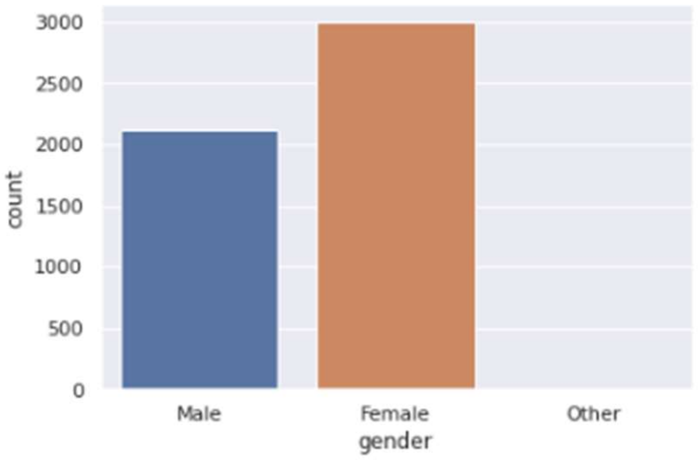
1. Unique value
2. Drop

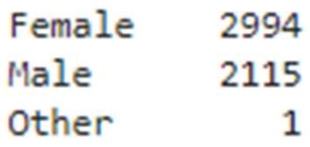
11 columns, 5110 row

EDA

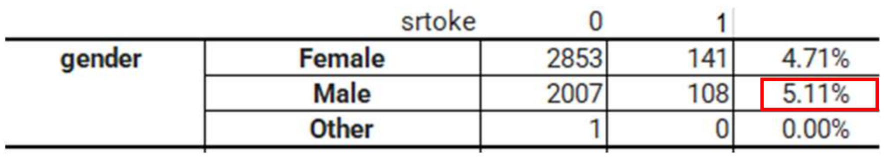
2. Gender

"Male", "Female" or "Other"





```
Female      2994
Male        2115
Other          1
```



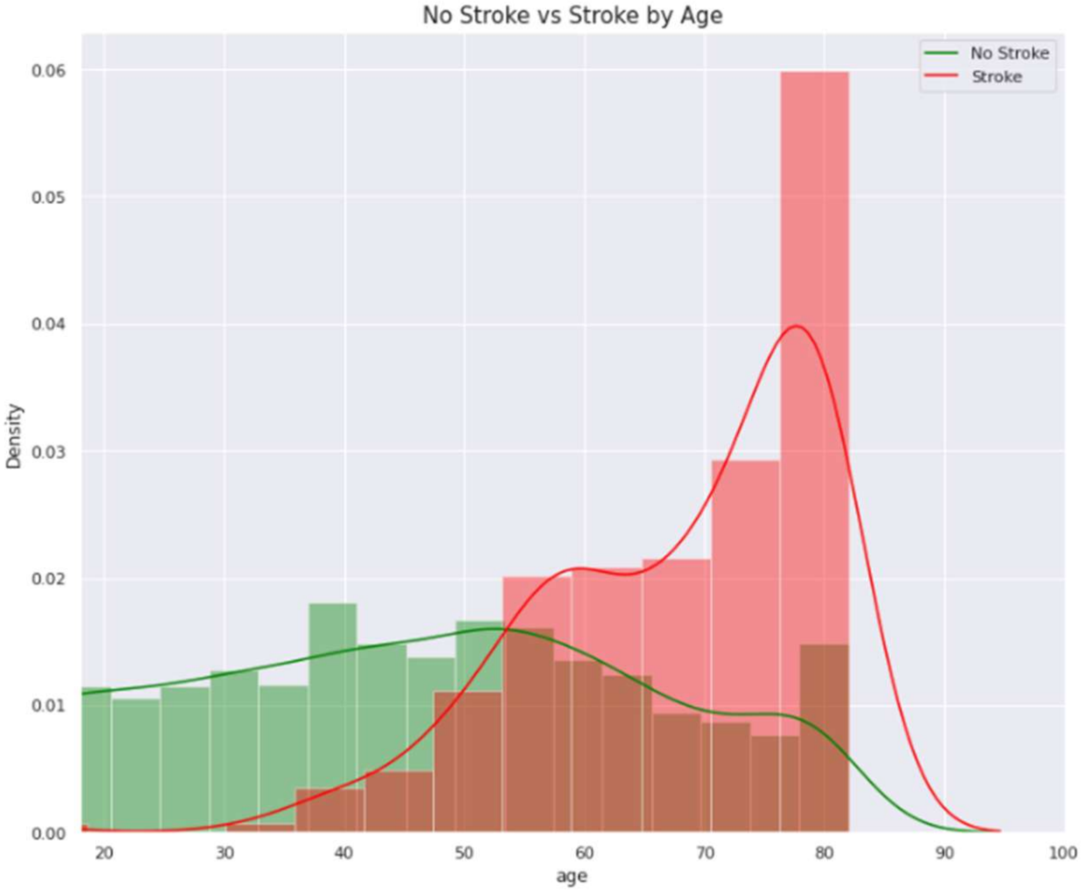
	srtoke	0	1	
gender	Female	2853	141	4.71%
	Male	2007	108	5.11%
	Other	1	0	0.00%

11 columns, 5109 row

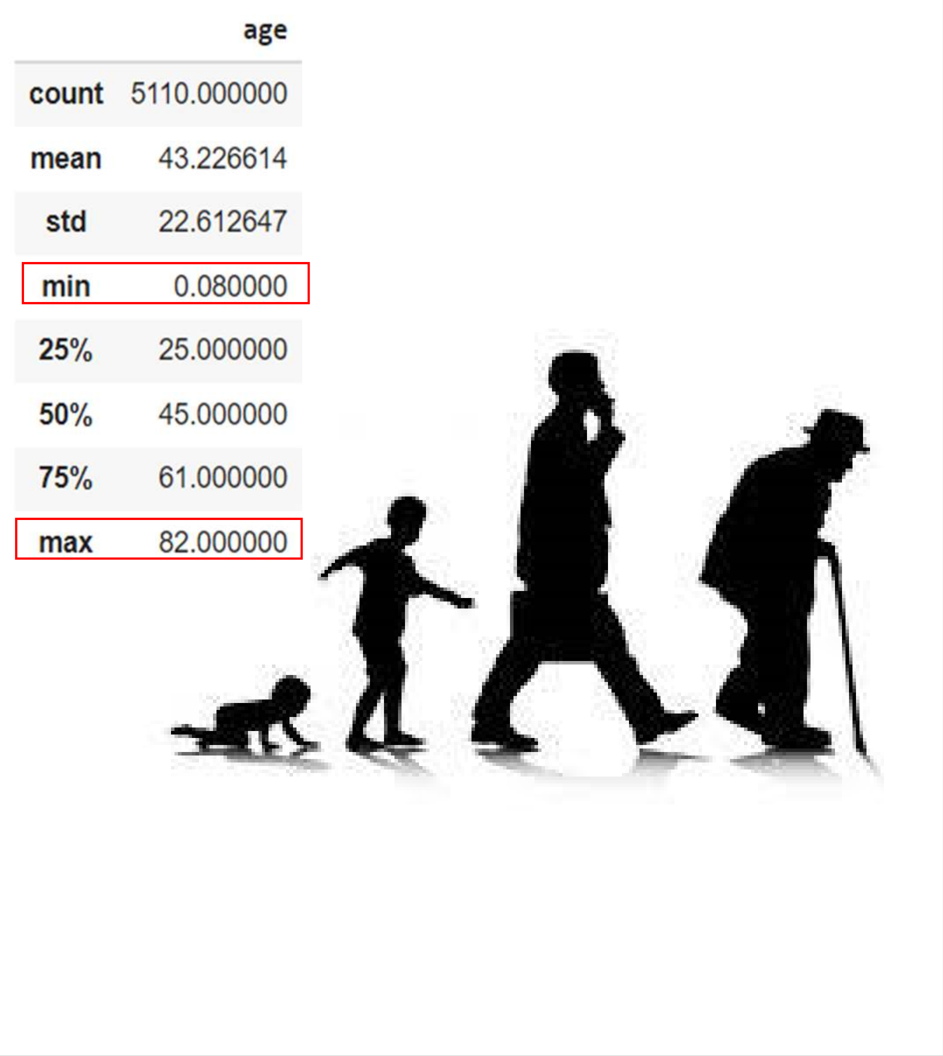
EDA

3. Age:

Age of the patient



11 columns, 5109 row



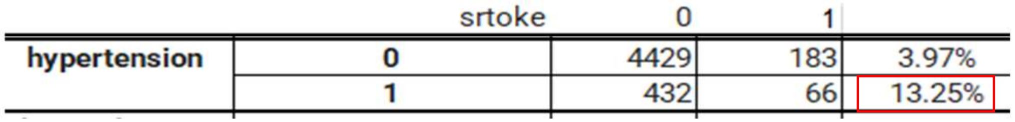
	age
count	5110.000000
mean	43.226614
std	22.612647
min	0.080000
25%	25.000000
50%	45.000000
75%	61.000000
max	82.000000

EDA

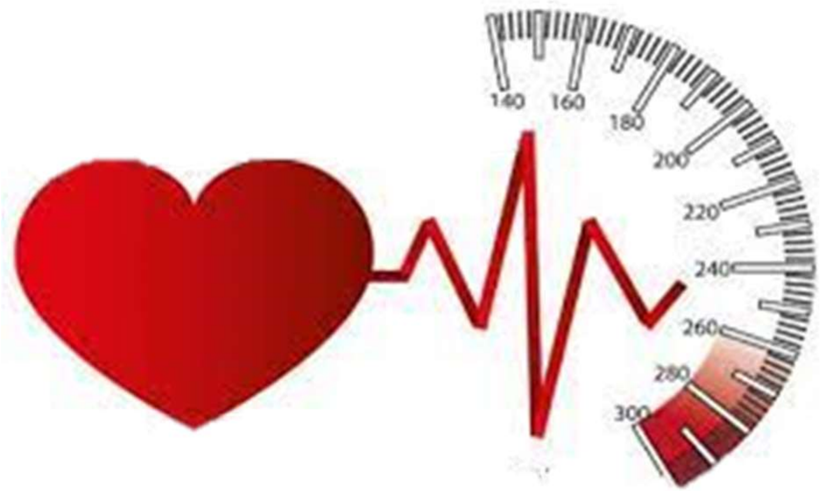
4. Hypertension :

0 - No hypertension

1 - Hypertension



	srtoke	0	1	
hypertension	0	4429	183	3.97%
	1	432	66	13.25%



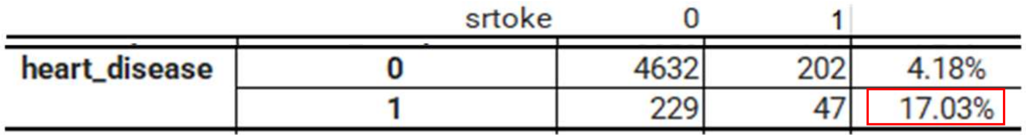
11 columns, 5109 row

EDA

5. Heart disease :

0 - No heart diseases

1 - Heart disease



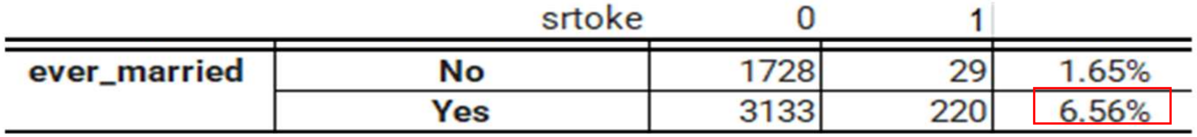
	srtoke	0	1	
heart_disease	0	4632	202	4.18%
	1	229	47	17.03%

11 columns, 5109 row

EDA

6. Ever married :

"No" or "Yes"



	srtoke	0	1	
ever_married	No	1728	29	1.65%
	Yes	3133	220	6.56%

11 columns, 5109 row

EDA

7. Work type :

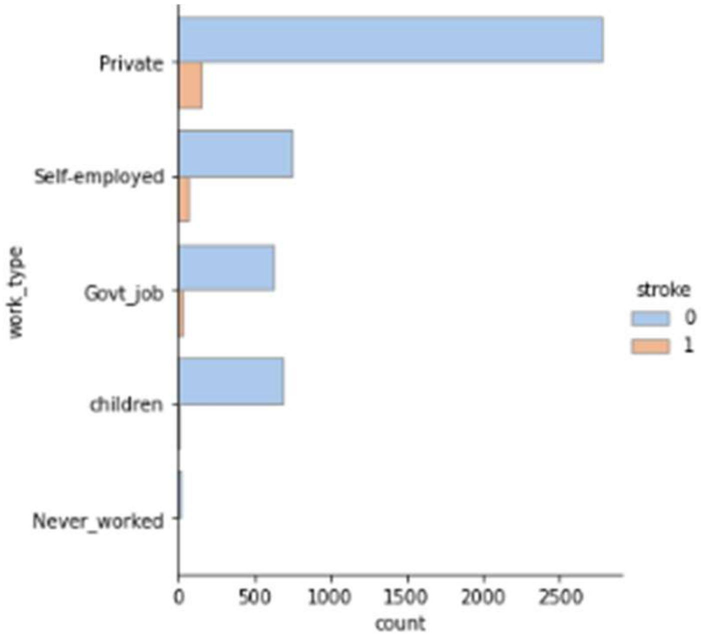
"Private"

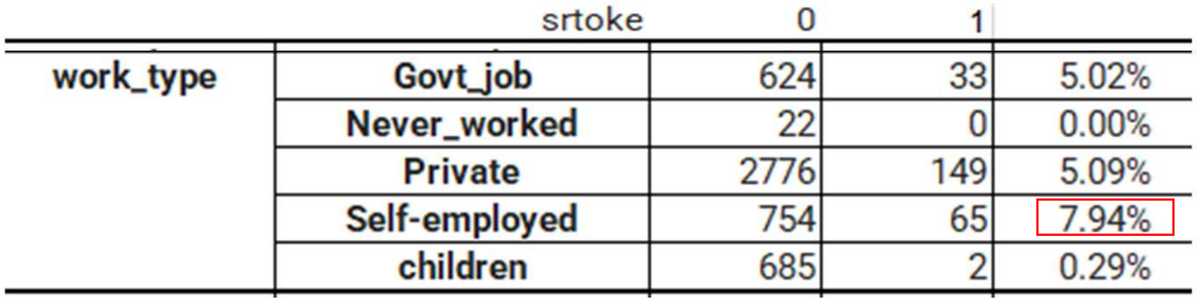
"Self-employed“

"Govt_jov“

"Children“

"Never_worked"





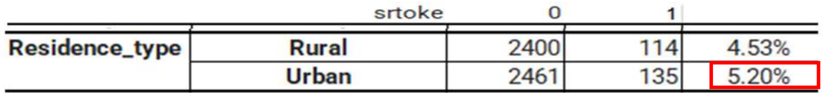
	srtoke	0	1	
work_type	Govt_job	624	33	5.02%
	Never_worked	22	0	0.00%
	Private	2776	149	5.09%
	Self-employed	754	65	7.94%
	children	685	2	0.29%

11 columns, 5109 row

EDA

9. Residence type :

"Rural" or "Urban"



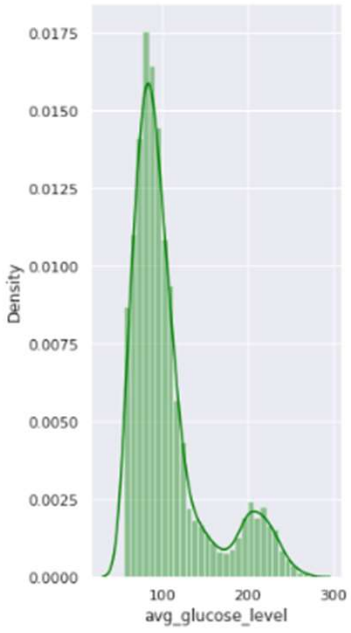
	srtoke	0	1	
Residence_type	Rural	2400	114	4.53%
	Urban	2461	135	5.20%

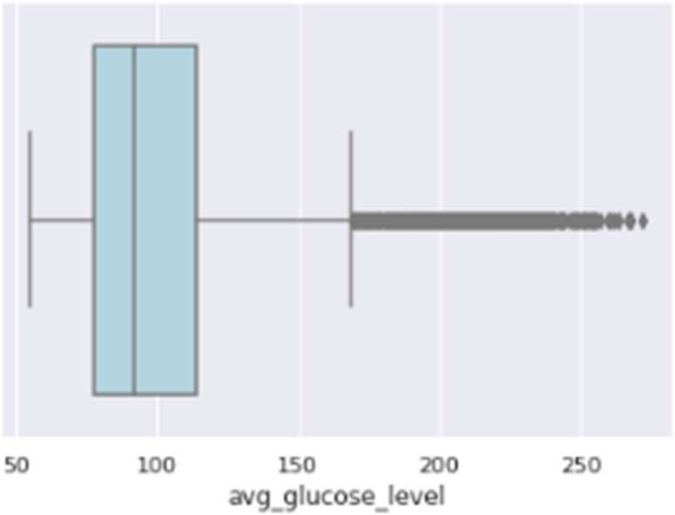
11 columns, 5109 row

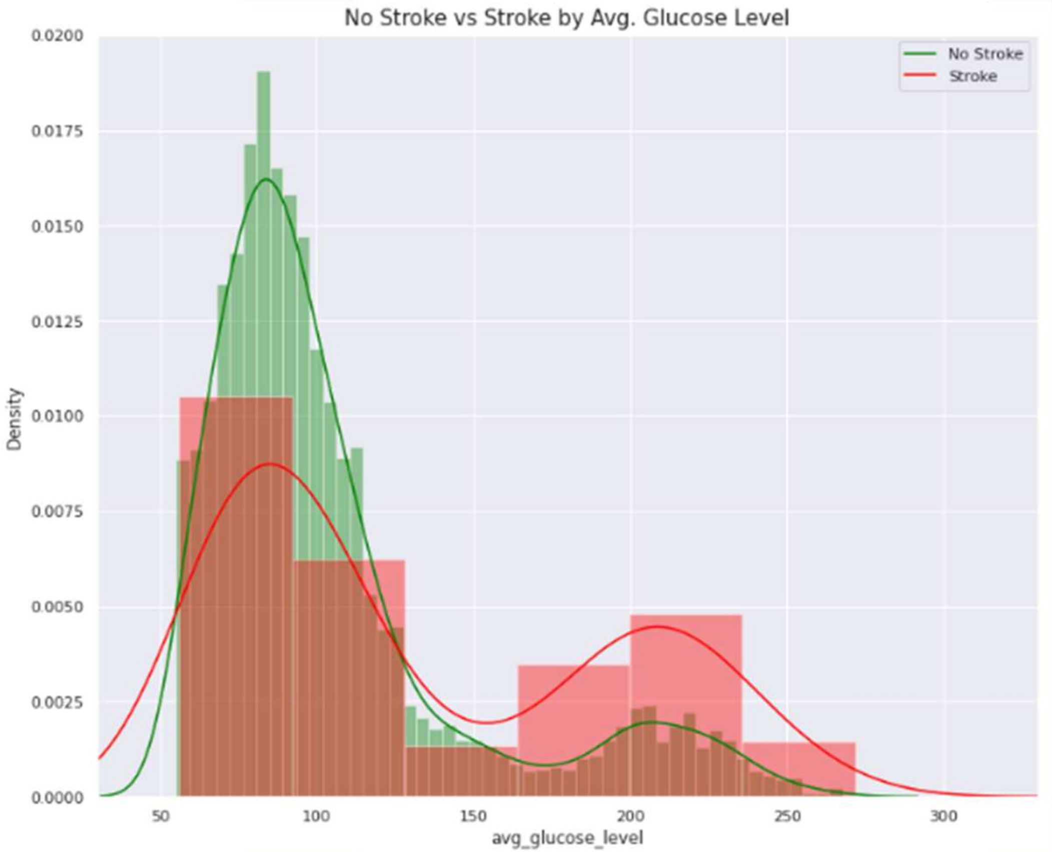
EDA

9. Avg Glucose level:

Average glucose level in blood







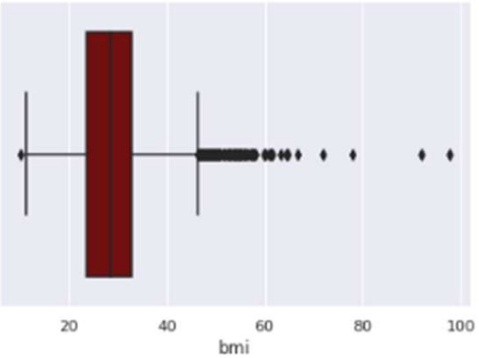
11 columns, 5109 row

EDA

10. BMI:

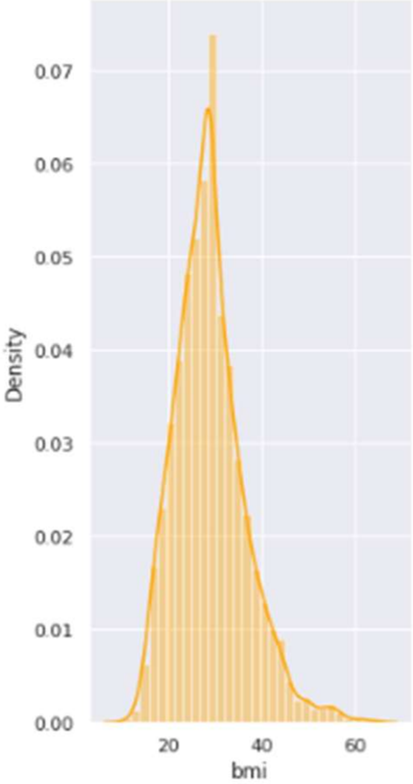
Body Mass Index

201 null - Replacing the missing values with mean



Outlier:

Decision - Drop 5 max bmi

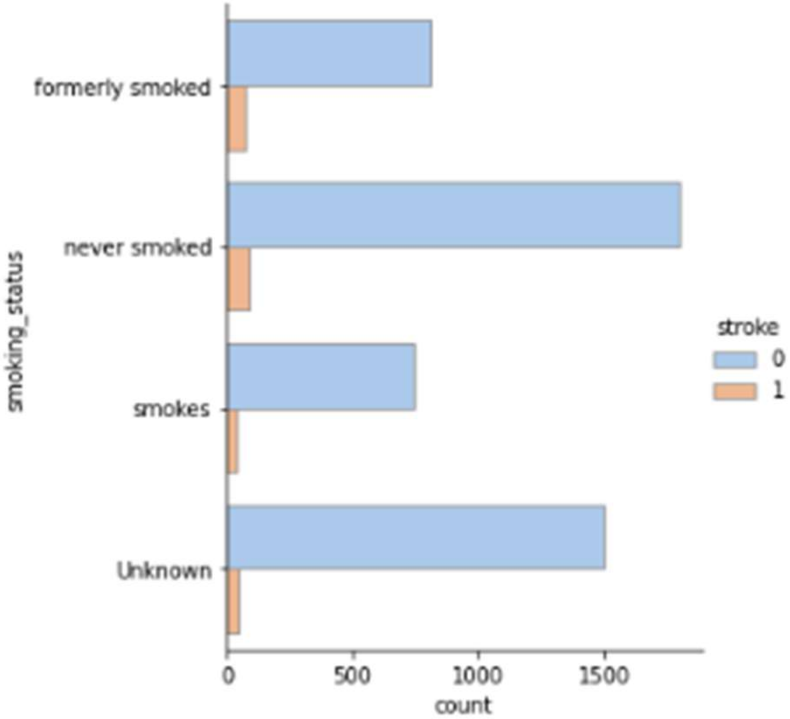


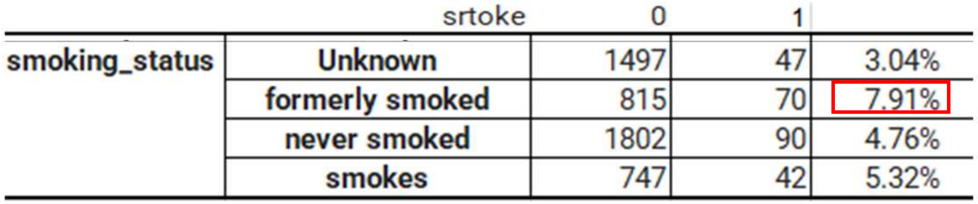
11 columns, 5104 row

EDA

10. Smoking Status:

"Formerly smoked“
"Never smoked“
"Smokes“
 "Unknown"





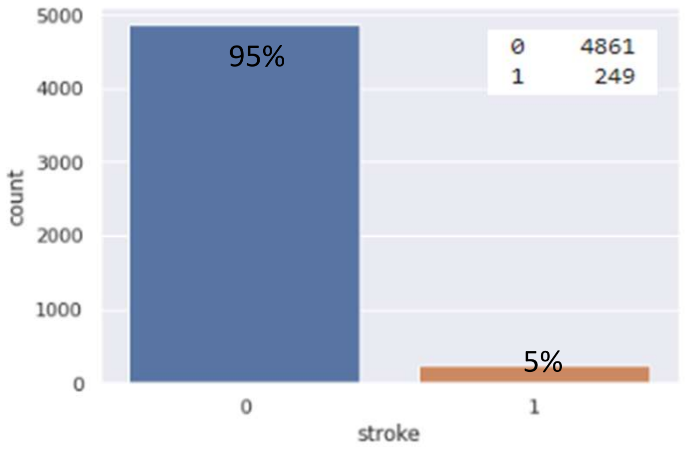
	srtoke	0	1	
smoking_status	Unknown	1497	47	3.04%
	formerly smoked	815	70	7.91%
	never smoked	1802	90	4.76%
	smokes	747	42	5.32%

Target variable

12. Stroke:

0 = No stroke

1 = Stroke

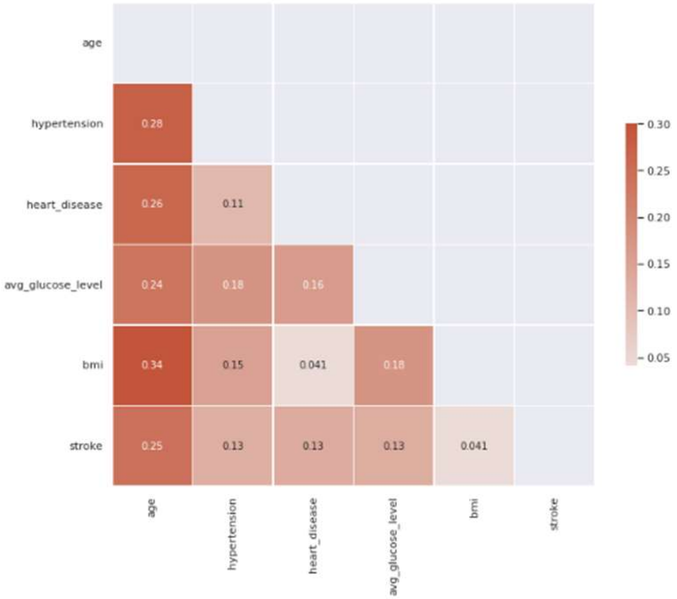


```
0     4861
1      249
```

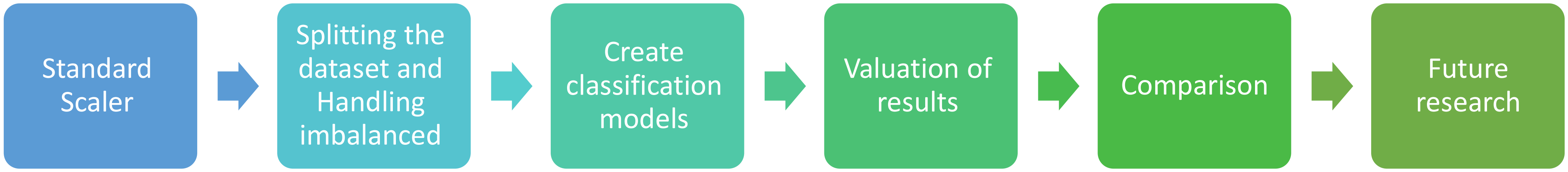
Conclusion and the next steps

Age has the highest impact on stroke, even though the stroke also depends on the other variables,

Such as: glucose level, heart disease, blood pressure, smoking and even type of work and area of residence



Conclusion and the next steps



Standard Scaler

Splitting the dataset and Handling imbalanced

Create classification models

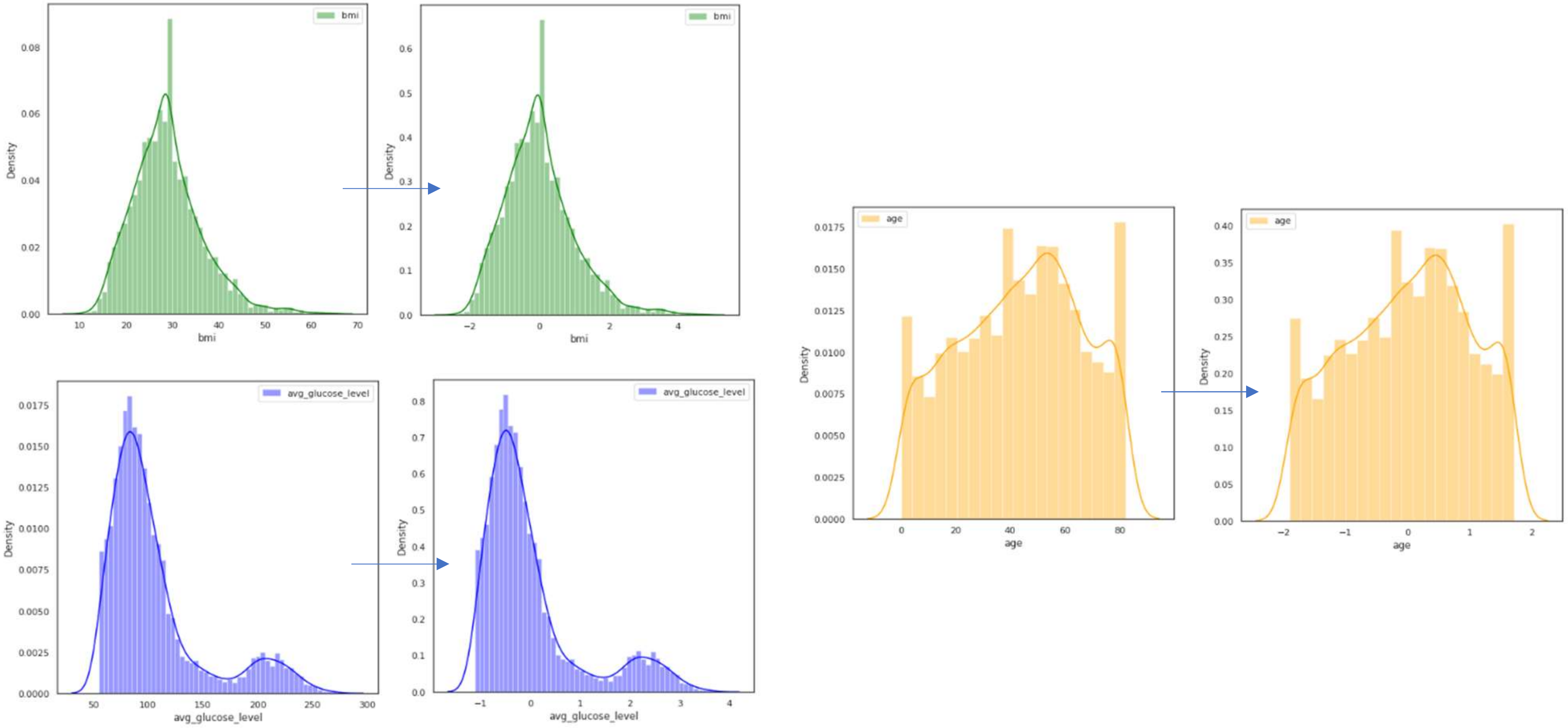
Valuation of results

Comparison

Future research

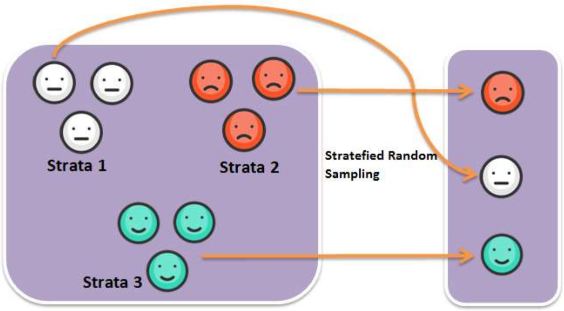
Standard Scaler

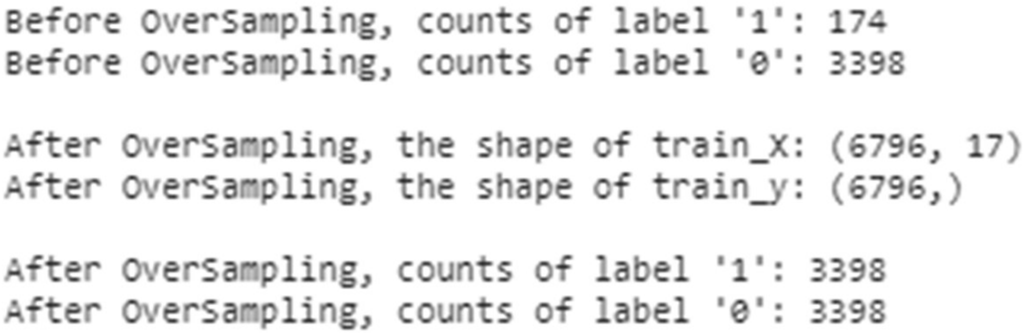
- ➤The models With and Without Standart Scaler



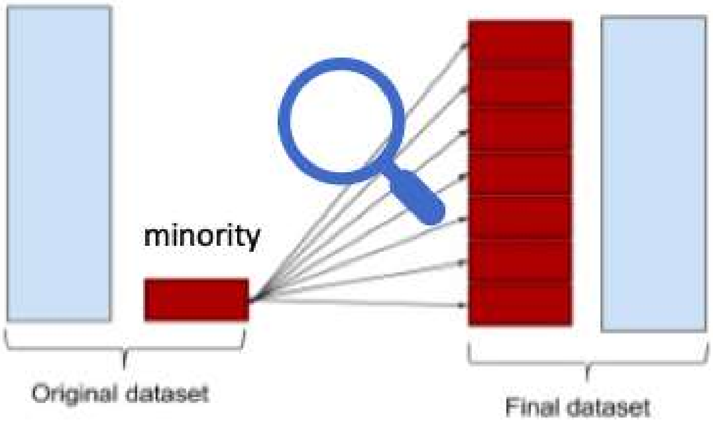
Splitting data and Handling imbalanced

- Splitting dataset 30-70
- stratify (95-5)
- Over Sampling using SMOTE (50-50 train data)





```
Before OverSampling, counts of label '1': 174
Before OverSampling, counts of label '0': 3398

After OverSampling, the shape of train_X: (6796, 17)
After OverSampling, the shape of train_y: (6796,)

After OverSampling, counts of label '1': 3398
After OverSampling, counts of label '0': 3398
```



classification models

- ➢Logistic Regression
- ➢SVM
- ➢Random Forest

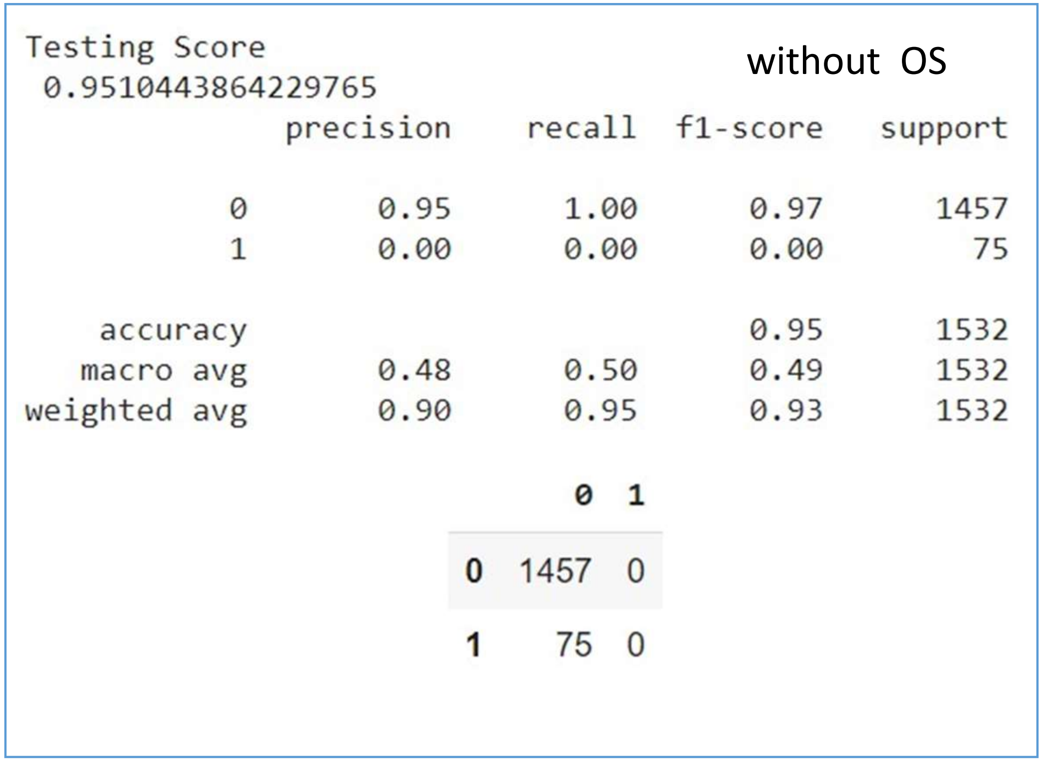
valuation of results

- Precision
- Recall
- F1

and connection

Logistic Regression

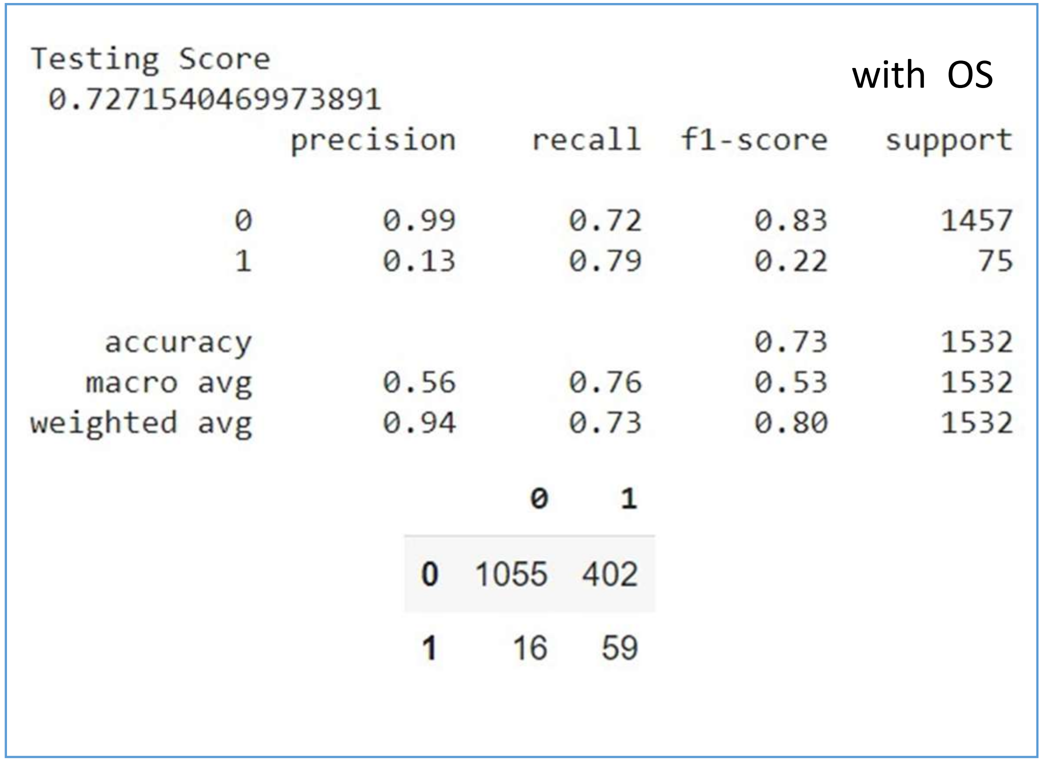
➢With and without Over Sampling



without OS

```
Testing Score
 0.9510443864229765
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      1457
           1       0.00      0.00      0.00        75

    accuracy                           0.95      1532
   macro avg       0.48      0.50      0.49      1532
weighted avg       0.90      0.95      0.93      1532
```

	0	1
0	1457	0
1	75	0

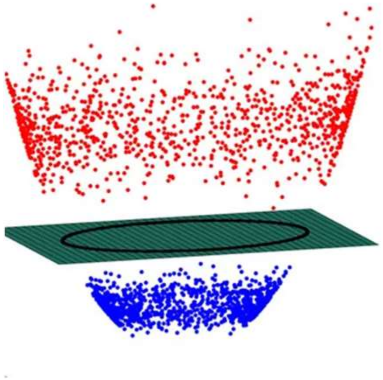


with OS

```
Testing Score
 0.7271540469973891
              precision    recall  f1-score   support

           0       0.99      0.72      0.83      1457
           1       0.13      0.79      0.22        75

    accuracy                           0.73      1532
   macro avg       0.56      0.76      0.53      1532
weighted avg       0.94      0.73      0.80      1532
```

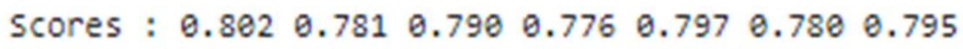
	0	1
0	1055	402
1	16	59

SVM



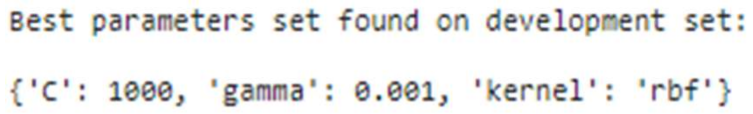
- With and without Over Sampling
- With and without Standart Scaler

- Cross Validation

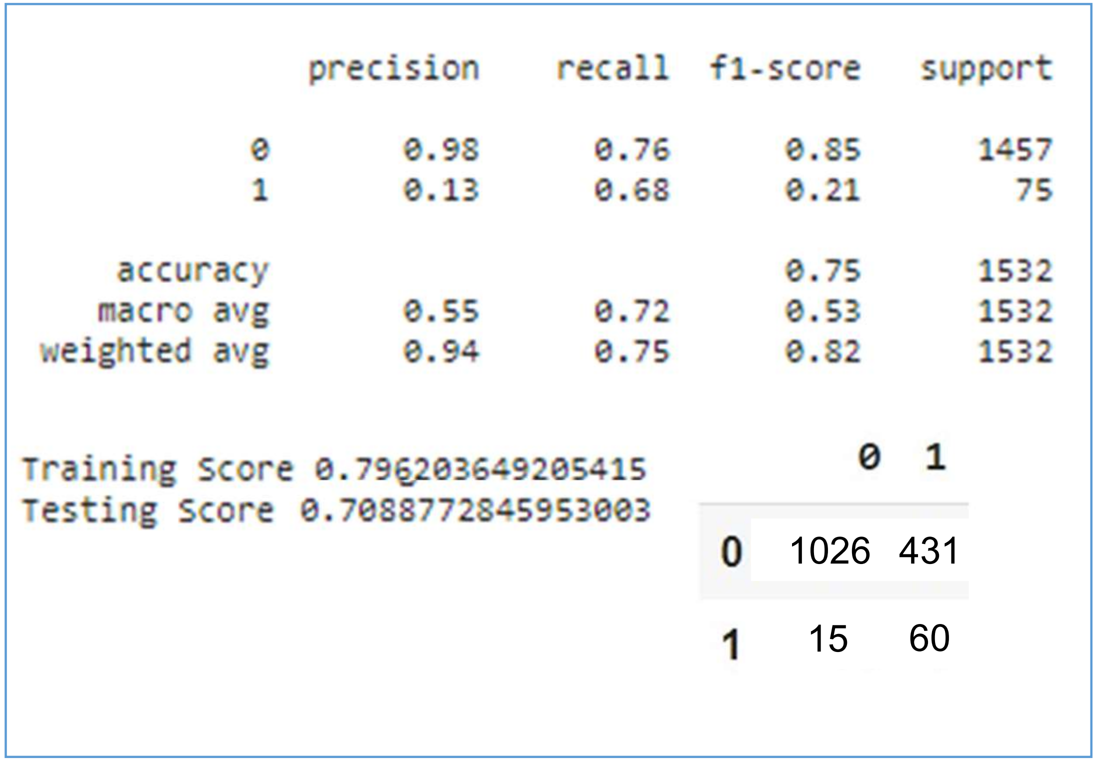


```
Scores : 0.802 0.781 0.790 0.776 0.797 0.780 0.795
```

- Grid Search



```
Best parameters set found on development set:

{'C': 1000, 'gamma': 0.001, 'kernel': 'rbf'}
```



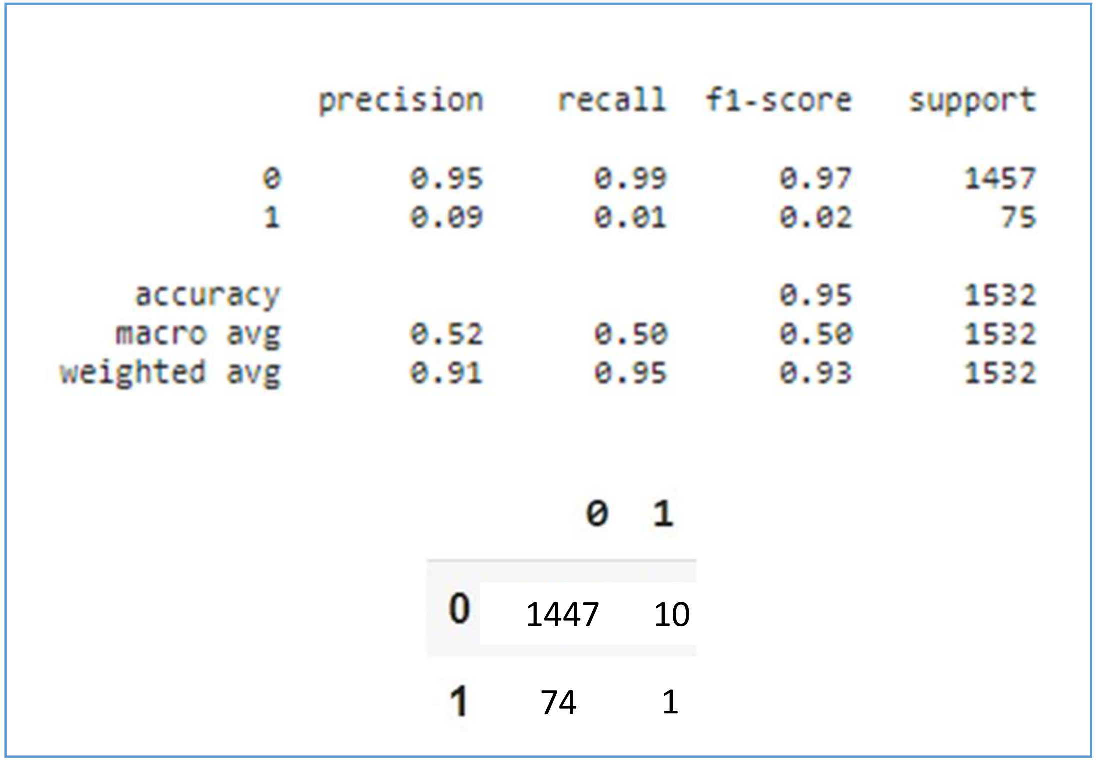
```
              precision    recall  f1-score   support

           0       0.98      0.76      0.85      1457
           1       0.13      0.68      0.21        75

    accuracy                           0.75      1532
   macro avg       0.55      0.72      0.53      1532
weighted avg       0.94      0.75      0.82      1532


Training Score 0.796203649205415
Testing Score  0.7088772845953003
```

	0	1
0	1026	431
1	15	60

Random Forest

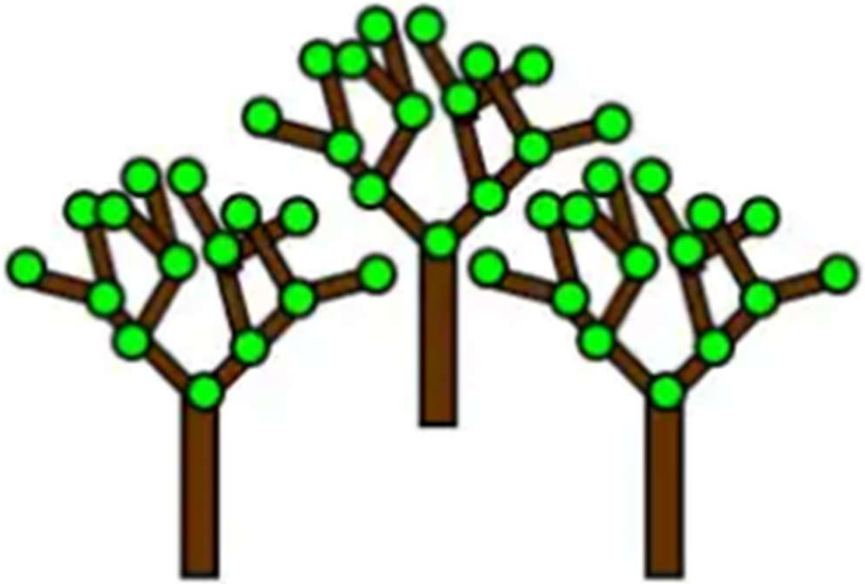
- With and without Over Sampling
- Grid Search

```
Best Parameters :  {'criterion': 'gini', 'n_estimators': 150, 'random_state': 0}
```



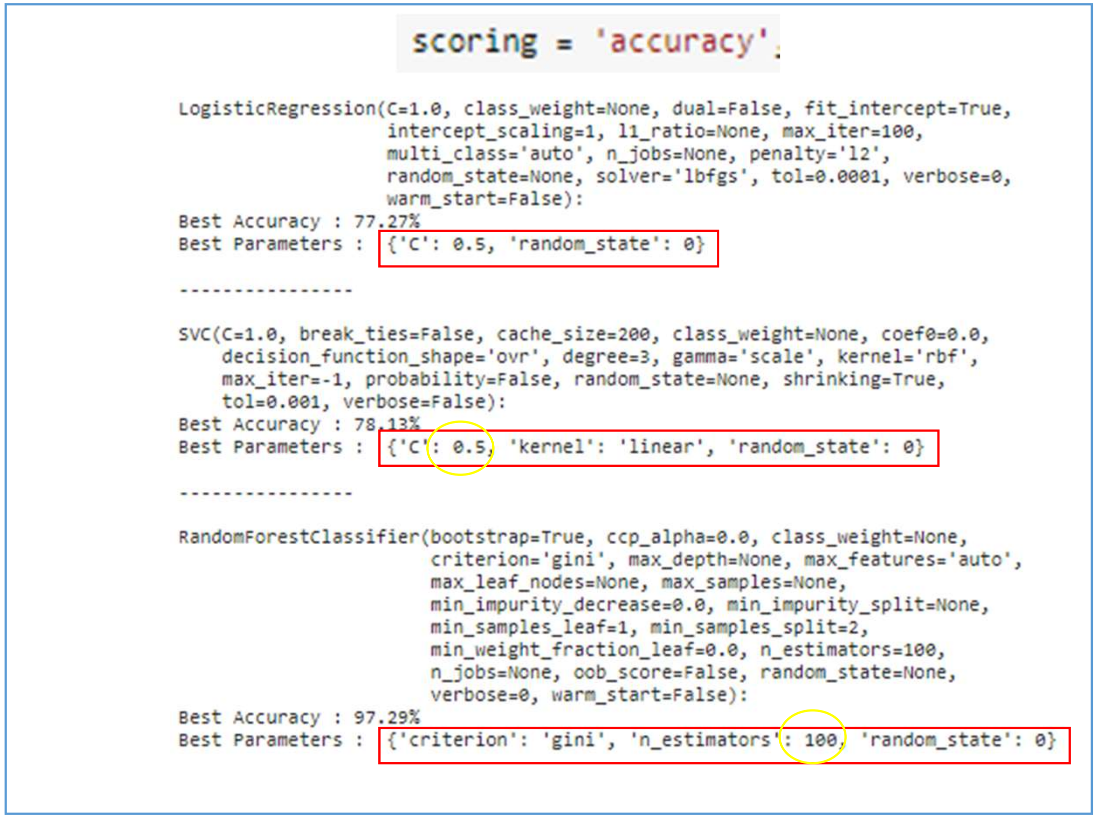
```
              precision    recall  f1-score   support

           0       0.95      0.99      0.97      1457
           1       0.09      0.01      0.02        75

    accuracy                           0.95      1532
   macro avg       0.52      0.50      0.50      1532
weighted avg       0.91      0.95      0.93      1532
```

	0	1
0	1447	10
1	74	1



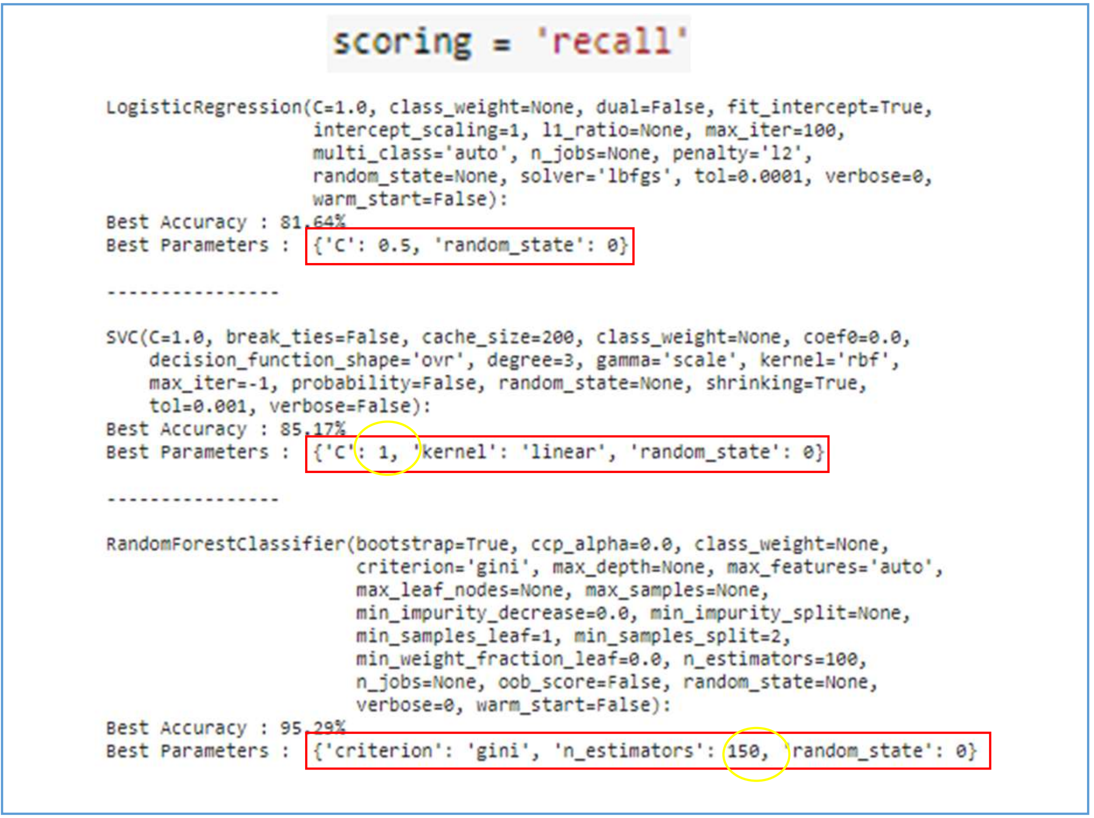
Present main results and comparison of methods

Grid Search



```
scoring = 'accuracy',
```

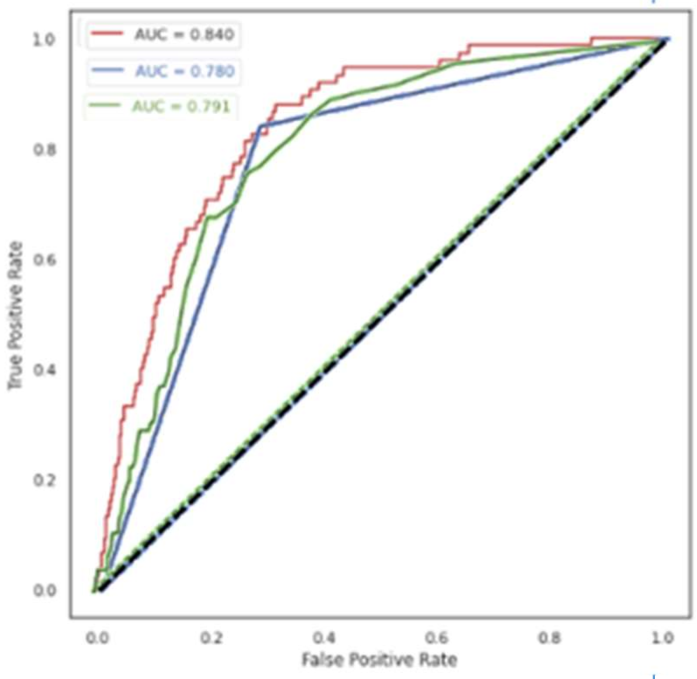
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False):
Best Accuracy : 77.27%
Best Parameters :  {'C': 0.5, 'random_state': 0}

----------------

SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False):
Best Accuracy : 78.13%
Best Parameters :  {'C': 0.5, 'kernel': 'linear', 'random_state': 0}

----------------

RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False):
Best Accuracy : 97.29%
Best Parameters :  {'criterion': 'gini', 'n_estimators': 100, 'random_state': 0}
```

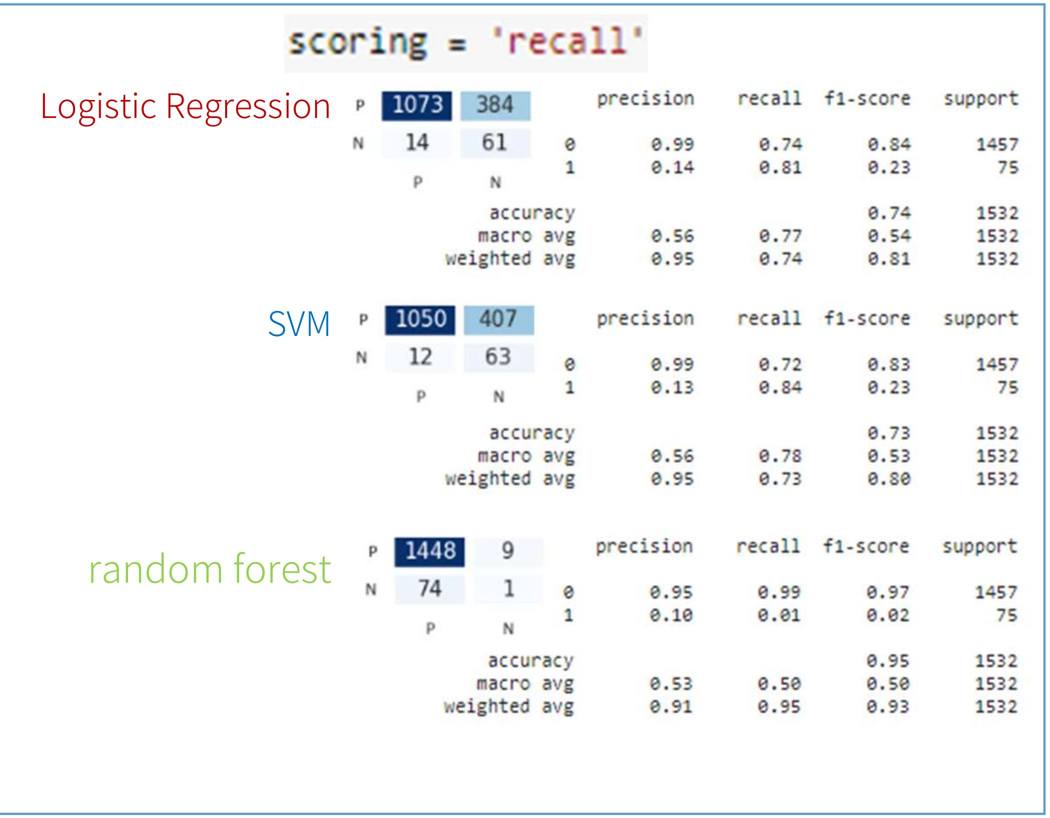


```
scoring = 'recall'
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False):
Best Accuracy : 81.64%
Best Parameters :  {'C': 0.5, 'random_state': 0}

----------------

SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False):
Best Accuracy : 85.17%
Best Parameters :  {'C': 1, 'kernel': 'linear', 'random_state': 0}

----------------

RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False):
Best Accuracy : 95.29%
Best Parameters :  {'criterion': 'gini', 'n_estimators': 150, 'random_state': 0}
```

Present main results and comparison of methods





```
scoring = 'recall'
```

Logistic Regression

```
P  1073  384
N    14   61
      P    N

              precision    recall  f1-score   support

           0       0.99      0.74      0.84      1457
           1       0.14      0.81      0.23        75

    accuracy                           0.74      1532
   macro avg       0.56      0.77      0.54      1532
weighted avg       0.95      0.74      0.81      1532
```

SVM

```
P  1050  407
N    12   63
      P    N

              precision    recall  f1-score   support

           0       0.99      0.72      0.83      1457
           1       0.13      0.84      0.23        75

    accuracy                           0.73      1532
   macro avg       0.56      0.78      0.53      1532
weighted avg       0.95      0.73      0.80      1532
```

random forest

```
P  1448    9
N    74    1
      P    N

              precision    recall  f1-score   support

           0       0.95      0.99      0.97      1457
           1       0.10      0.01      0.02        75

    accuracy                           0.95      1532
   macro avg       0.53      0.50      0.50      1532
weighted avg       0.91      0.95      0.93      1532
```

summary and Future research

- Scaling VS Threshold
- Defining additional / other variables in Grid Search
- Examination of additional models
- Combine scores and determining weight for each of them
- ensemble methods

Thanks

Tomer badug

Shirli miller

Judi Eliya