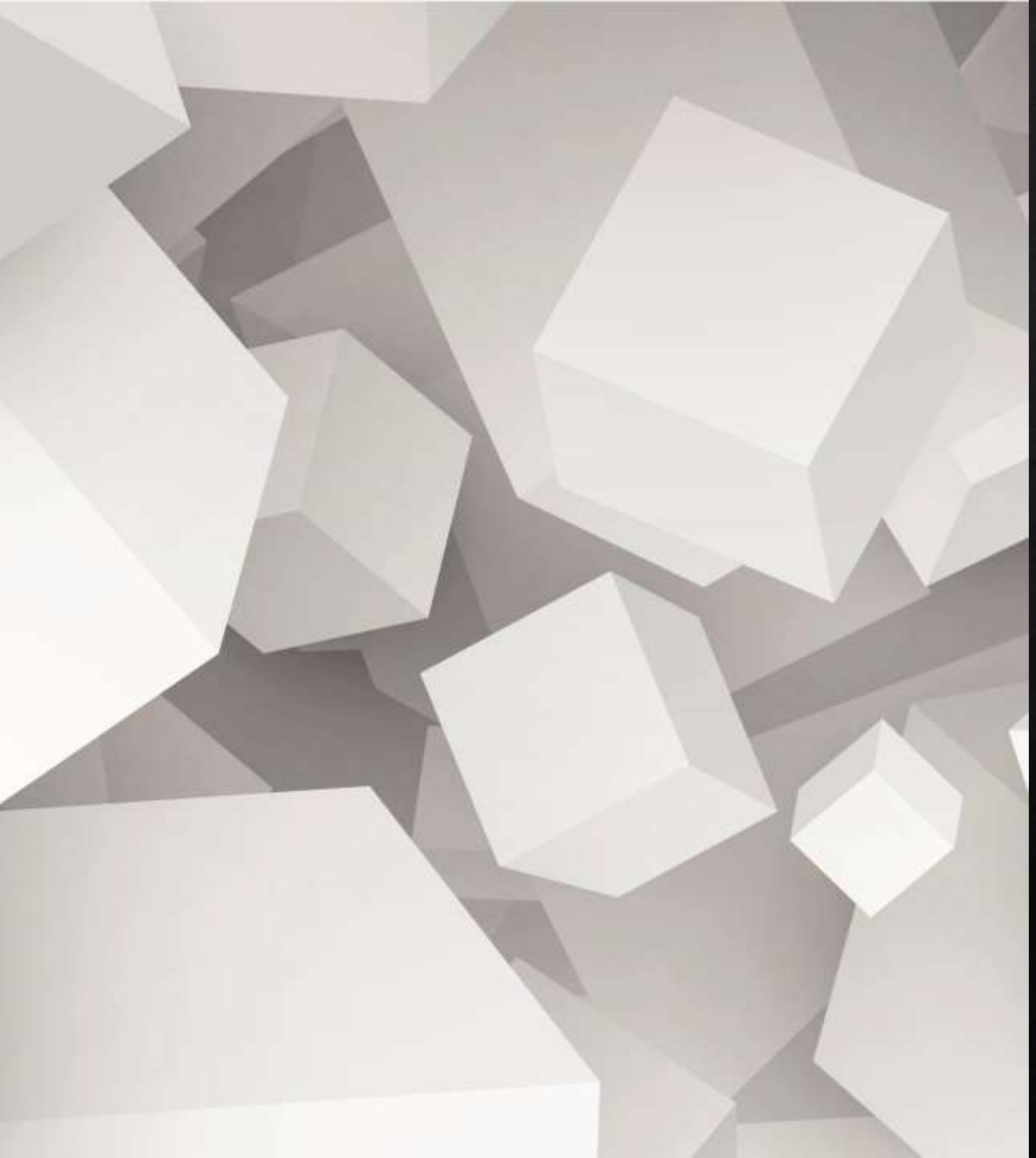


KSI Project

Group 7

Members:

- Ronald Saenz Huerta
- Ripudaman
- Karan Maria
- Manipal Sidhu
- Mahpara Rafia Radmy

Objective

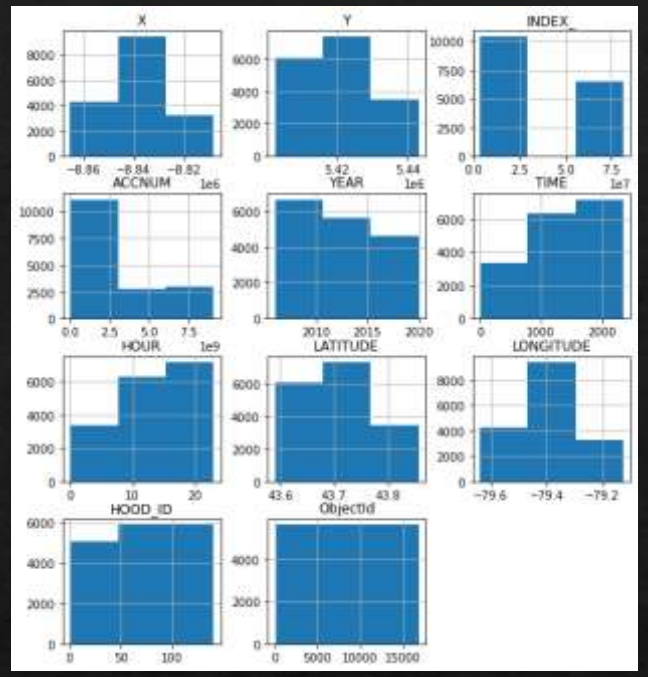
- Main objective of this project is to predict the condition which is responsible for a fatal accident
- There are various types of features present in the dataset like the physical and environmental condition of the accident location, geometric coordinate, collision vehicle, driver condition, time of day, and many more which can play an important role in defending the probability of survival

Data modeling

- We check for Null values
- We have checked the data set if it's balanced and found it to be highly imbalanced
- We've extracted the months and years from the date to find seasonal trends
- Removed duplicate columns like neighborhood and Division
- We've assumed that columns with very few values will not affect our model much
- They were too many classes in some columns which were grouped to improve performance

Data Exploration Features

```
Data Shape
(16860, 57)
```




```
Column: INVTYPE  - Len: 19  - Values:  ['Driver' 'Pedestrian' 'Motorcycle Driver' 'Passenger' 'Vehicle Owner'
 'Other Property Owner' 'Other' 'Cyclist' 'Truck Driver'
 'Motorcycle Passenger' nan 'Driver - Not Hit' 'In-Line Skater'
 'Moped Driver' 'Wheelchair' 'Pedestrian - Not Hit' 'Trailer Owner'
 'Witness' 'Cyclist Passenger']

Driver                    7051
Pedestrian                2794
Passenger                 1867
Vehicle Owner             1404
Cyclist                    724
Motorcycle Driver          605
Truck Driver               314
Other Property Owner       222
Other                      152
Motorcycle Passenger        32
Moped Driver                27
Driver - Not Hit            17
Wheelchair                  13
In-Line Skater               5
Trailer Owner                2
Cyclist Passenger            2
Pedestrian - Not Hit         1
Witness                      1
Name: INVTYPE, dtype: int64
```

```
Column: LIGHT  - Len: 9  - Values:  ['Daylight' 'Dark' 'Dawn, artificial' 'Dusk, artificial' 'Dusk'
 'Dark, artificial' 'Dawn' 'Daylight, artificial' 'Other']

Daylight              8783
Dark                  3201
Dark, artificial      2572
Dusk                   214
Dusk, artificial       164
Daylight, artificial   121
Dawn                    97
Dawn, artificial        87
Other                    6
Name: LIGHT, dtype: int64
```

Clean values

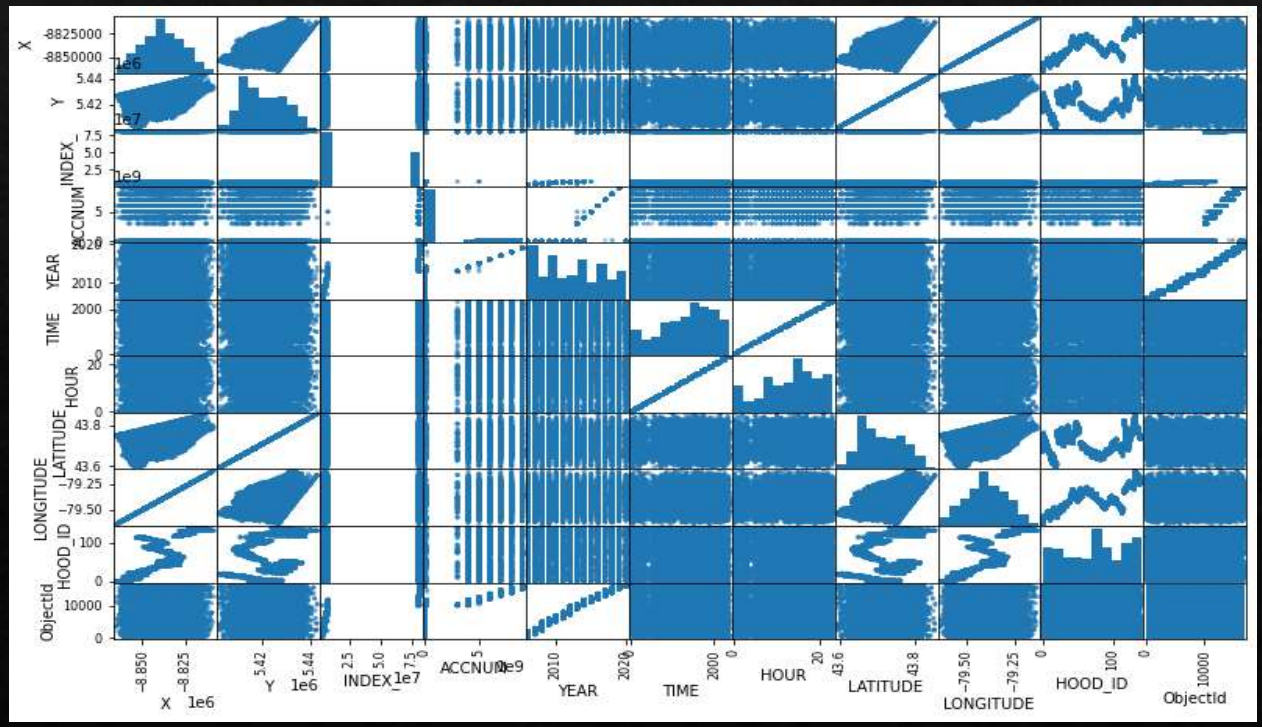
Null Values

```
Data - Null Values
X                  0
Y                  0
INDEX_             0
ACCNUM             0
YEAR               0
DATE               0
TIME               0
HOUR               0
STREET1            0
STREET2         1510
OFFSET         14114
ROAD_CLASS       497
DISTRICT         141
WARDNUM          196
DIVISION         196
LATITUDE           0
LONGITUDE          0
LOCCOORD         105
ACCLOC          5450
TRAFFCTL          29
VISIBILITY        18
LIGHT              0
RDSFCOND          23
ACCLASS            0
IMPACTYPE          4
INVTYPE           12
INVAGE             0
INJURY          1612
FATAL_NO       16147
INITDIR         4894
VEHTYPE         2813
MANOEUVER       7233
DRIVACT         8398
DRIVCOND        8396
PEDTYPE        14074
PEDACT         14081
PEDCOND        14025
CYCLISTYPE     16160
CYCACT         16153
CYCCOND        16154
```

Plots

Distribution with Histograms

Correlation with Scatter Plot



Data Exploration Features

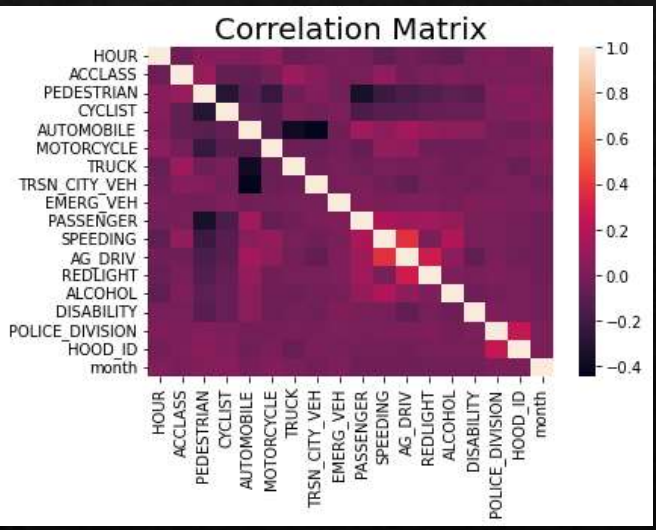
Stats to analyze mean, standard deviation

Index	X	Y	INDEX	ACCNUM	YEAR	TIME	HOUR	LATITUDE	LONGITUDE	HOOD_ID	ObjectId
count	16860	16860	16860	16860	16860	16860	16860	16860	16860	16860	16860
mean	-8.83827e+06	5.42082e+06	3.47522e+07	2.26346e+09	2012.26	1352.11	13.2411	43.7109	-79.3955	74.0036	8430.5
std	11594.9	8664.36	3.65713e+07	3.26384e+09	4.2528	629.727	6.30268	0.0562536	0.104159	41.4115	4867.21
min	-8.86531e+06	5.40226e+06	3.36321e+06	25301	2006	0	0	43.5903	-79.6384	1	1
25%	-8.8464e+06	5.41335e+06	5.35871e+06	1.06514e+06	2009	913	9	43.6624	-79.4686	39	4215.75
50%	-8.83836e+06	5.41964e+06	7.47428e+06	1.2932e+06	2012	1442	14	43.7033	-79.3963	77	8430.5
75%	-8.82953e+06	5.42791e+06	8.06319e+07	5.00181e+09	2016	1845	18	43.7569	-79.317	112	12645.2
max	-8.80825e+06	5.4431e+06	8.1542e+07	9.08535e+09	2020	2359	23	43.8554	-79.1259	140	16860

Index	HOUR	PEDESTRIAN	CYCLIST	AUTOMOBILE	MOTORCYCLE	TRUCK	TRSN_CITY_VEH	EMERG_VEH	PASSENGER	SPEEDING	AG_DRIV	REDLIGHT	ALCOHOL	DISABILITY	HOOD_ID	month
count	15245	15245	15245	15245	15245	15245	15245	15245	15245	15245	15245	15245	15245	15245	15245	15245
mean	13.2516	0.42368	0.113021	0.902066	0.0867826	0.0623155	0.0590357	0.00157429	0.338718	0.126861	0.506855	0.0787143	0.0418498	0.0264349	74.1303	6.82092
std	6.26679	0.494157	0.316628	0.297235	0.281525	0.241736	0.235699	0.0396473	0.473629	0.332828	0.499969	0.269301	0.200252	0.16043	41.2387	3.29805
min	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
25%	9	0	0	1	0	0	0	0	0	0	0	0	0	0	39	4
50%	14	0	0	1	0	0	0	0	0	0	1	0	0	0	77	7
75%	18	1	0	1	0	0	0	0	1	0	1	0	0	0	112	10
max	23	1	1	1	1	1	1	1	1	1	1	1	1	1	140	12

Correlation with ACCLASS

```
ACCLASS              1.000000
TRUCK                0.114711
PEDESTRIAN           0.100861
SPEEDING             0.089580
TRSN_CITY_VEH        0.048213
ALCOHOL              0.021518
HOOD_ID              0.015462
POLICE_DIVISION      0.007411
REDLIGHT            -0.000108
month               -0.001364
PASSENGER           -0.003197
DISABILITY          -0.004044
MOTORCYCLE          -0.012923
EMERG_VEH           -0.015988
AG_DRIV             -0.029194
HOUR                -0.037810
CYCLIST             -0.078454
AUTOMOBILE          -0.084198
Name: ACCLASS, dtype: float64
```

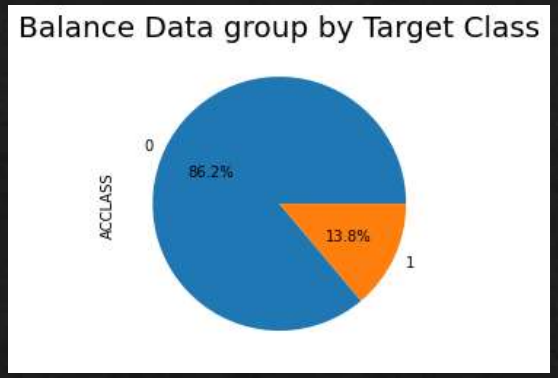


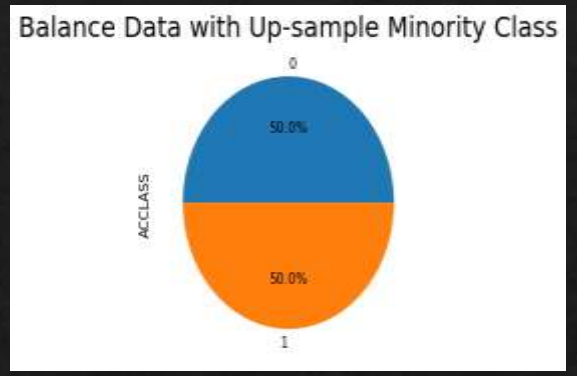
We drop those columns, and finally we worked with 28 features

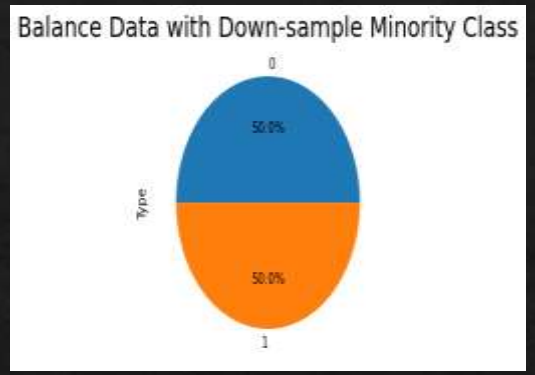
```
# Drop columns that are not required

# A lot of different values
drop_columns=['INDEX_','ObjectId','ACCNUM', 'X', 'Y','STREET1', 'STREET2', 'LATITUDE', 'LONGITUDE']
# Duplicated with HOOD_ID and POLICE_DIVISION
drop_columns+=['NEIGHBOURHOOD', 'DIVISION']
# A lot of null values
drop_columns+=["OFFSET", "PEDTYPE", "PEDACT", "PEDCOND", "CYCLISTYPE", "CYCACT", "CYCCOND", "FATAL_NO"]
# Own analysis
drop_columns+=['TIME', 'YEAR', 'DATE', 'WARDNUM', 'INITDIR', 'INVAGE', 'INJURY']
```

Imbalanced Data



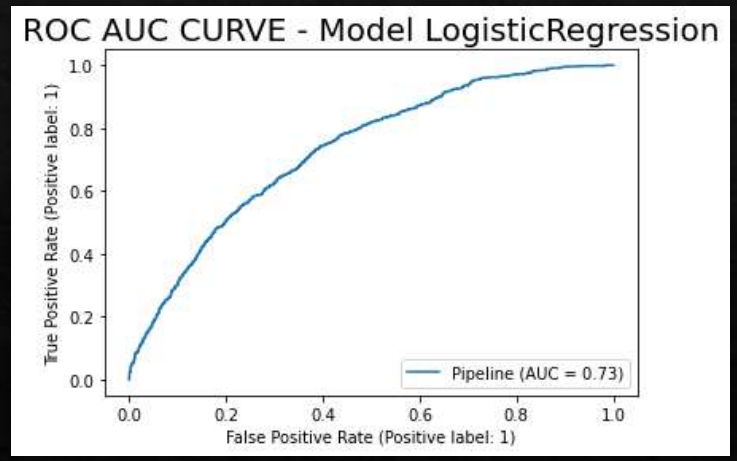


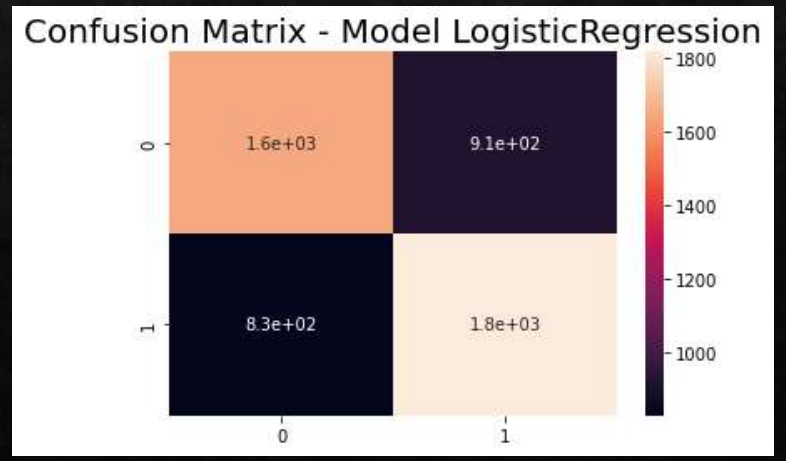


Logistic Regression Model	Imbalanced Data	Balanced with up-sample minority	Balanced with down-sample minority
Accuracy	0.8646377770426729	0.6719781907541085	0.6731963688485427
Precision	0.05207835642618251	0.6896022116418369	0.6779741997133302
Recall	0.6374269005847953	0.6661226911950152	0.6715570279223853
ROC AUC	0.7523322939754811	0.6721921264619507	0.6732121849693933
ACCLASS – VALUE 0	13022	13022	2093
ACCLASS – VALUE 1	2093	13022	2093

Logistic regression

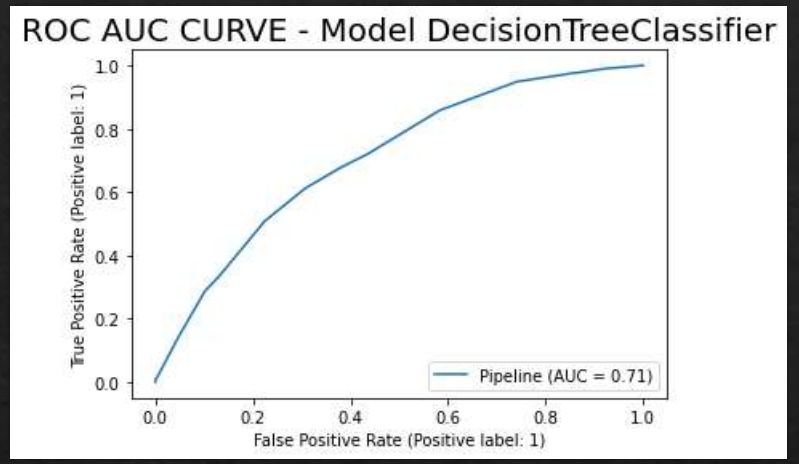
- Test Precision: 0.6658116526200073
- Test Recall: 0.6864374763883642
- Test F1 Score: 0.6759672619047619
- Test ROC AUC Score: 0.6652327897164303
- Test Accuracy Score: 0.6655788059128431

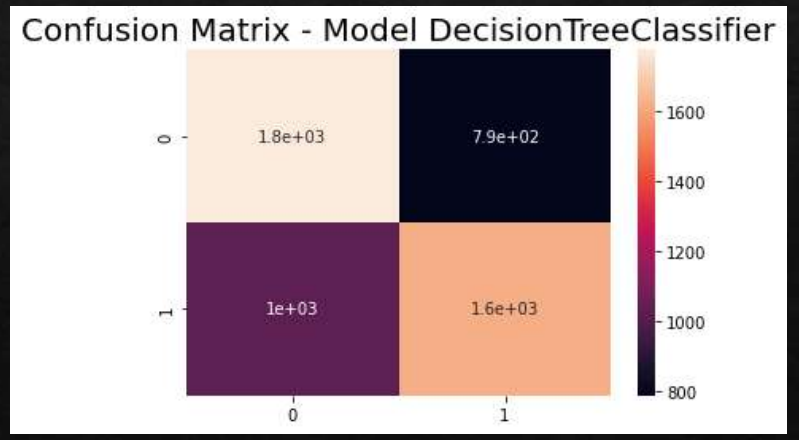




Decision tree classifier

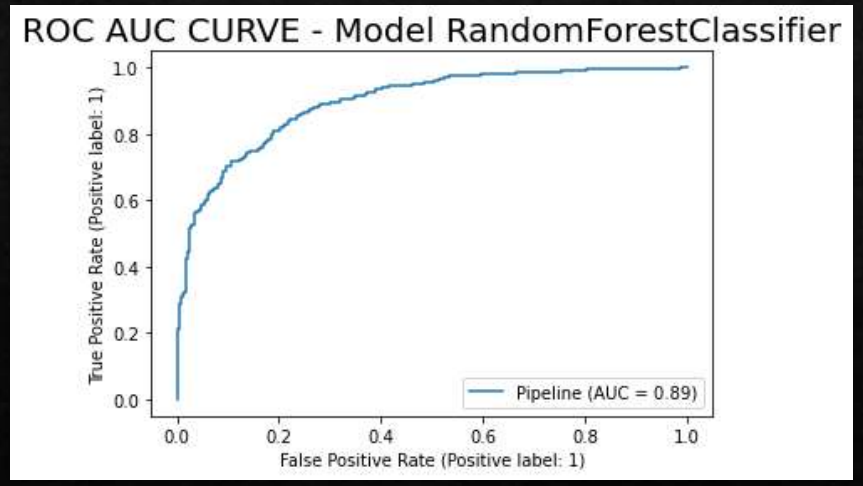
- Test Precision: 0.6730369754881596
- Test Recall: 0.612013600302229
- Test F1 Score: 0.6410763751483973
- Test ROC AUC Score: 0.6524158555765633
- Test Accuracy Score: 0.6517565751583797

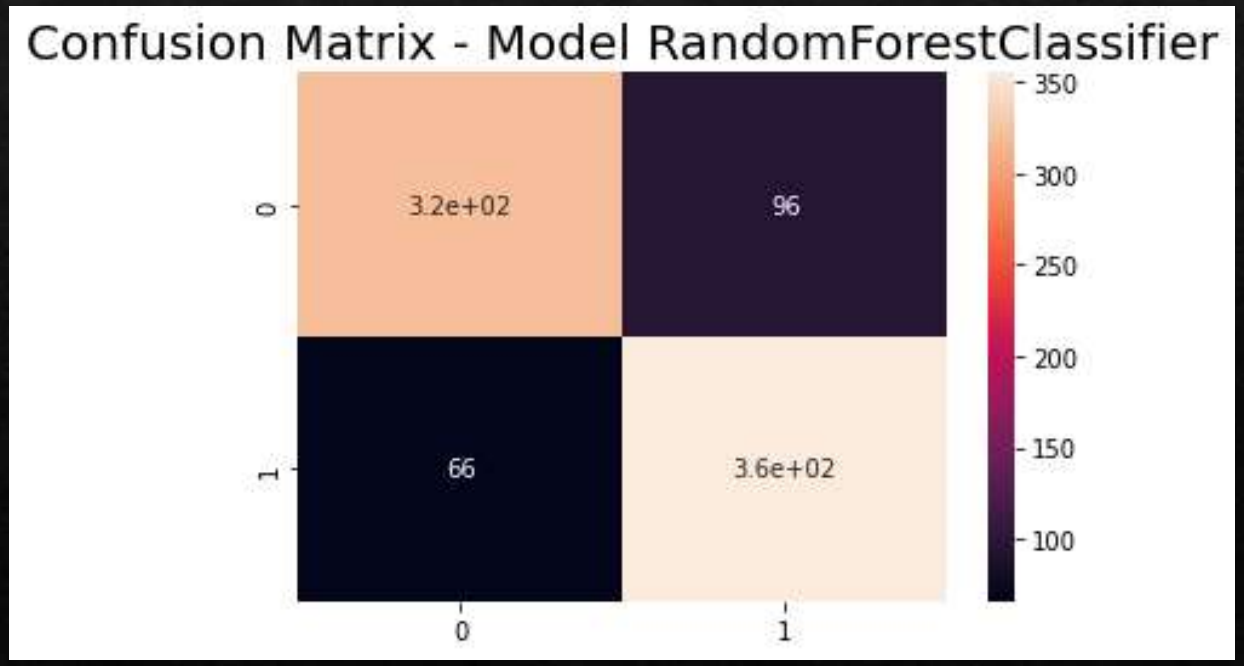




Random forest classifier

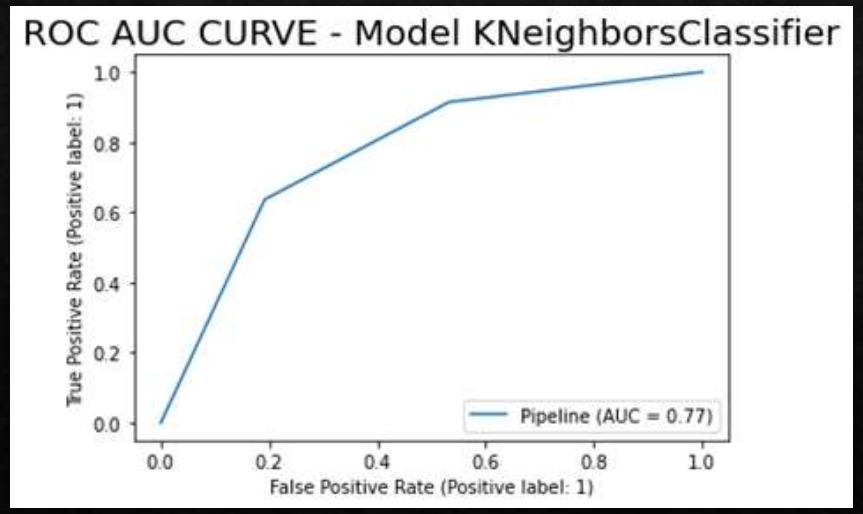
- Test Precision: 0.7871396895787139
- Test Recall: 0.8432304038004751
- Test F1 Score: 0.8142201834862386
- Test ROC AUC Score: 0.8065072882311728
- Test Accuracy Score: 0.8066825775656324

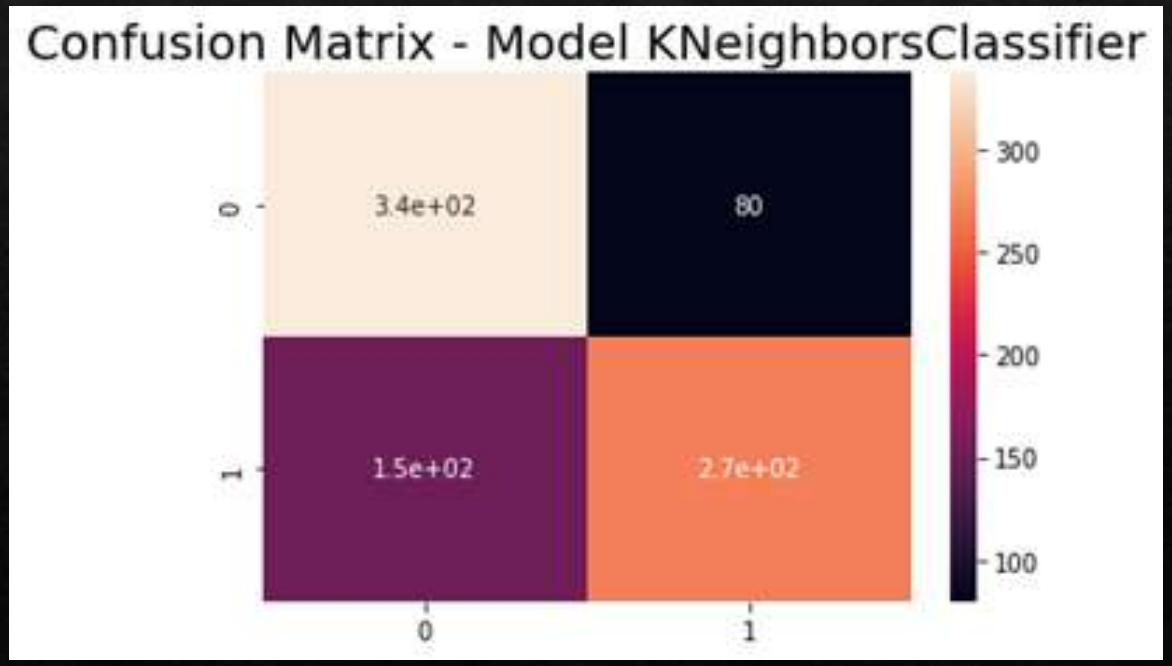




K Neighbors Classifier

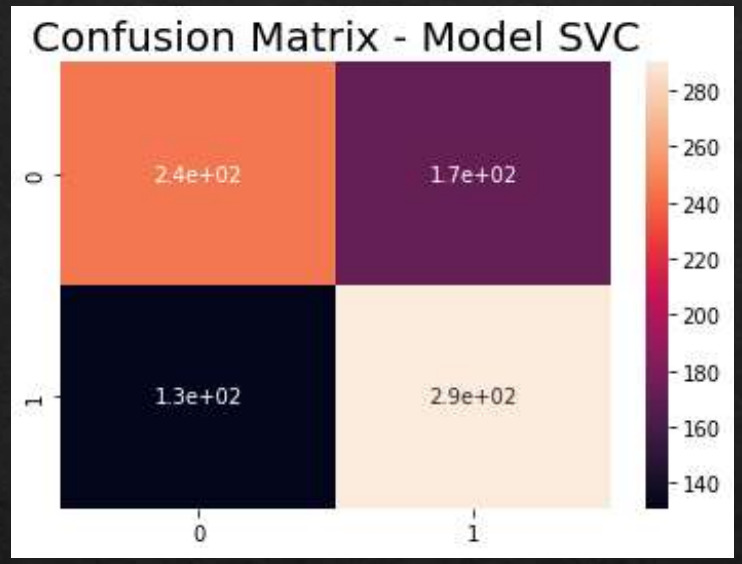
- Test Precision: 0.7701149425287356
- Test Recall: 0.6365795724465558
- Test F1 Score: 0.6970091027308193
- Test ROC AUC Score: 0.7223665248323906
- Test Accuracy Score: 0.7219570405727923

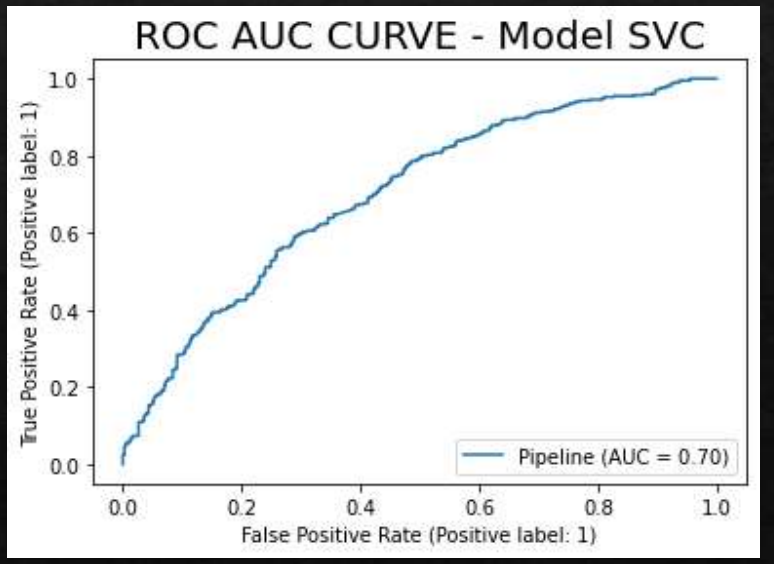




SVM

- Test Precision: 0.6277056277056277
- Test Recall: 0.6888361045130641
- Test F1 Score: 0.6568516421291052
- Test ROC AUC Score: 0.6381830402661244
- Test Accuracy Score: 0.6384248210023866



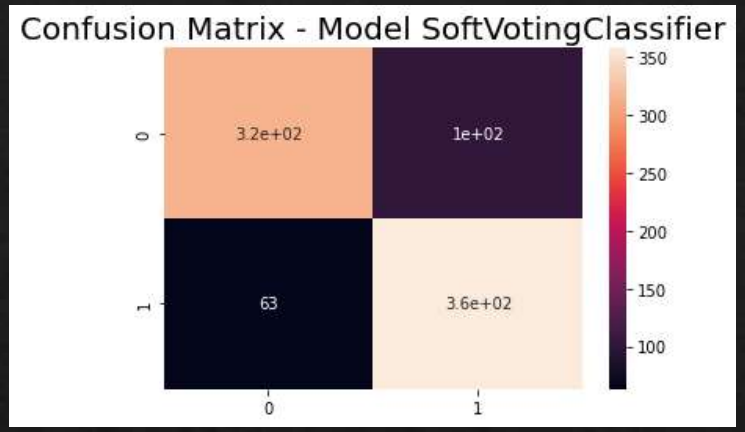


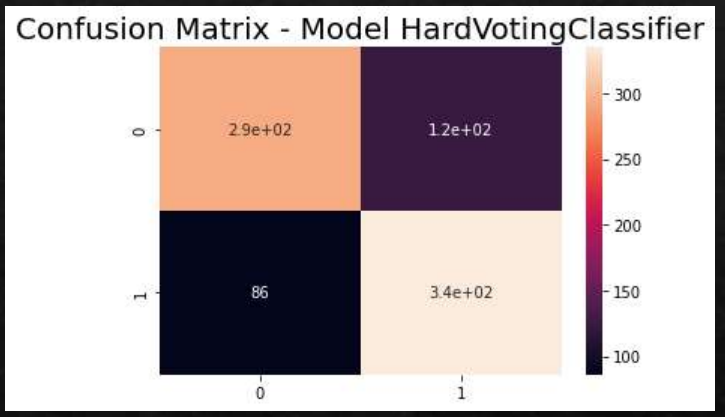
Hard Voting

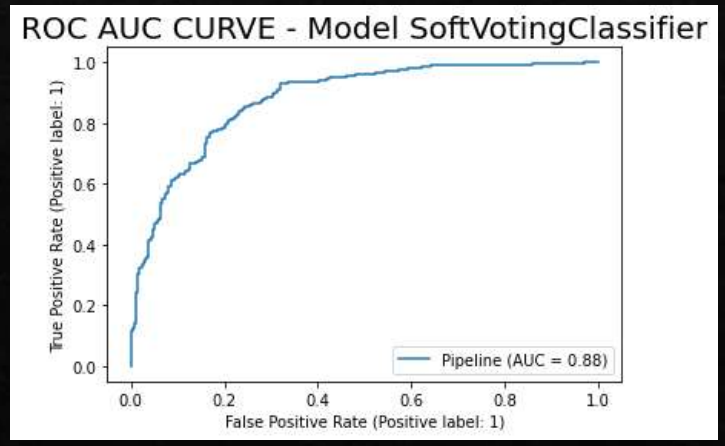
- Test Precision: 0.7314410480349345
- Test Recall: 0.7957244655581948
- Test F1 Score: 0.7622298065984072
- Test ROC AUC Score: 0.7503802183906082
- Test Accuracy Score: 0.7505966587112172

Soft Voting

- Test Precision: 0.7782608695652173
- Test Recall: 0.850356294536817
- Test F1 Score: 0.8127128263337116
- Test ROC AUC Score: 0.8028759889950272
- Test Accuracy Score: 0.8031026252983293







Model Stats

	Logistic Regression	Random Forest	Decision Tree	K Neighbors	SVC	Hard Voting	Soft Voting
Accuracy	66.55%	**80.67%**	65.18%	72.20%	63.84%	75.06%	**80.31%**
Precision	66.58%	**78.71%**	67.30%	77.01%	62.77%	73.14%	**77.83%**
Recall	68.64%	84.32%	61.20%	63.66%	68.88%	79.57%	85.04%
F1	67.60%	81.42%	64.11%	69.70%	65.69%	76.22%	81.27%
ROC AUC	66.52%	80.65%	65.24%	72.24%	63.82%	75.04%	80.29%

DEMO

- API
  - Flask
  - Python
- Website
  - HTML, JavaScript, jQuery, CSS