# Capstone Project: CAR Crash SEVERITY

Sasith Rajasooriya

# OUTLINE

- INTRODUCTION
- DATA SOURCE AND DATA SET
- PREPROCESSING
  - Data Cleaning
  - Exploratory Analysis
  - Dimensionality Reduction effort (PCA)PREPROCESSING
- SAMPLING AND BALANCED SAMPLING
- CLASSIFICATION
- SUMMARY AND CONCLUSIONS

# INTRODUCTION

- A machine learning model which is successful in predicting the severity with a great accuracy will make sure following important informative measures are taken efficiently.
- Successful machine learning model will elaborate a set of conditions (ex: temperature, wind speed, other weather conditions etc.) which are found to be highly related to accidents with higher level of severity (ex: degree 4 accidents.
- Traffic Police officials and patrolling security officials can be informed about these conditions prior and make sure additional security and road/highway support teams are allocated on these areas.

# Problem

- “What is the severity of the probable accident in the USA under a set of certain conditions (features)?
- How can we classify to train and test effective machine learning models?

# DATA SOURCE AND THE DATASET

- **Data Source**
  - In this project I use the accident data in the United States in the “US Accidents” data set available in kaggle.com.
  - The data set is maintained by its owner Sobhan Moosavi. The data set has 3513740 accident data recorded from 2016-02-01 to 2019-12-31, and further updated annually.
  - The data set is developed from “Traffic Streaming Report APIs” and collected using streaming reports on important traffic events.

# DATA SOURCE AND THE DATASET

- **Data Source**
  - In this project I use the accident data in the United States in the “US Accidents” data set available in kaggle.com.
  - The data set is maintained by its owner Sobhan Moosavi. The data set has 3513740 accident data recorded from 2016-02-01 to 2019-12-31, and further updated annually.
  - The data set is developed from “Traffic Streaming Report APIs” and collected using streaming reports on important traffic events.

# DATA SOURCE AND THE DATASET

- **Data Set**
  - The data set is of the dimension of 3513740 rows and 49 columns. Each row had measured and recorded attributes of an accident occurred. Different data types was observed as mentioned below.

# METHODOLOGY
# PREPROCESSING :Cleaning the data set

- **Removing irrelevant and unusable variables.**
  - Ex: Start_Time
  - City
  - Description
- **Large portion of values are missing**
  - Ex: Windchill / Precepitation
- **To avoid Multicolinearity**
  - Ex: Sunrise_Sunset

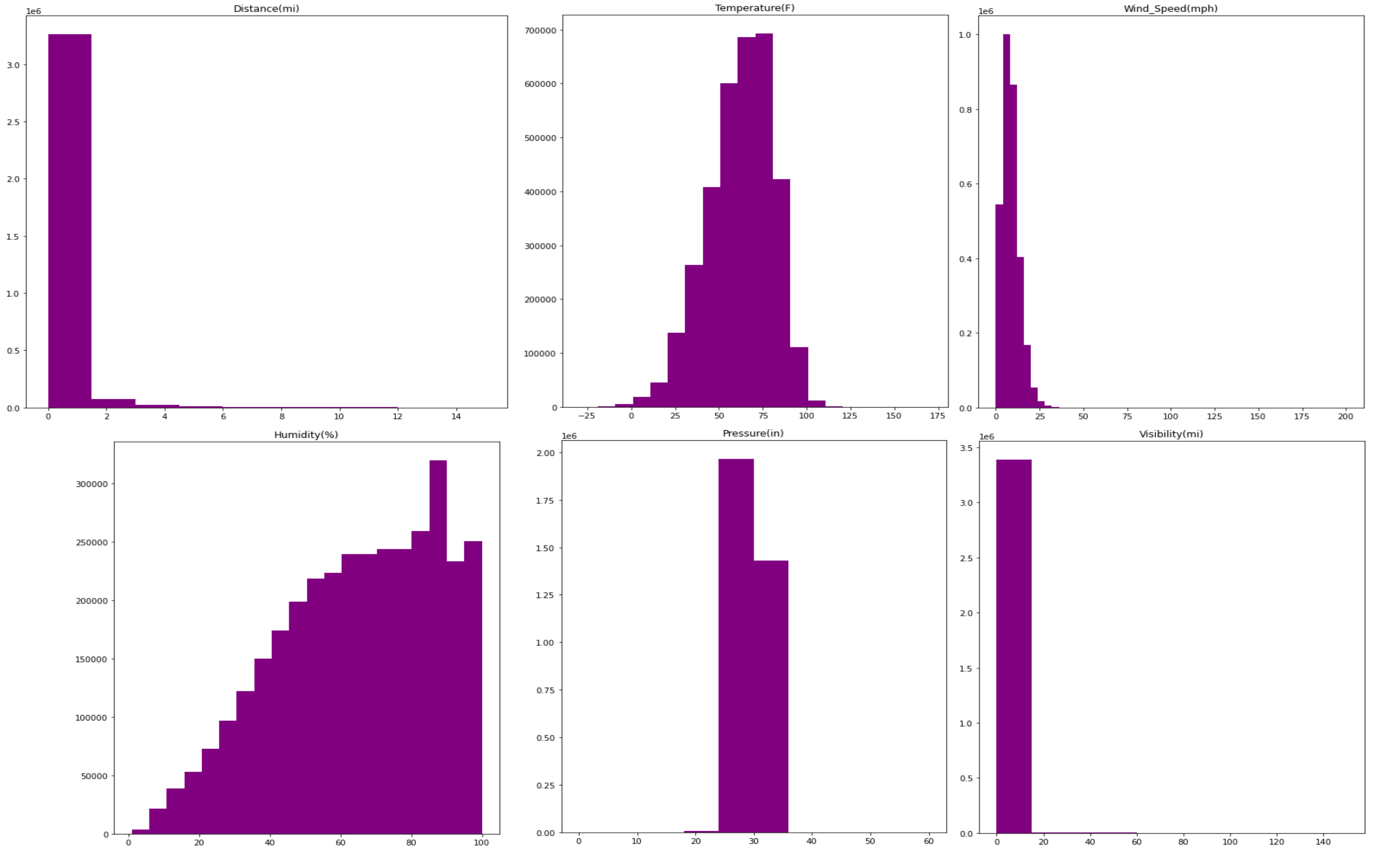
# PREPROCESSING :
# Dealing with the Missing data

- Out of the 21 columns, 9 columns did not have any missing values, which is very good. Other 12 columns had missing values that we have to deal with. After the removal of irrelevant variables, only 10 left with the missing values to deal with.
  - Variables “Precipitation” and “Wind_Chill” were missing a large portion of data values.  These variables are dropped
  - Attribute Wind_speed is also missing over 10% of the values recorded. We will replace these missing values by the average wind speed.
  - Other records with missing attributes values will be removed.

# PREPROCESSING :DATA ENTRY ERRORS

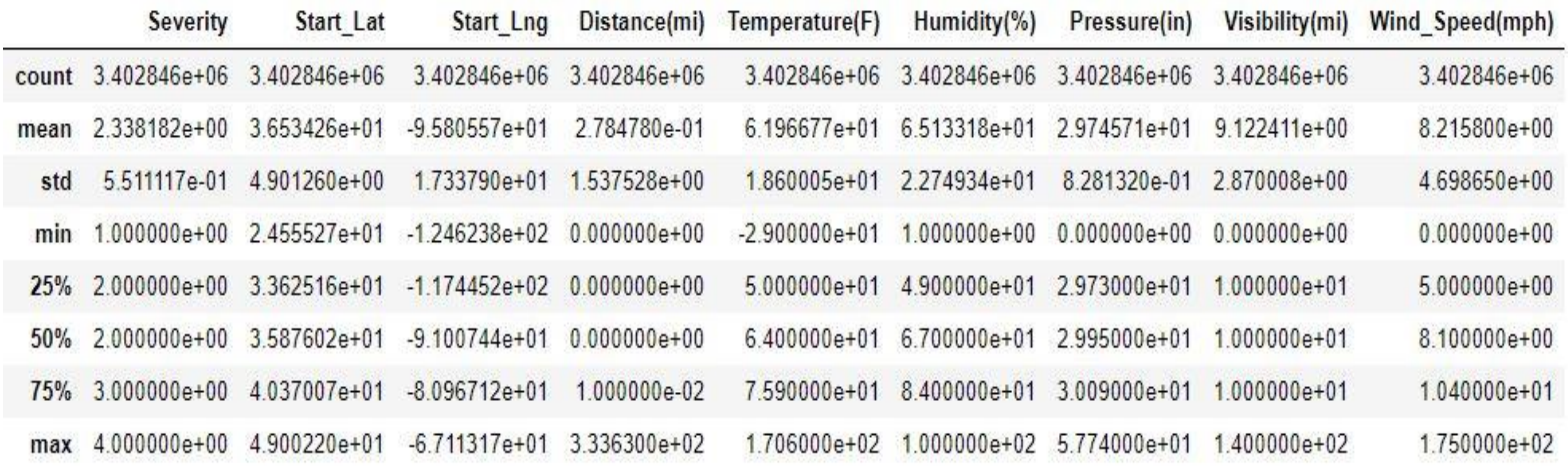
- Windspeed
  - As per real records, the highest recorded Wind Speed in the history of US is around 231 mph, which is in 1984. Therefore, keeping that in mind, we removed accident records with “Wind speed” recorded over 200 mph for better accuracy

# PREPROCESSING :Exploratory analysis



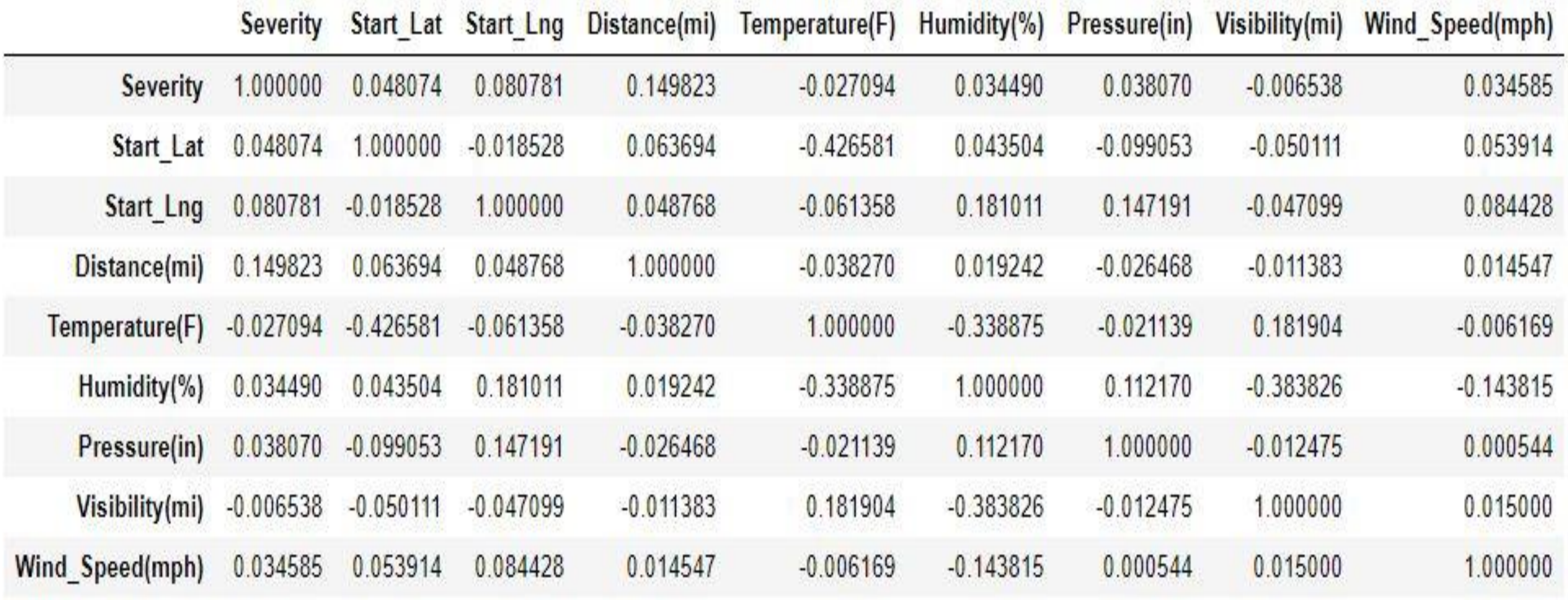
Distribution of Quantitative Variables

# PREPROCESSING :Summaries



| | Severity | Start_Lat | Start_Lng | Distance(mi) | Temperature(F) | Humidity(%) | Pressure(in) | Visibility(mi) | Wind_Speed(mph) |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 3.402846e+06 | 3.402846e+06 | 3.402846e+06 | 3.402846e+06 | 3.402846e+06 | 3.402846e+06 | 3.402846e+06 | 3.402846e+06 | 3.402846e+06 |
| **mean** | 2.338182e+00 | 3.653426e+01 | -9.580557e+01 | 2.784780e-01 | 6.196677e+01 | 6.513318e+01 | 2.974571e+01 | 9.122411e+00 | 8.215800e+00 |
| **std** | 5.511117e-01 | 4.901260e+00 | 1.733790e+01 | 1.537528e+00 | 1.860005e+01 | 2.274934e+01 | 8.281320e-01 | 2.870008e+00 | 4.698650e+00 |
| **min** | 1.000000e+00 | 2.455527e+01 | -1.246238e+02 | 0.000000e+00 | -2.900000e+01 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| **25%** | 2.000000e+00 | 3.362516e+01 | -1.174452e+02 | 0.000000e+00 | 5.000000e+01 | 4.900000e+01 | 2.973000e+01 | 1.000000e+01 | 5.000000e+00 |
| **50%** | 2.000000e+00 | 3.587602e+01 | -9.100744e+01 | 0.000000e+00 | 6.400000e+01 | 6.700000e+01 | 2.995000e+01 | 1.000000e+01 | 8.100000e+00 |
| **75%** | 3.000000e+00 | 4.037007e+01 | -8.096712e+01 | 1.000000e-02 | 7.590000e+01 | 8.400000e+01 | 3.009000e+01 | 1.000000e+01 | 1.040000e+01 |
| **max** | 4.000000e+00 | 4.900220e+01 | -6.711317e+01 | 3.336300e+02 | 1.706000e+02 | 1.000000e+02 | 5.774000e+01 | 1.400000e+02 | 1.750000e+02 |

# PREPROCESSING : No strong Correlation



| | Severity | Start_Lat | Start_Lng | Distance(mi) | Temperature(F) | Humidity(%) | Pressure(in) | Visibility(mi) | Wind_Speed(mph) |
|---|---|---|---|---|---|---|---|---|---|
| Severity | 1.000000 | 0.048074 | 0.080781 | 0.149823 | -0.027094 | 0.034490 | 0.038070 | -0.006538 | 0.034585 |
| Start_Lat | 0.048074 | 1.000000 | -0.018528 | 0.063694 | -0.426581 | 0.043504 | -0.099053 | -0.050111 | 0.053914 |
| Start_Lng | 0.080781 | -0.018528 | 1.000000 | 0.048768 | -0.061358 | 0.181011 | 0.147191 | -0.047099 | 0.084428 |
| Distance(mi) | 0.149823 | 0.063694 | 0.048768 | 1.000000 | -0.038270 | 0.019242 | -0.026468 | -0.011383 | 0.014547 |
| Temperature(F) | -0.027094 | -0.426581 | -0.061358 | -0.038270 | 1.000000 | -0.338875 | -0.021139 | 0.181904 | -0.006169 |
| Humidity(%) | 0.034490 | 0.043504 | 0.181011 | 0.019242 | -0.338875 | 1.000000 | 0.112170 | -0.383826 | -0.143815 |
| Pressure(in) | 0.038070 | -0.099053 | 0.147191 | -0.026468 | -0.021139 | 0.112170 | 1.000000 | -0.012475 | 0.000544 |
| Visibility(mi) | -0.006538 | -0.050111 | -0.047099 | -0.011383 | 0.181904 | -0.383826 | -0.012475 | 1.000000 | 0.015000 |
| Wind_Speed(mph) | 0.034585 | 0.053914 | 0.084428 | 0.014547 | -0.006169 | -0.143815 | 0.000544 | 0.015000 | 1.000000 |

Chi-Square test results were statistically significant.
p-value: [0. 0.]Significance level: 0.05Degree of freedom: 125

# PREPROCESSING :*Chi-Square Test for independence: Drop Sunrise_sunset !*

- “Weather condition” and the “Sunrise_Sunset”
  - Test results : Not independent

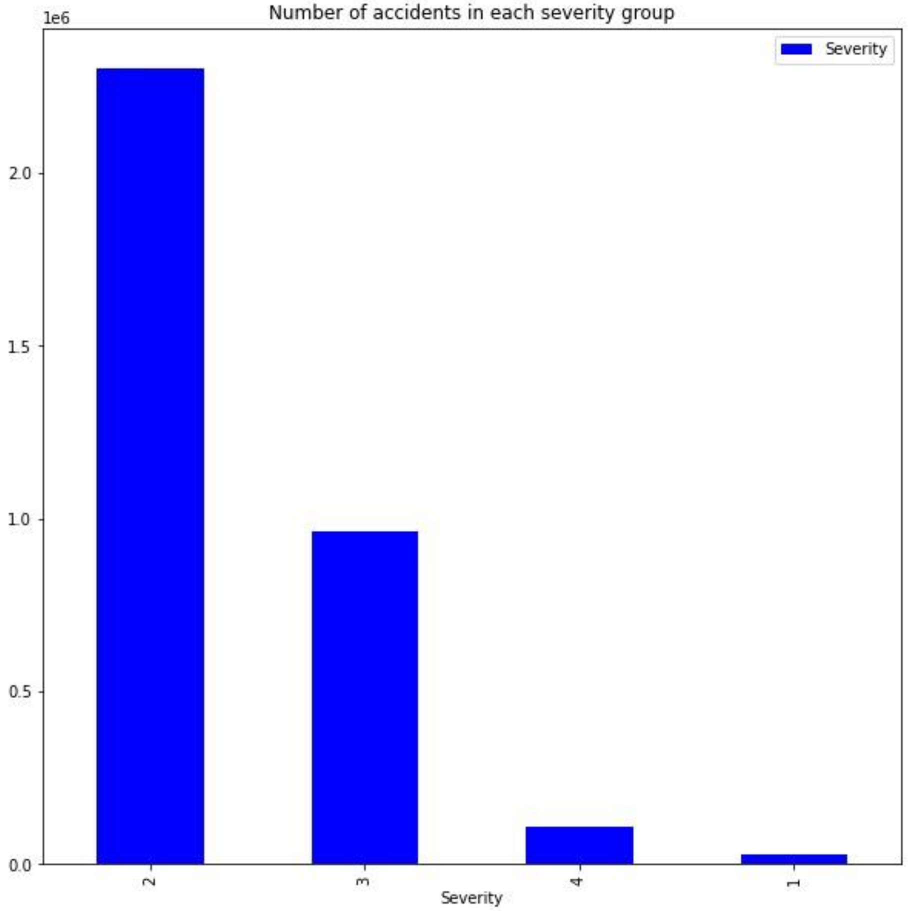
    Chi-Square test results were statistically significant.

    p-value: [0. 0.]

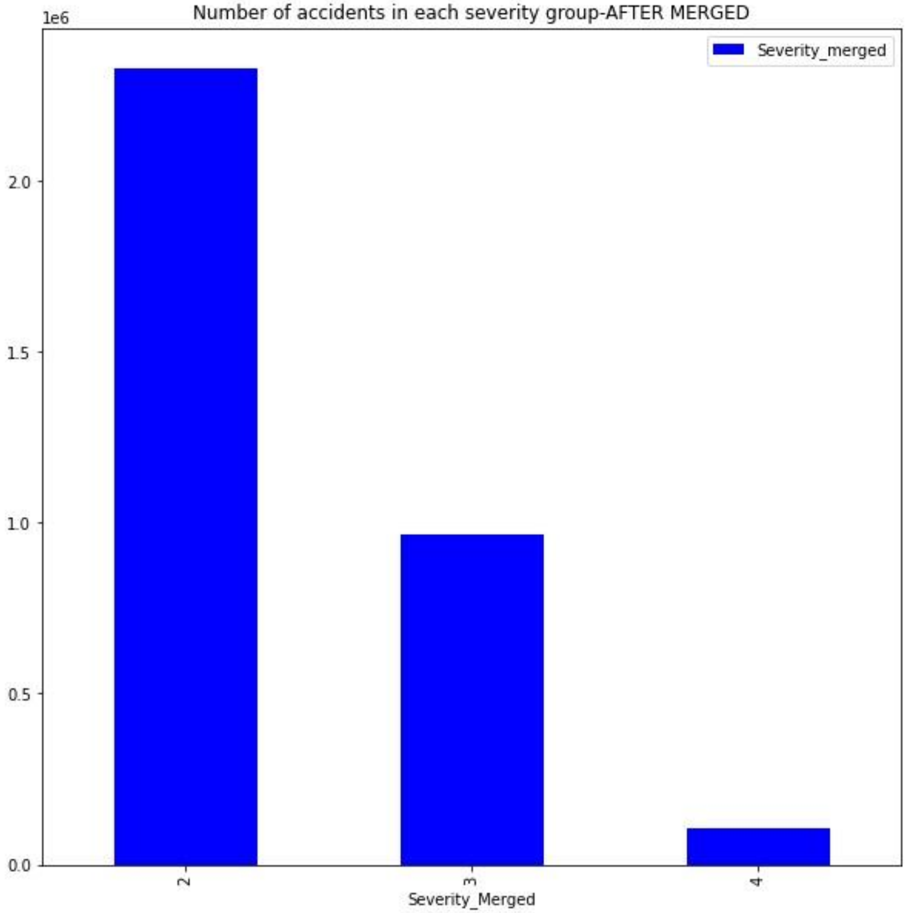
    Significance level: 0.05

    Degree of freedom: 125

# Merge Severity class: Severity 4 priority



Number of accidents in each severity group



Number of accidents in each severity group-AFTER MERGED

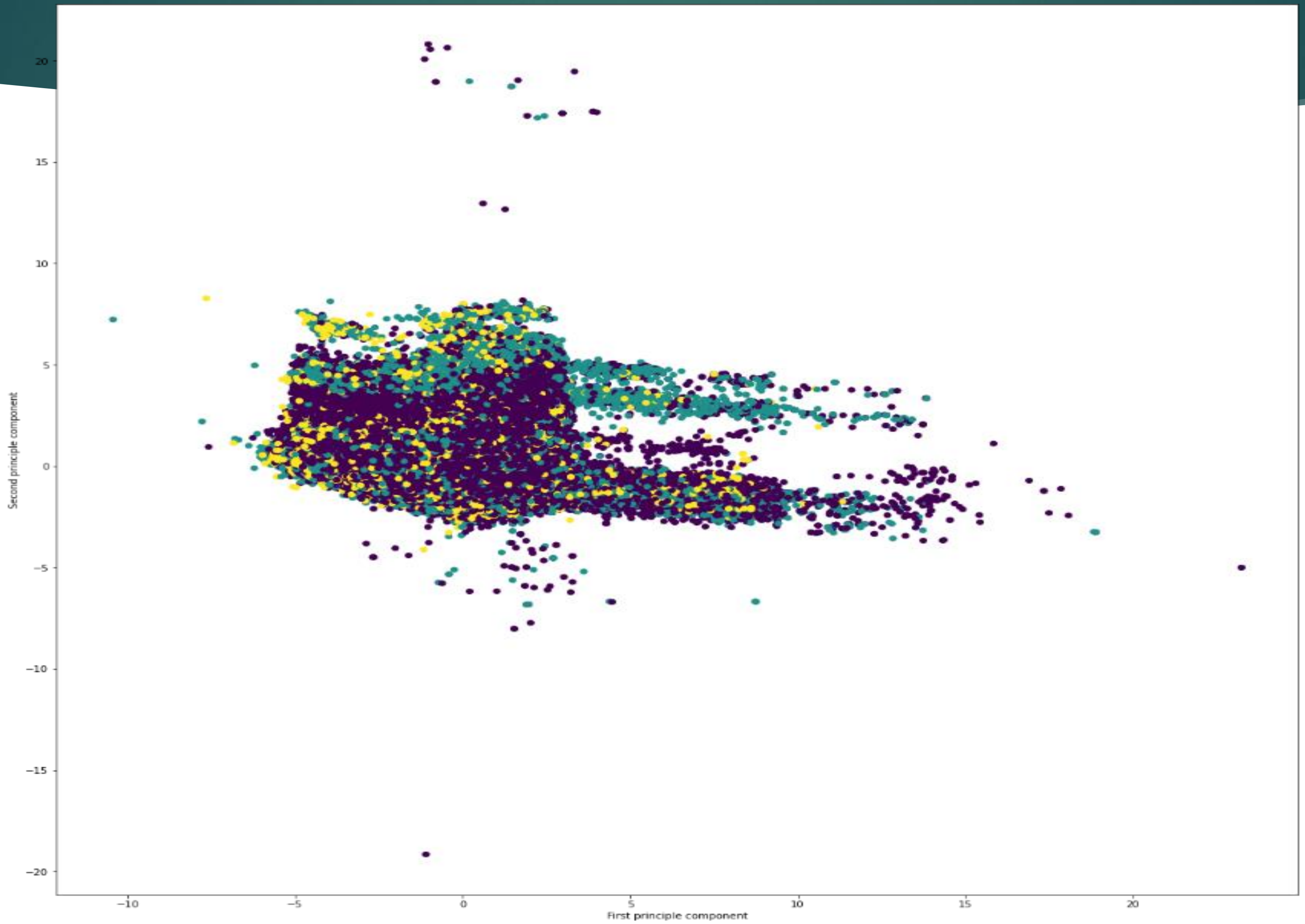
# Target Feature

| Severity | Count | % |
|---|---|---|
| 1 | 28540 | 0.008387 |
| 2 | 2302403 | 0.676611 |
| 3 | 964484 | 0.283435 |
| 4 | 107419 | 0.031567 |

# PREPROCESSING : categorical feature "Weather_Condition"

- “Weather condition” we have in our data frame has 126 different categories (levels).
- Out of 126 categories only 8 categories have occurred at frequencies higher than 100,000.
- Therefore, this column was re-organized by bringing down from 126 to 20 levels with the highest frequencies
- all other conditions in to one category called "Other Weather Conditions".

# PREPROCESSING : Dimensionality Reduction effort (PCA)



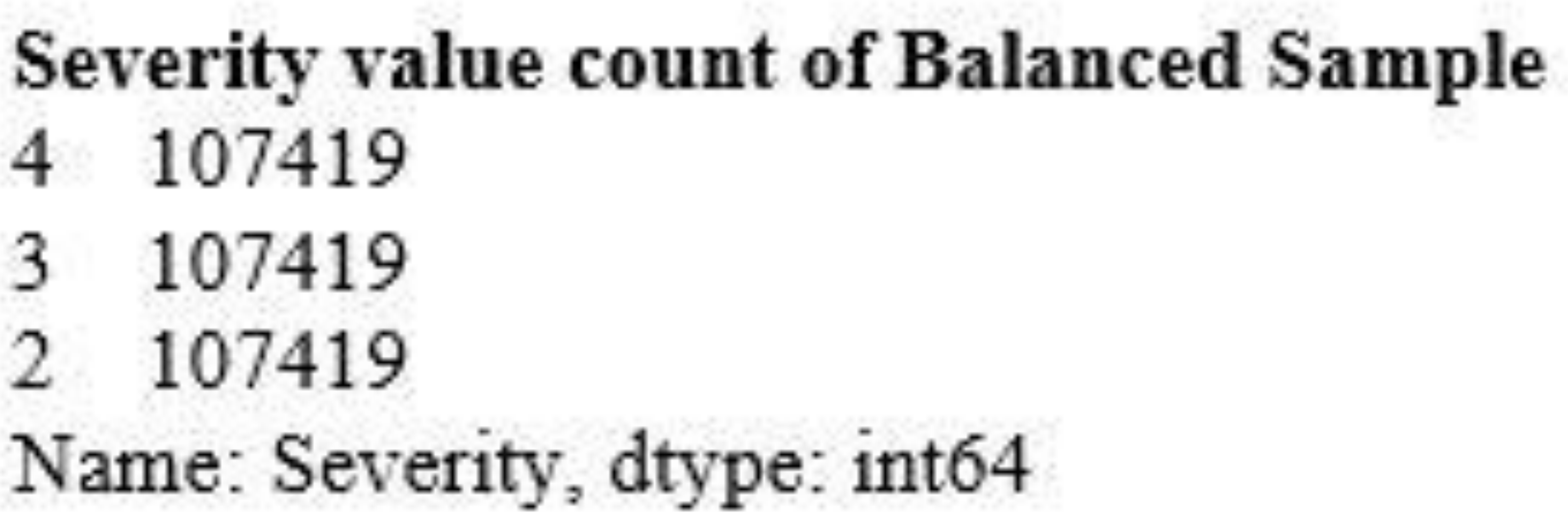
# PREPROCESSING :PCA not Success

- REDUCING THE DIMENTIONALITY of the accidents severity feature set. Explained variances did not add up to at least 95% of the variance (first component about 81.6% and the second component 5.6% didn’t add up to at least 95% of the variation).

# METHODOLOGY
# SAMPLING AND BALANCING THE SAMPLES

- **Balancing the Data**
- data set is not a balanced one. It is mostly observed that accident severity data is highly imbalanced. Therefore, it was needed to take steps to make the data set balanced using steps such as Random "Under-sampling" from majority subsets.
- **Making a Main Random Sample**
- For processing convenience, first a random sample of the size 15000 was obtained from our cleaned data frame. Taking this step was compelled since when the entire data set or a larger sample was run without high power cloud computing or similar techniques, the Kernel did not run due to memory limitations

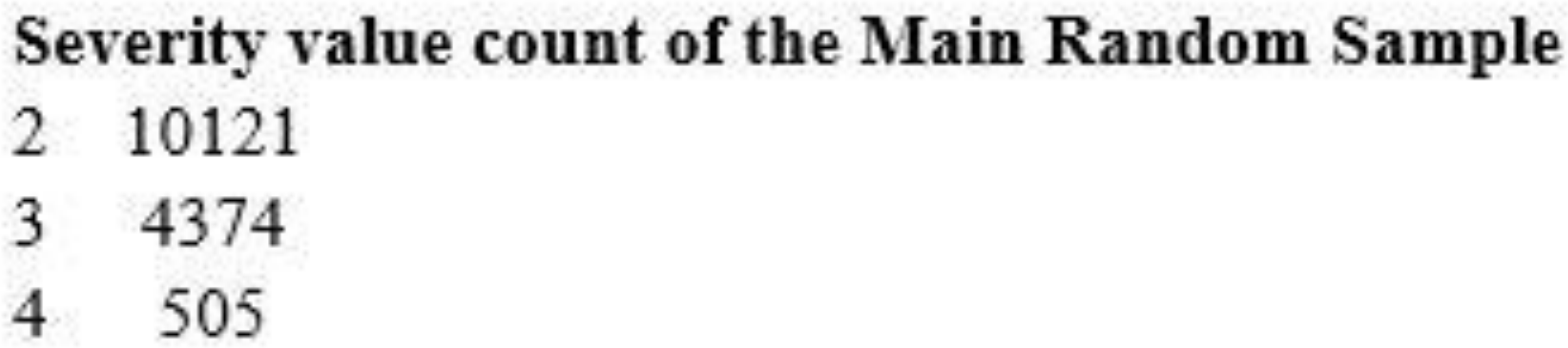
# SAMPLING AND BALANCING THE SAMPLES



```
Severity value count of Balanced Sample
4    107419
3    107419
2    107419
Name: Severity, dtype: int64
```

# SAMPLING AND BALANCING THE SAMPLES

- MAIN SAMPLE



**Severity value count of the Main Random Sample**

| | |
|---|---|
| 2 | 10121 |
| 3 | 4374 |
| 4 | 505 |

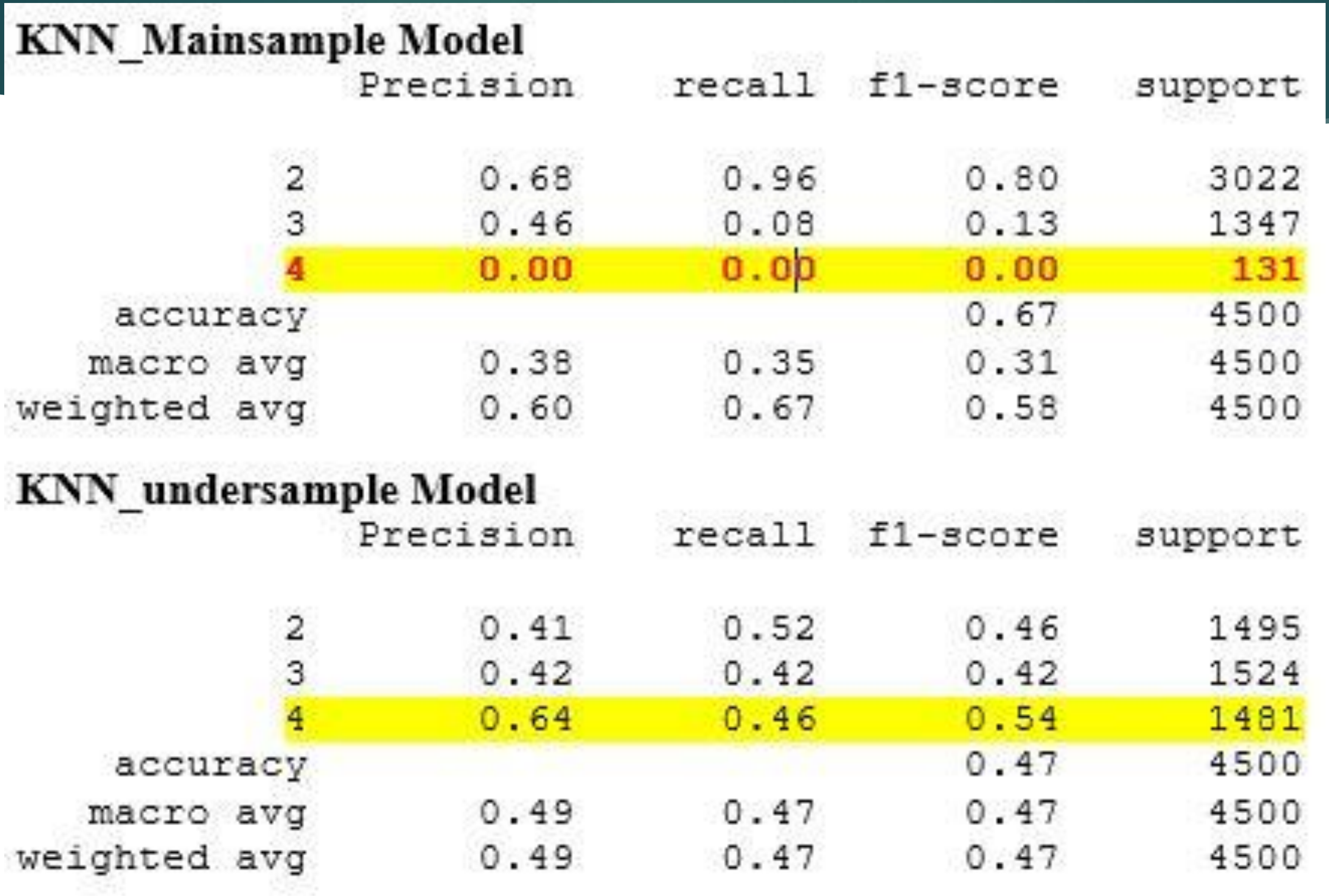
# TRAIN TEST SPLIT

- Train set – 70%
- Test Set – 30%

# METHODOLOGY : CLASSIFICATION

- KNN
- SVM
- Random Forrest

# RESULTS : CLASSIFICATION : KNN



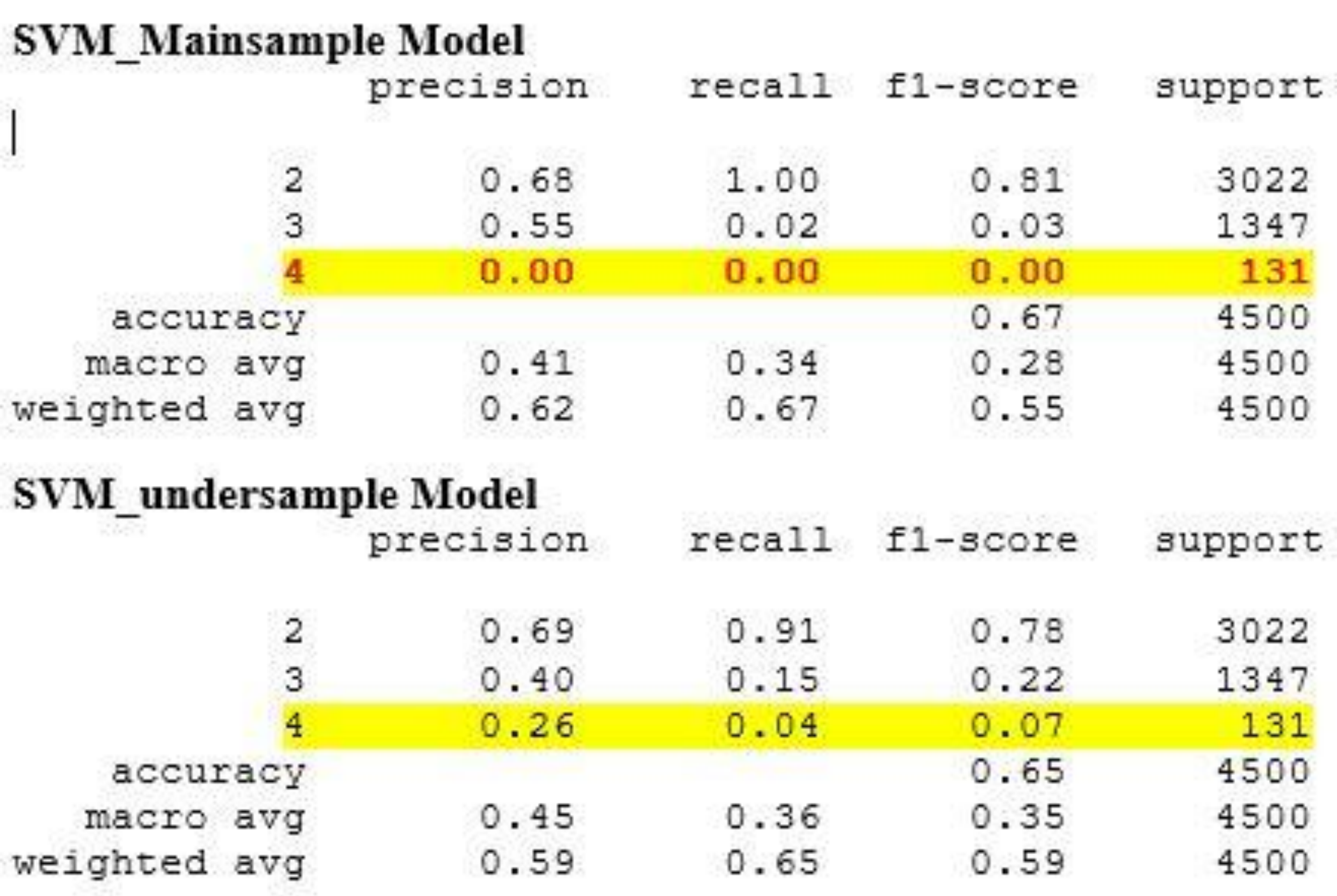
**KNN_Mainsample Model**

```
              Precision    recall  f1-score   support

           2       0.68      0.96      0.80      3022
           3       0.46      0.08      0.13      1347
           4       0.00      0.00      0.00       131
    accuracy                           0.67      4500
   macro avg       0.38      0.35      0.31      4500
weighted avg       0.60      0.67      0.58      4500
```

**KNN_undersample Model**

```
              Precision    recall  f1-score   support

           2       0.41      0.52      0.46      1495
           3       0.42      0.42      0.42      1524
           4       0.64      0.46      0.54      1481
    accuracy                           0.47      4500
   macro avg       0.49      0.47      0.47      4500
weighted avg       0.49      0.47      0.47      4500
```

# RESULTS : CLASSIFICATION : SVM



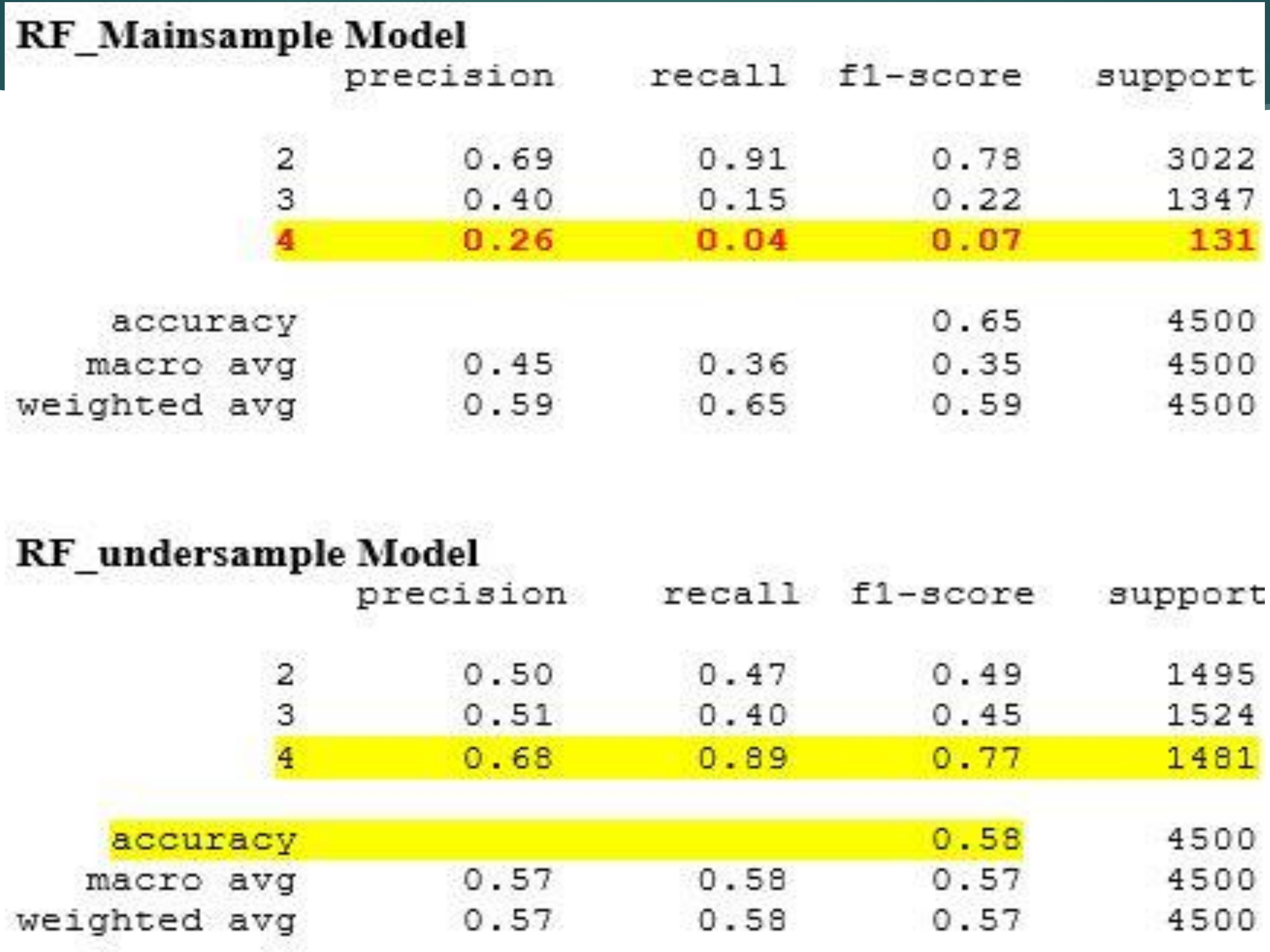
SVM_Mainsample Model

```
              precision    recall  f1-score   support

           2       0.68      1.00      0.81      3022
           3       0.55      0.02      0.03      1347
           4       0.00      0.00      0.00       131
    accuracy                           0.67      4500
   macro avg       0.41      0.34      0.28      4500
weighted avg       0.62      0.67      0.55      4500
```

SVM_undersample Model

```
              precision    recall  f1-score   support

           2       0.69      0.91      0.78      3022
           3       0.40      0.15      0.22      1347
           4       0.26      0.04      0.07       131
    accuracy                           0.65      4500
   macro avg       0.45      0.36      0.35      4500
weighted avg       0.59      0.65      0.59      4500
```

# RESULTS : CLASSIFICATION : Random Forrest

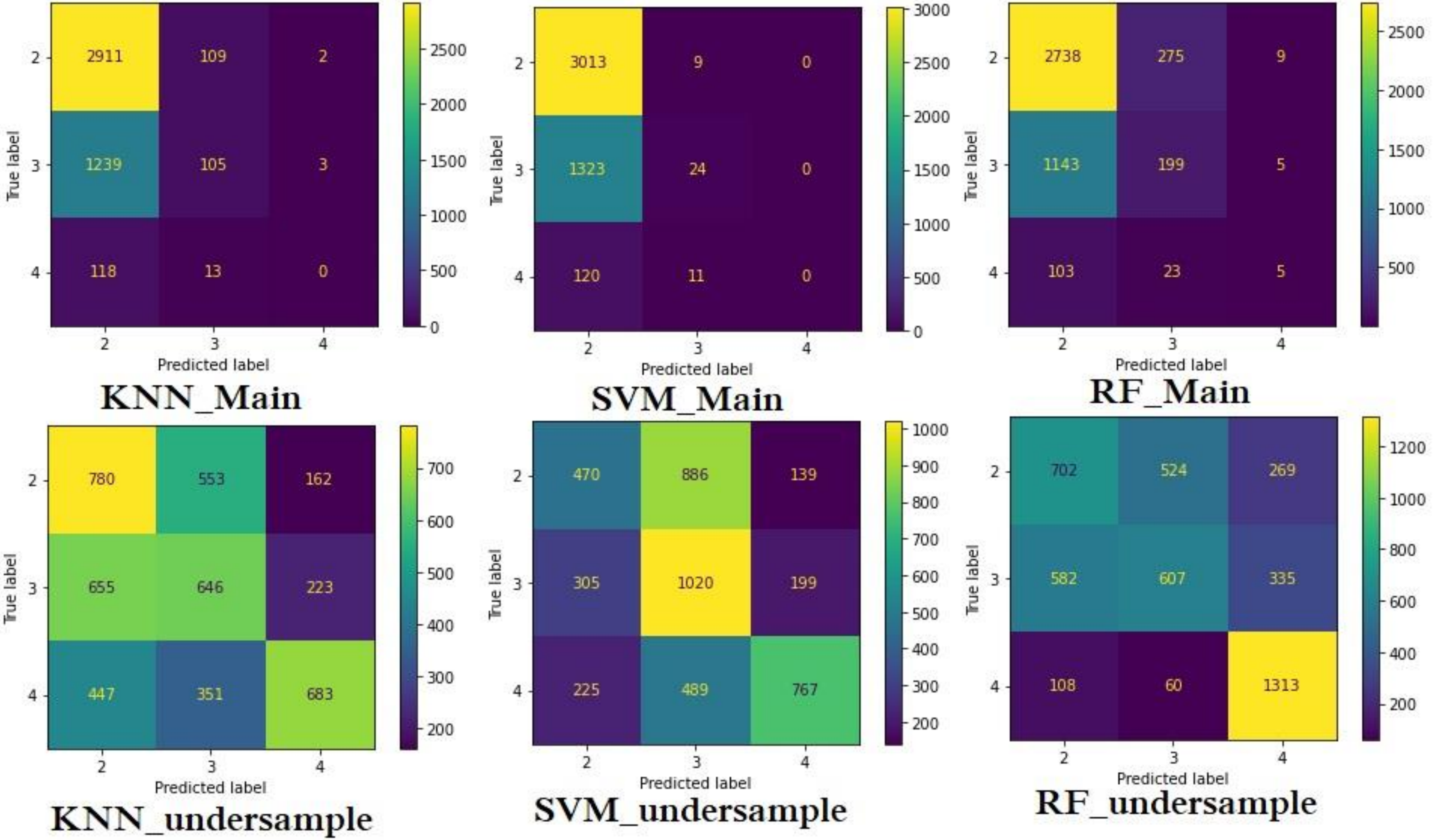


**RF_Mainsample Model**

```
              precision    recall  f1-score   support

           2       0.69      0.91      0.78      3022
           3       0.40      0.15      0.22      1347
           4       0.26      0.04      0.07       131

    accuracy                           0.65      4500
   macro avg       0.45      0.36      0.35      4500
weighted avg       0.59      0.65      0.59      4500
```

**RF_undersample Model**

```
              precision    recall  f1-score   support

           2       0.50      0.47      0.49      1495
           3       0.51      0.40      0.45      1524
           4       0.68      0.89      0.77      1481

    accuracy                           0.58      4500
   macro avg       0.57      0.58      0.57      4500
weighted avg       0.57      0.58      0.57      4500
```



**KNN_Main**

| True label / Predicted label | 2 | 3 | 4 |
|---|---|---|---|
| 2 | 2911 | 109 | 2 |
| 3 | 1239 | 105 | 3 |
| 4 | 118 | 13 | 0 |

**SVM_Main**

| True label / Predicted label | 2 | 3 | 4 |
|---|---|---|---|
| 2 | 3013 | 9 | 0 |
| 3 | 1323 | 24 | 0 |
| 4 | 120 | 11 | 0 |

**RF_Main**

| True label / Predicted label | 2 | 3 | 4 |
|---|---|---|---|
| 2 | 2738 | 275 | 9 |
| 3 | 1143 | 199 | 5 |
| 4 | 103 | 23 | 5 |

**KNN_undersample**

| True label / Predicted label | 2 | 3 | 4 |
|---|---|---|---|
| 2 | 780 | 553 | 162 |
| 3 | 655 | 646 | 223 |
| 4 | 447 | 351 | 683 |

**SVM_undersample**

| True label / Predicted label | 2 | 3 | 4 |
|---|---|---|---|
| 2 | 470 | 886 | 139 |
| 3 | 305 | 1020 | 199 |
| 4 | 225 | 489 | 767 |

**RF_undersample**

| True label / Predicted label | 2 | 3 | 4 |
|---|---|---|---|
| 2 | 702 | 524 | 269 |
| 3 | 582 | 607 | 335 |
| 4 | 108 | 60 | 1313 |

**Confusion Matrix: Model comparision**

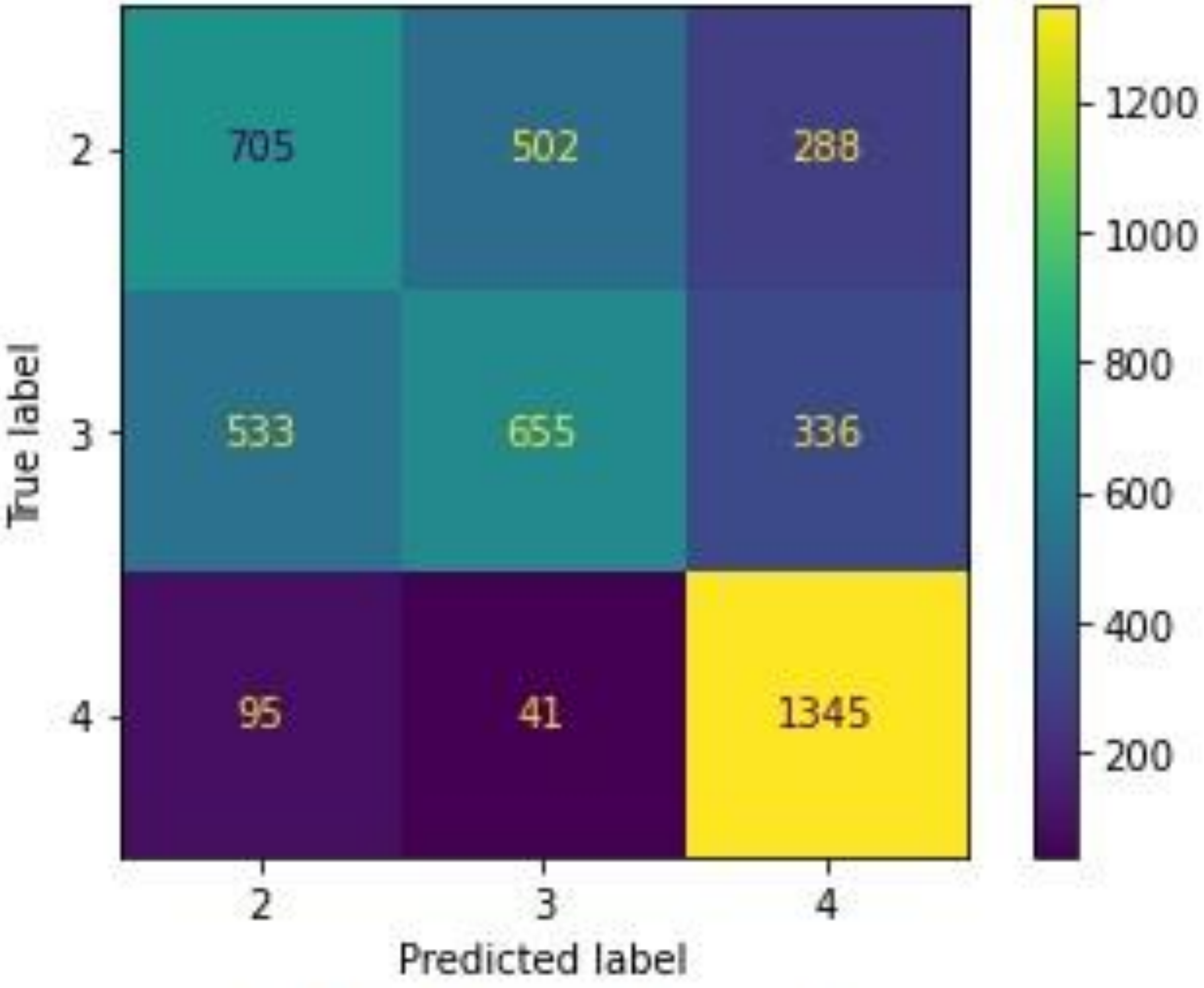
# RESULTS : CLASSIFICATION: BEST MODEL

- RF_Grid_undersample Model Performances

**RF_Grid_undersample Model Performances**

```
              precision    recall  f1-score   support

           2       0.53      0.47      0.50      1495
           3       0.55      0.43      0.48      1524
           4       0.68      0.91      0.78      1481

    accuracy                           0.60      4500
   macro avg       0.59      0.60      0.59      4500
weighted avg       0.59      0.60      0.59      4500
```

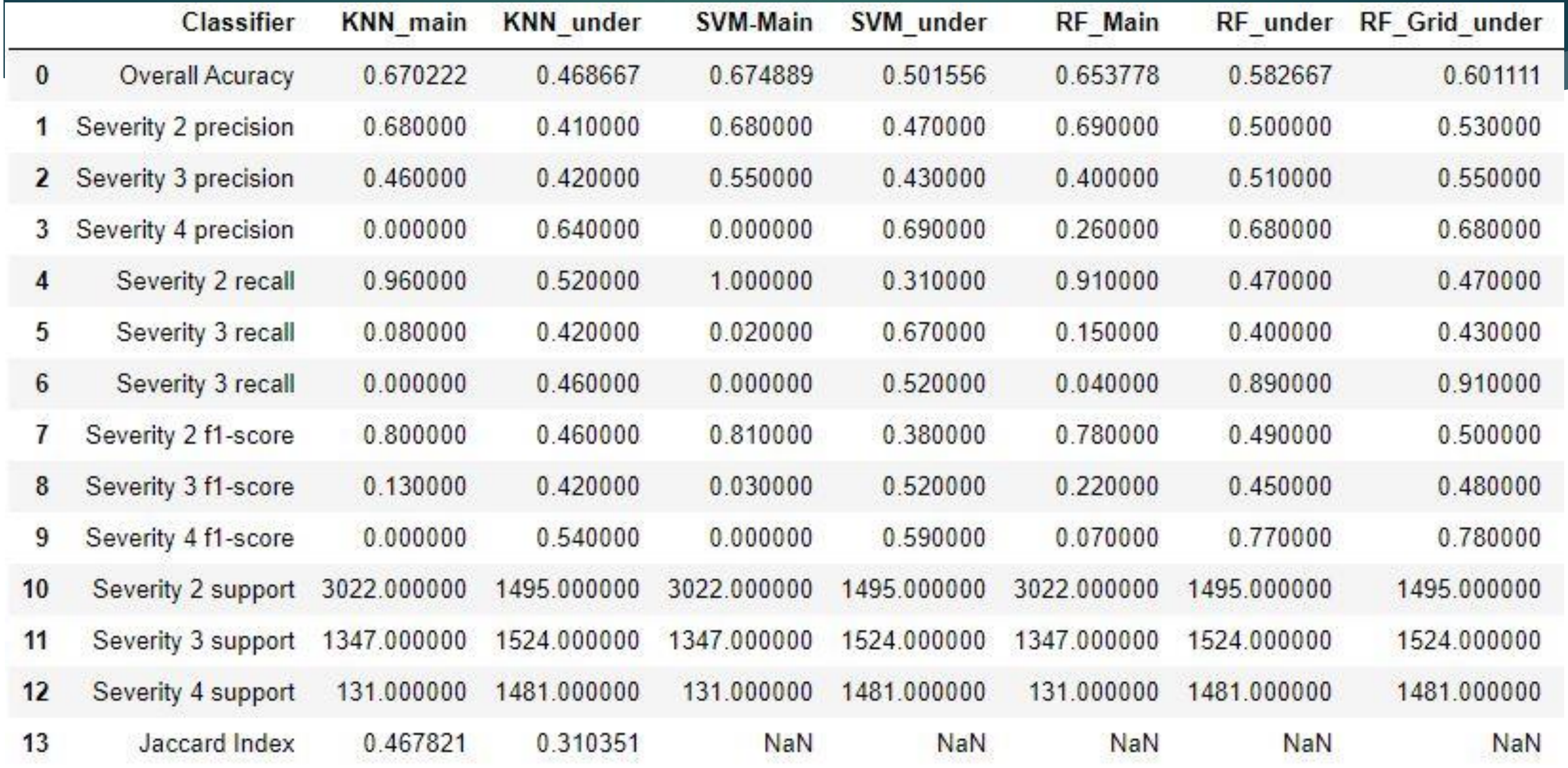
# RESULTS : CLASSIFICATION: BEST MODEL



| True label \ Predicted label | 2 | 3 | 4 |
|---|---|---|---|
| 2 | 705 | 502 | 288 |
| 3 | 533 | 655 | 336 |
| 4 | 95 | 41 | 1345 |

RF_Grid_under

# PERFORMANCES : DISCUSSION



| | Classifier | KNN_main | KNN_under | SVM-Main | SVM_under | RF_Main | RF_under | RF_Grid_under |
|---|---|---|---|---|---|---|---|---|
| 0 | Overall Acuracy | 0.670222 | 0.468667 | 0.674889 | 0.501556 | 0.653778 | 0.582667 | 0.601111 |
| 1 | Severity 2 precision | 0.680000 | 0.410000 | 0.680000 | 0.470000 | 0.690000 | 0.500000 | 0.530000 |
| 2 | Severity 3 precision | 0.460000 | 0.420000 | 0.550000 | 0.430000 | 0.400000 | 0.510000 | 0.550000 |
| 3 | Severity 4 precision | 0.000000 | 0.640000 | 0.000000 | 0.690000 | 0.260000 | 0.680000 | 0.680000 |
| 4 | Severity 2 recall | 0.960000 | 0.520000 | 1.000000 | 0.310000 | 0.910000 | 0.470000 | 0.470000 |
| 5 | Severity 3 recall | 0.080000 | 0.420000 | 0.020000 | 0.670000 | 0.150000 | 0.400000 | 0.430000 |
| 6 | Severity 3 recall | 0.000000 | 0.460000 | 0.000000 | 0.520000 | 0.040000 | 0.890000 | 0.910000 |
| 7 | Severity 2 f1-score | 0.800000 | 0.460000 | 0.810000 | 0.380000 | 0.780000 | 0.490000 | 0.500000 |
| 8 | Severity 3 f1-score | 0.130000 | 0.420000 | 0.030000 | 0.520000 | 0.220000 | 0.450000 | 0.480000 |
| 9 | Severity 4 f1-score | 0.000000 | 0.540000 | 0.000000 | 0.590000 | 0.070000 | 0.770000 | 0.780000 |
| 10 | Severity 2 support | 3022.000000 | 1495.000000 | 3022.000000 | 1495.000000 | 3022.000000 | 1495.000000 | 1495.000000 |
| 11 | Severity 3 support | 1347.000000 | 1524.000000 | 1347.000000 | 1524.000000 | 1347.000000 | 1524.000000 | 1524.000000 |
| 12 | Severity 4 support | 131.000000 | 1481.000000 | 131.000000 | 1481.000000 | 131.000000 | 1481.000000 | 1481.000000 |
| 13 | Jaccard Index | 0.467821 | 0.310351 | NaN | NaN | NaN | NaN | NaN |

# CONCLUSIONS

- RF with a balanced sample clearly performed well and predict much better.
- The best model therefore is the RF_Grid_under model.
- Both KNN and SVM models are highly affected by the imbalance of the samples as all the performance indicators with respect to the minority class is nothing but zero for both these models.

# THANK YOU!