Introduction to Business Problem:

To educate people about the impact and risk of bad habit and lack of disciple while driving, a whole analytics section is to be conducted in order to classify new estimated cases of human/environment anomaly that will potentially lead to different classes of injuries and fatalities and Public & personal properties damages.

Data Introduction:

The data provide a history of accident recorded by the government, which rates the accident severity by looking at different factors as:

Independent Variables:

- Counts of Involved: Persons, Pedestrian, bicycles, vehicles count.
- Conditions of: Environment, Road, Junction type, Light .
- Potential causes flags: Inattention, drugs or alcohols, speeding, pedestrian right Flags.
- Injuries: number of total injuries, serious injuries and fatalities.
- Timeframe and place: Date and Place pf the accident.
- Others: Collision type, parked car hit.

Dependent Variables:

Based on all factors described above, a raking (severity code) is given to the accident.

- Severity Code and its description.

Below attached a link of the Metadata that can explain some details about data features

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

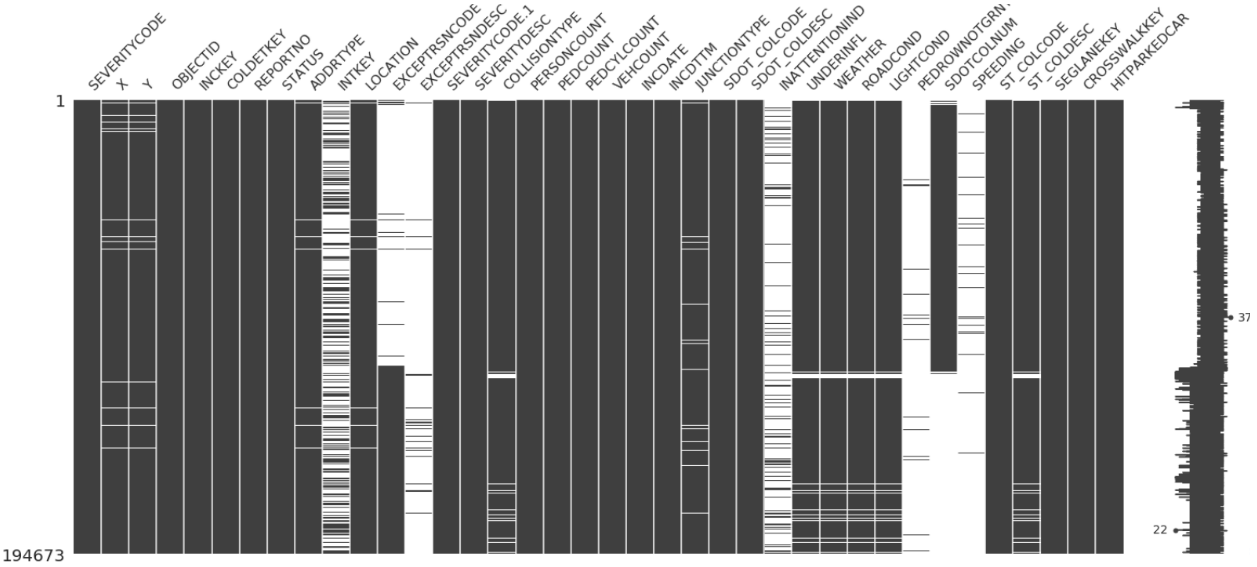
I. Methodology:

1. Data wrangling & feature selection
2. EDA
3. Modeling
4. Results

II. Data wrangling & feature selection

The initial data represents too much missing, useless/duplicated features.

The bars below show the overall data quality per features:

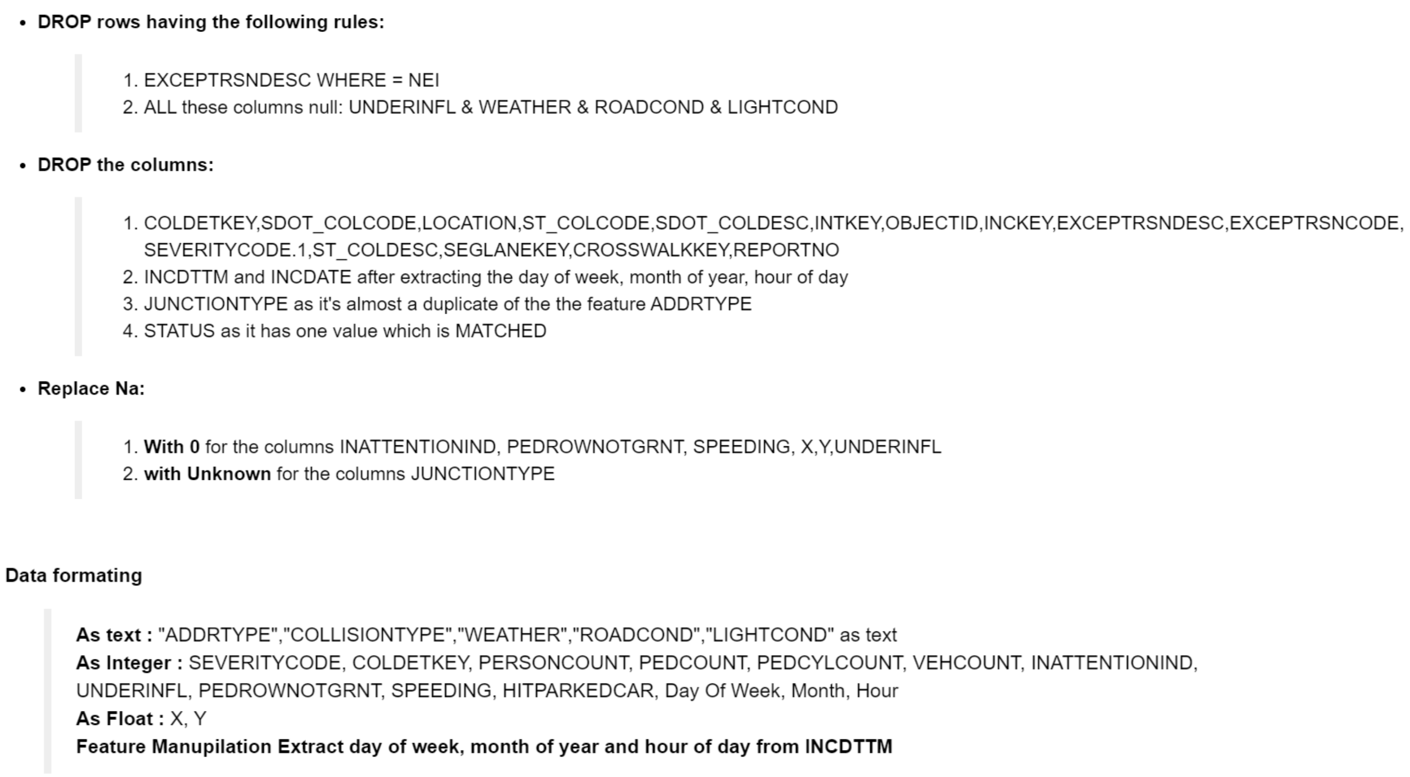


- We see clearly some columns have so many missing value. The top 3 are:
  - Intentioned
  - Speeding
  - Pedrownotgrn

  The reason why the potential important columns are null is the fact the null values represent a "No" Flags. **In the data wrangling strategy, we decide to fill null with 0 value.**

Other columns that represent a significant missing record needs to be dropped as they aren't a considered feature for our model.

We defined the below steps for Data wrangling steps :



- **DROP rows having the following rules:**

> 1. EXCEPTRSNDESC WHERE = NEI
> 2. ALL these columns null: UNDERINFL & WEATHER & ROADCOND & LIGHTCOND

- **DROP the columns:**

> 1. COLDETKEY,SDOT_COLCODE,LOCATION,ST_COLCODE,SDOT_COLDESC,INTKEY,OBJECTID,INCKEY,EXCEPTRSNDESC,EXCEPTRSNCODE, SEVERITYCODE.1,ST_COLDESC,SEGLANEKEY,CROSSWALKKEY,REPORTNO
> 2. INCDTTM and INCDATE after extracting the day of week, month of year, hour of day
> 3. JUNCTIONTYPE as it's almost a duplicate of the the feature ADDRTYPE
> 4. STATUS as it has one value which is MATCHED

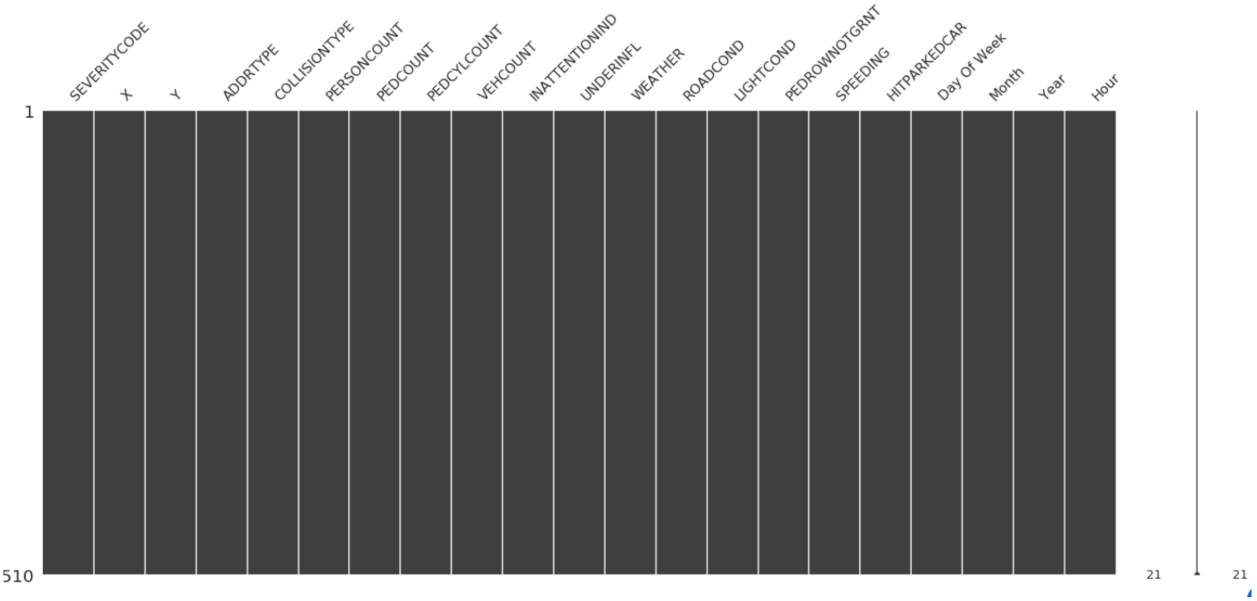
- **Replace Na:**

> 1. **With 0** for the columns INATTENTIONIND, PEDROWNOTGRNT, SPEEDING, X,Y,UNDERINFL
> 2. **with Unknown** for the columns JUNCTIONTYPE

Data formating

> **As text :** "ADDRTYPE","COLLISIONTYPE","WEATHER","ROADCOND","LIGHTCOND" as text
> **As Integer :** SEVERITYCODE, COLDETKEY, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, HITPARKEDCAR, Day Of Week, Month, Hour
> **As Float :** X, Y
> **Feature Manupilation Extract day of week, month of year and hour of day from INCDTTM**

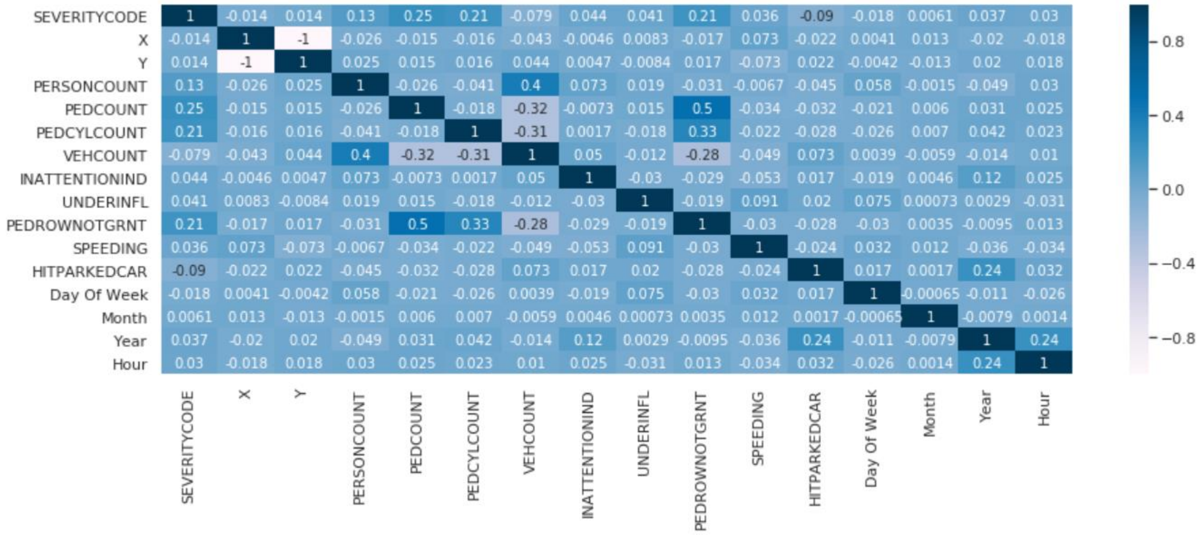
All processing done, here the kept features that we will rely on as input for model:



We clearly see zero missing data.

III. EDA:

Pearson Correlation Matrix:



	SEVERITYCODE	X	Y	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INATTENTIONIND	UNDERINFL	PEDROWNOTGRNT	SPEEDING	HITPARKEDCAR	Day Of Week	Month	Year	Hour
SEVERITYCODE	1	-0.014	0.014	0.13	0.25	0.21	-0.079	0.044	0.041	0.21	0.036	-0.09	-0.018	0.0061	0.037	0.03
X	-0.014	1	-1	-0.026	-0.015	-0.016	-0.043	-0.0046	0.0083	-0.017	0.073	-0.022	0.0041	0.013	-0.02	-0.018
Y	0.014	-1	1	0.025	0.015	0.016	0.044	0.0047	-0.0084	0.017	-0.073	0.022	-0.0042	-0.013	0.02	0.018
PERSONCOUNT	0.13	-0.026	0.025	1	-0.026	-0.041	0.4	0.073	0.019	-0.031	-0.0067	-0.045	0.058	-0.0015	-0.049	0.03
PEDCOUNT	0.25	-0.015	0.015	-0.026	1	-0.018	-0.32	-0.0073	0.015	0.5	-0.034	-0.032	-0.021	0.006	0.031	0.025
PEDCYLCOUNT	0.21	-0.016	0.016	-0.041	-0.018	1	-0.31	0.0017	-0.018	0.33	-0.022	-0.028	-0.026	0.007	0.042	0.023
VEHCOUNT	-0.079	-0.043	0.044	0.4	-0.32	-0.31	1	0.05	-0.012	-0.28	-0.049	0.073	0.0039	-0.0059	-0.014	0.01
INATTENTIONIND	0.044	-0.0046	0.0047	0.073	-0.0073	0.0017	0.05	1	-0.03	-0.029	-0.053	0.017	-0.019	0.0046	0.12	0.025
UNDERINFL	0.041	0.0083	-0.0084	0.019	0.015	-0.018	-0.012	-0.03	1	-0.019	0.091	0.02	0.075	0.00073	0.0029	-0.031
PEDROWNOTGRNT	0.21	-0.017	0.017	-0.031	0.5	0.33	-0.28	-0.029	-0.019	1	-0.03	-0.028	-0.03	0.0035	-0.0095	0.013
SPEEDING	0.036	0.073	-0.073	-0.0067	-0.034	-0.022	-0.049	-0.053	0.091	-0.03	1	-0.024	0.032	0.012	-0.036	-0.034
HITPARKEDCAR	-0.09	-0.022	0.022	-0.045	-0.032	-0.028	0.073	0.017	0.02	-0.028	-0.024	1	0.017	0.0017	0.24	0.032
Day Of Week	-0.018	0.0041	-0.0042	0.058	-0.021	-0.026	0.0039	-0.019	0.075	-0.03	0.032	0.017	1	-0.00065	-0.011	-0.026
Month	0.0061	0.013	-0.013	-0.0015	0.006	0.007	-0.0059	0.0046	0.00073	0.0035	0.012	0.0017	-0.00065	1	-0.0079	0.0014
Year	0.037	-0.02	0.02	-0.049	0.031	0.042	-0.014	0.12	0.0029	-0.0095	-0.036	0.24	-0.011	-0.0079	1	0.24
Hour	0.03	-0.018	0.018	0.03	0.025	0.023	0.01	0.025	-0.031	0.013	-0.034	0.032	-0.026	0.0014	0.24	1

The matrix above shows that there is almost no perfect correlation between dependent - dependent variables or independent-dependent variable.

The most correlation we have are:

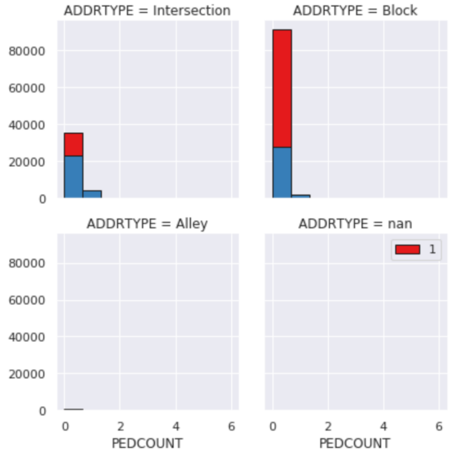
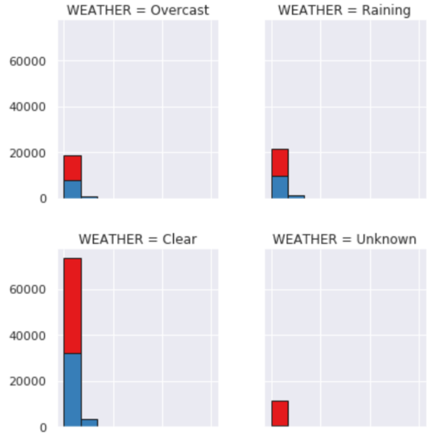
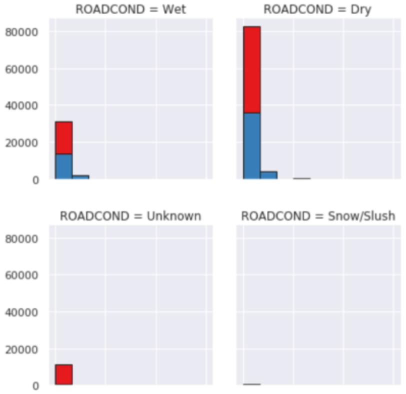
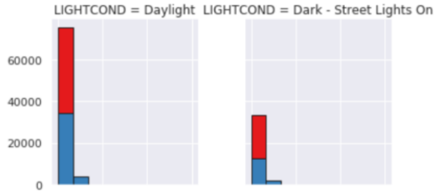
Independent variable:

- PEDROWNOTGRANT – PEDCYLCOUNT (0.5)
- VEHCOUNT – PERSONCOUNT (0.4)
- SPEEDING- PEDCYLCOUNT(0.33)

Dependent variable:

- SEVERIYCODE- PEDCOUNT(0.25)
- SEVERIYCODE- PEDCYLCOUNT(0.21)
- SEVERIYCODE- PEDROWNOTGRANT (0.21)

Analyses of categorical data on the dependent variables:

Categorical feature	PedCount Histogram
Collison address We see clearly that in terms of relative dangerous address, the intersections cause injury while the Block collisions have impact on properties.	
Weather It seems the Weather values has no impact of the split between injury or properties damage. Also the majority of collision happens during Clear weather.	
Road Road has the same profile like weather. Most Collison happens on the Dry Road. But the road condition, doesn't have any impact on the severity of the collision	
Light Light has the same profile like weather & road. Most Collison happens during daylight. But the light doesn't have any impact on the severity of the collision.	

IV. Model:

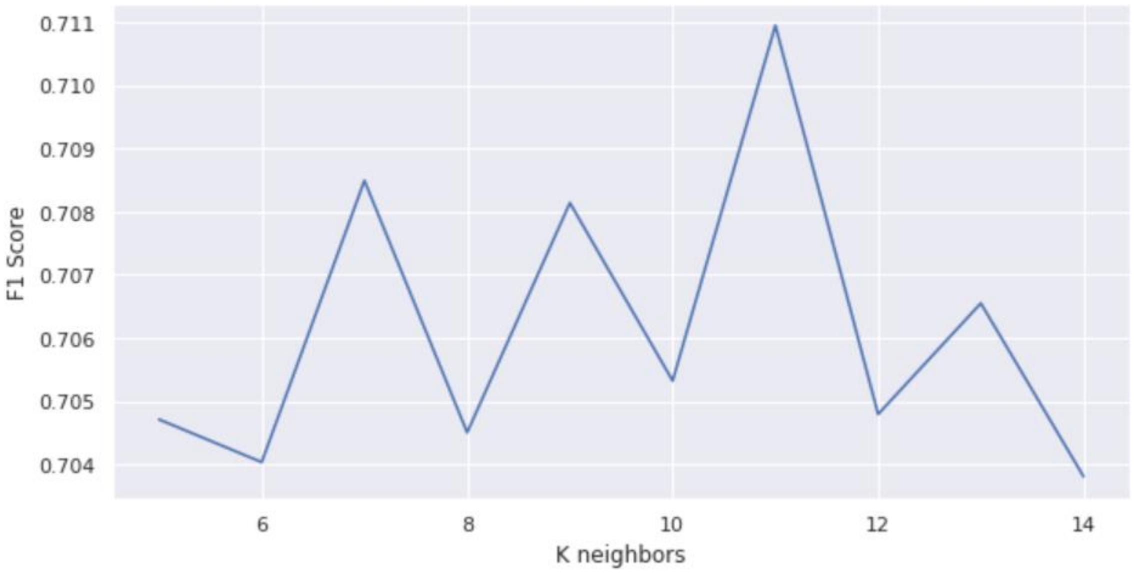
We decided in this section to use the KNN classification model.

The choice of the number of neighbors will be determined via iterative process.

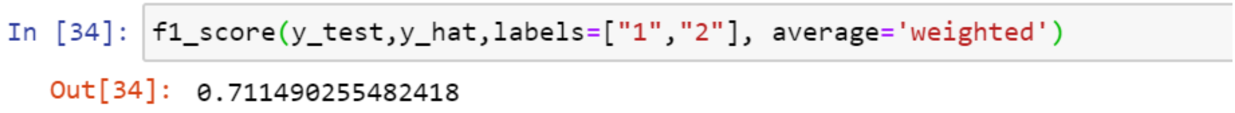
In order to prepare data for model, this is the steps designed in order to fit the model:

1. Converting categorical data to binary columns: Via one hot encoding using Pandas Get_dummies function
2. Normalizing data using the Zscaler using StandardScaler from sklearn.preprocessing Library
3. Split train and test data randomly: 80-20 using train_test_split function from sklearn
4. Repeat
    a. Choose K
    b. Fit model
    c. Predict model
    d. Measure accuracy using f1_score
5. Choose the K based on best f1_score.

All steps above done, here is the outcome of accuracy for k neighbors between [1,15] :

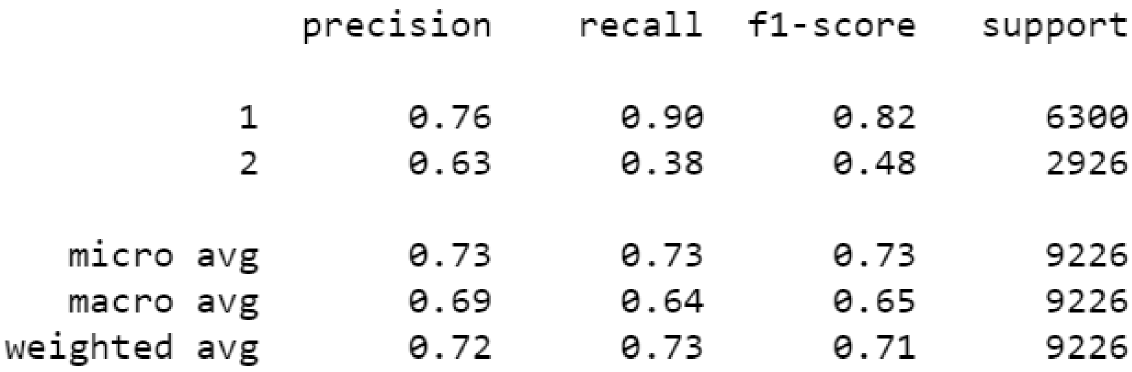


The F1 score for K= 11 is :



```
In [34]: f1_score(y_test,y_hat,labels=["1","2"], average='weighted')
Out[34]: 0.711490255482418
```

The evaluation metrics shows the following results:



```
              precision    recall  f1-score   support

           1       0.76      0.90      0.82      6300
           2       0.63      0.38      0.48      2926

   micro avg       0.73      0.73      0.73      9226
   macro avg       0.69      0.64      0.65      9226
weighted avg       0.72      0.73      0.71      9226
```

- **The model has a very interesting precision for both classes.**
- **The recall for class1 is nearly perfect while it’s lower for the 2 class.**
- **The f1 score is pretty much good in average**