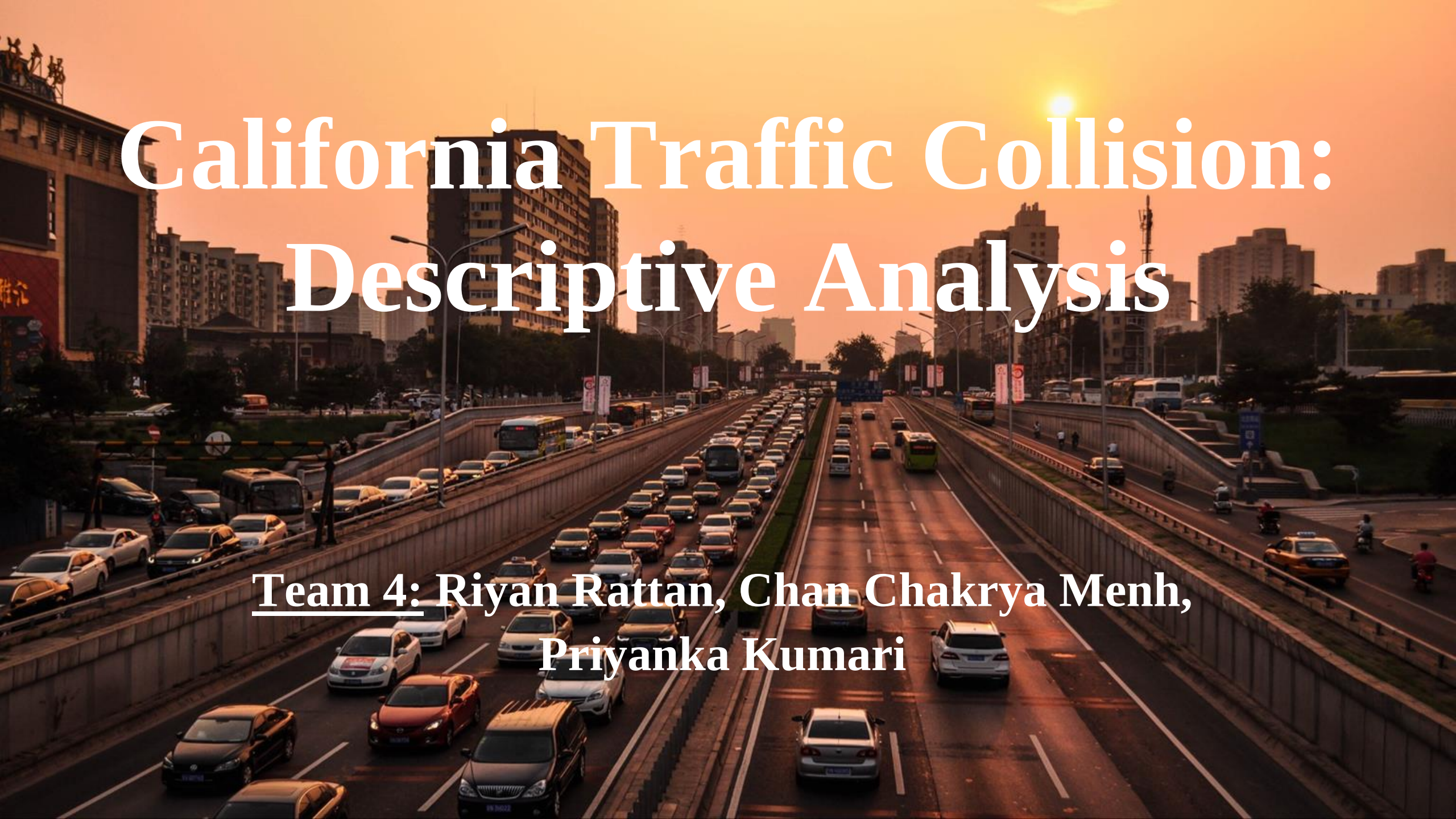


California Traffic Collision: Descriptive Analysis

Team 4: Riyan Rattan, Chan Chakrya Menh, Priyanka Kumari

Content

I. Introduction

Background:

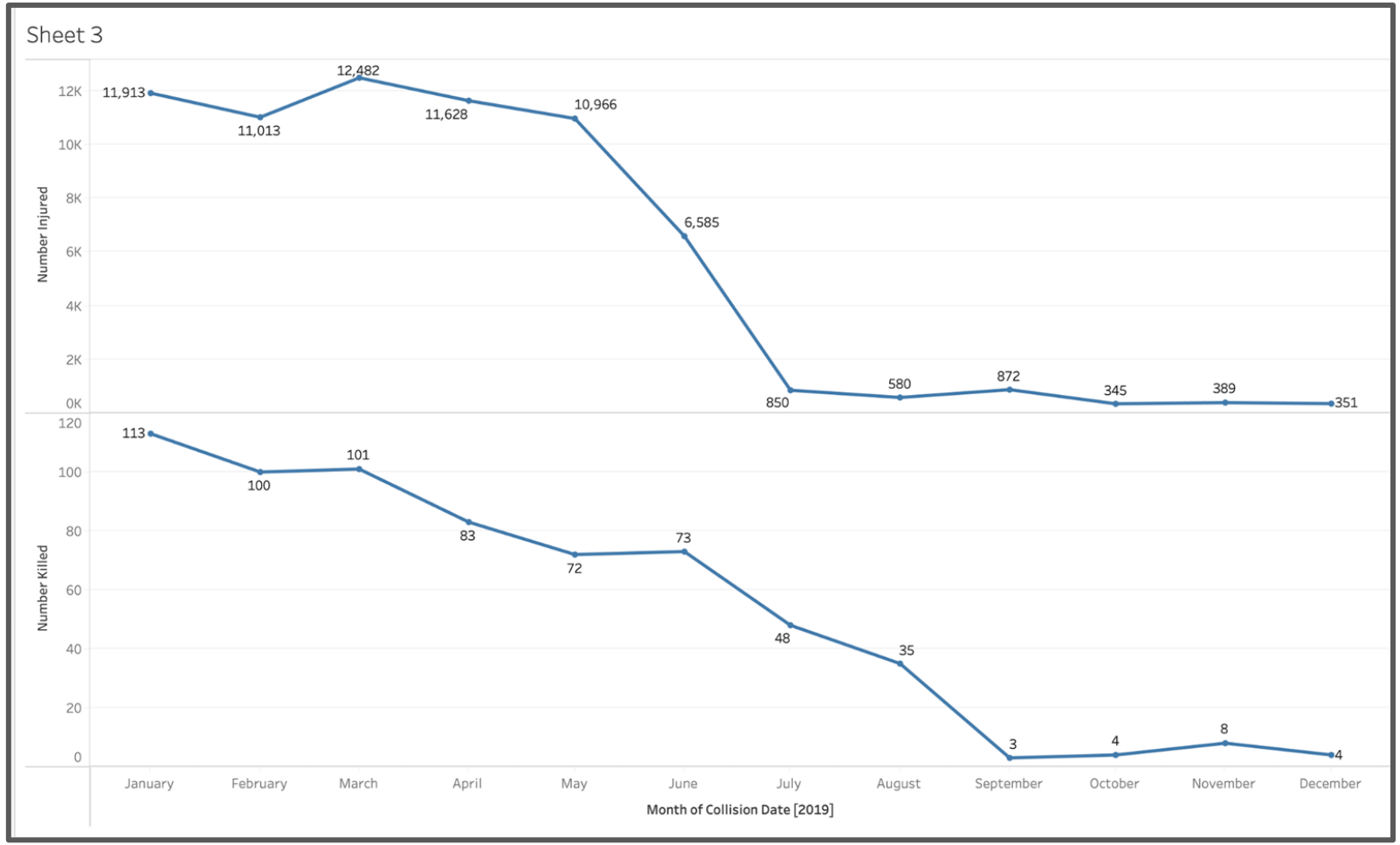
What is Traffic Collision?

A traffic collision or crash occurs when a vehicle collides with another vehicle, pedestrian, road barrier, or a stationary obstacle such as a tree or a utility pole. It may result in injury, death, vehicle damage, possession damage which causes death and disability, and financial burden.

The National Highway Traffic Safety Administration (NHTSA) disclosed its early estimation of traffic fatalities for 2021. NHTSA projects an estimated 42,915 individuals died in motor vehicle traffic crashes last year, a 10.5% expansion from the 38,824 fatalities in 2020. The projection is the highest fatalities since 2005 and the most significant annual percentage increase in the Fatality Analysis Reporting System history.

I. Introduction (cont.)
Number Killed Vs Number Injured by Month

In comparison, number of injured on monthly basis is more than number of killed from the dataset 10K from California 2019



1.1 Motive/ Goals

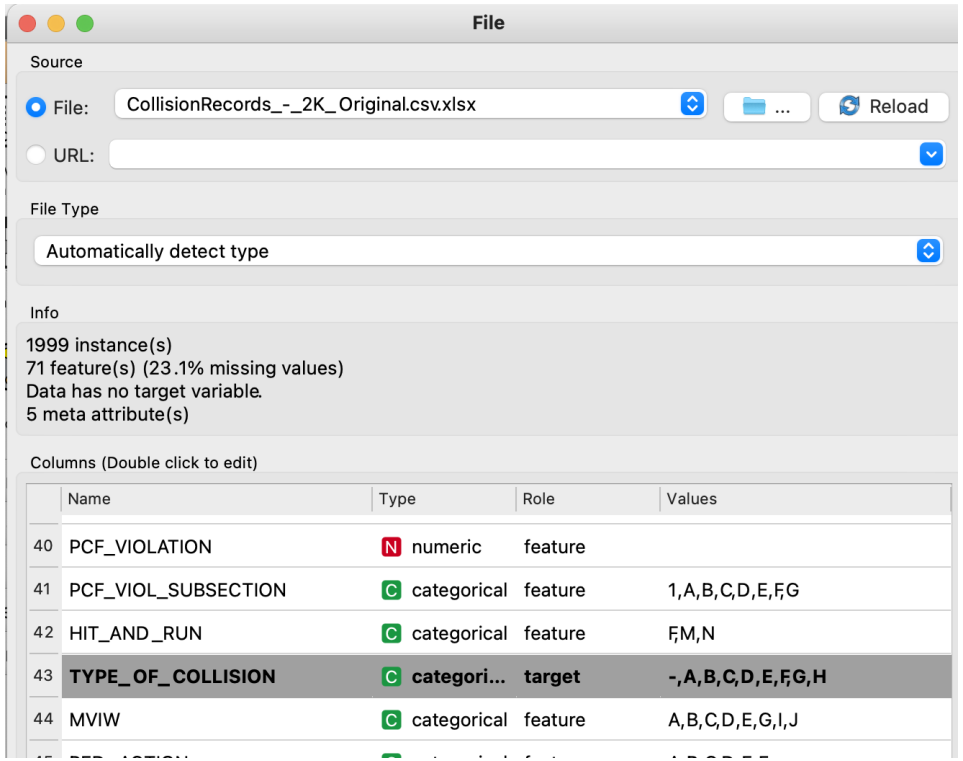
- The main goal of the project is to use machine learning to detect the collision type from California traffic collision dataset in 2019 and define the best mode for classification target.

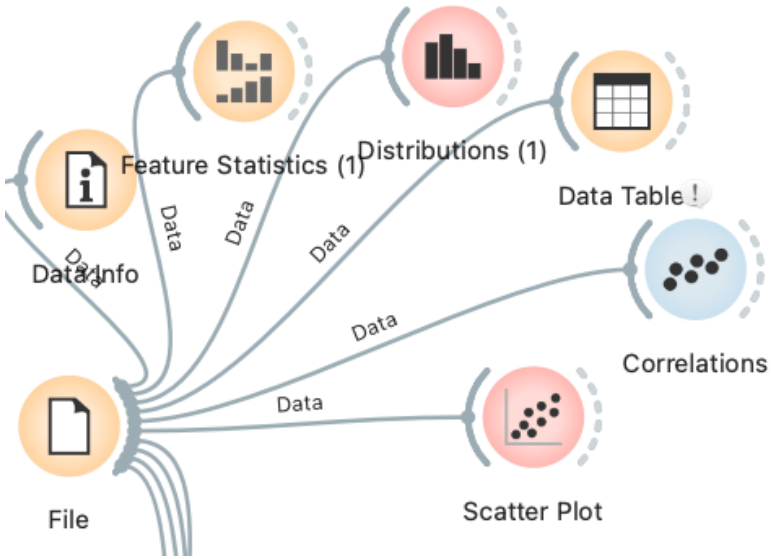
II. Research Questions

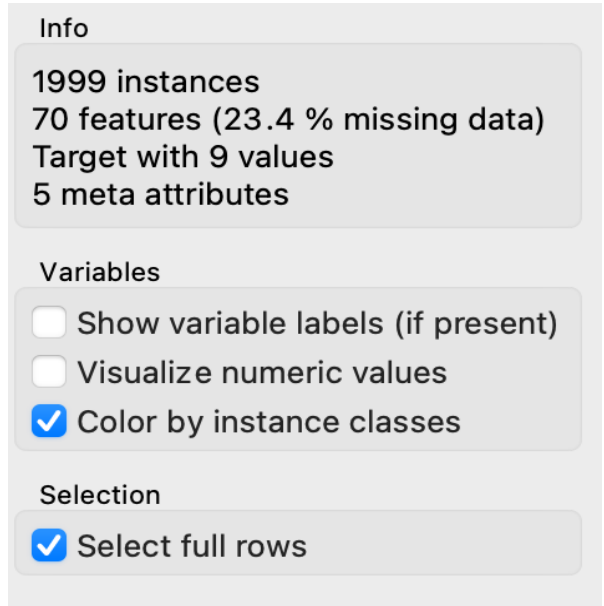
- What are collision type?
- Can we use machine learning to detect the type of collision?
- Which modeling would be best for classification target?
- Is there any relationship between the target and parameter?

III. Data Preparation
3.1. Review data set

- Load file on orange and inspect the data types, and attribute names.
- Select variable target ‘Type of collision’
- Data instances about 2k with 70 features, and about 23.4% missing values



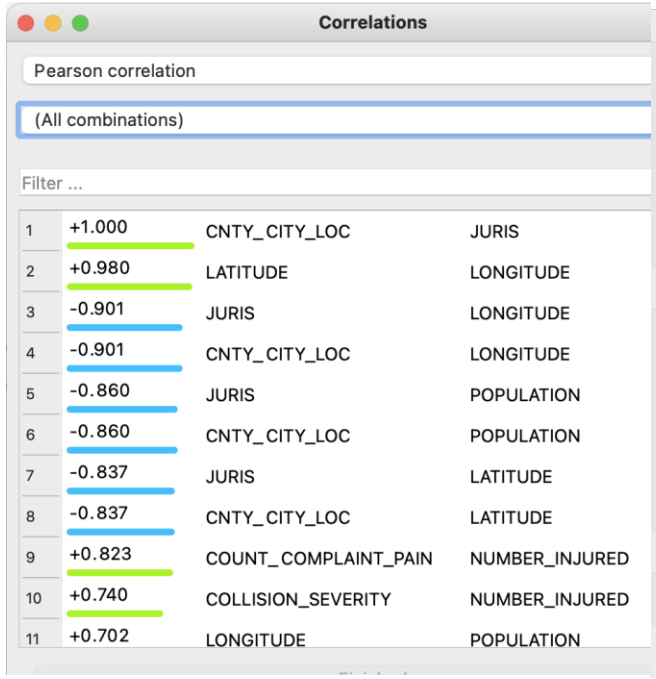


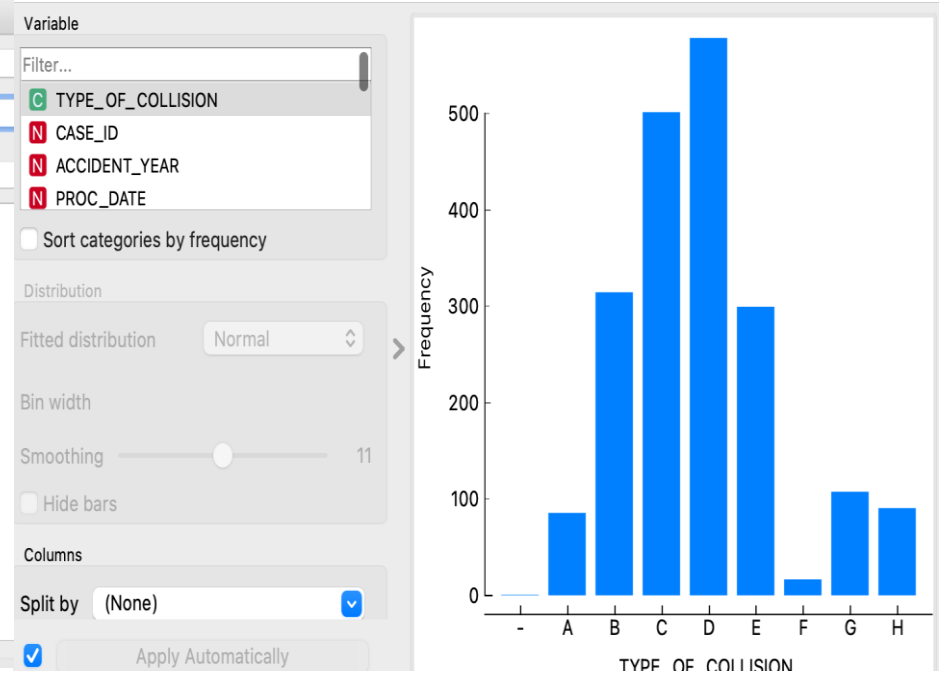


III. Data Preparation
3.2. Inspect the dataset

Inspect data table, feature statistic, correlation, and target distribution.
Notice somes categories and numeric types have missing values, and imbalance classification

	'PE_OF_COLLISIC	OFFICER_ID	PRIMARY_RD	SECONDARY_RD	F_VIOL_CATEGO	_VEHTYPE_AT_FA	CASE_ID	ACCIDENT_YEAR	PROC_DATE	JURIS	COLLISION_DATE	COI
1	C	975	SLOVER AV	ELM ST	3	22	8008498	2019	20190604	3604	20190514	
2	D	5163	CROWN VAL...	GOLDEN LA...	3	1	8008502	2019	20190710	3000	20190430	
3	D	27792	MAGNOLIA BL	BLUEBELL AV	9	3	8008506	2019	20190816	1942	20190529	
4	C	469	WOODSIDE ...	MIDDLEFIEL...	3	1	8008510	2019	20190801	4113	20190616	
5	D	630201	60TH ST WE...	AVENUE G	9	27	8008514	2019	20190820	1900	20190726	





III. Data Preparation
3.3. Classification balancing

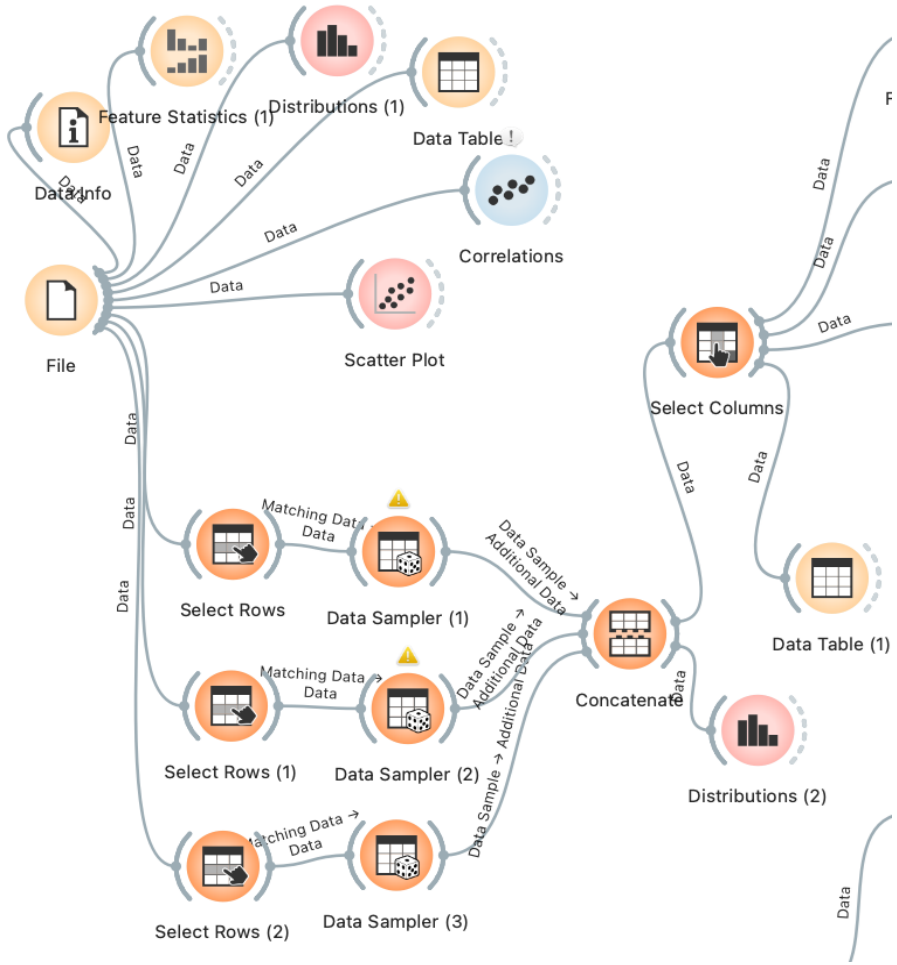
Since the target distribution is imbalancing, so we run classification balance by using select rows and data sampler to adjust data sample size of each classification.

1.A,G,H,F

2.B,E

3. C,D

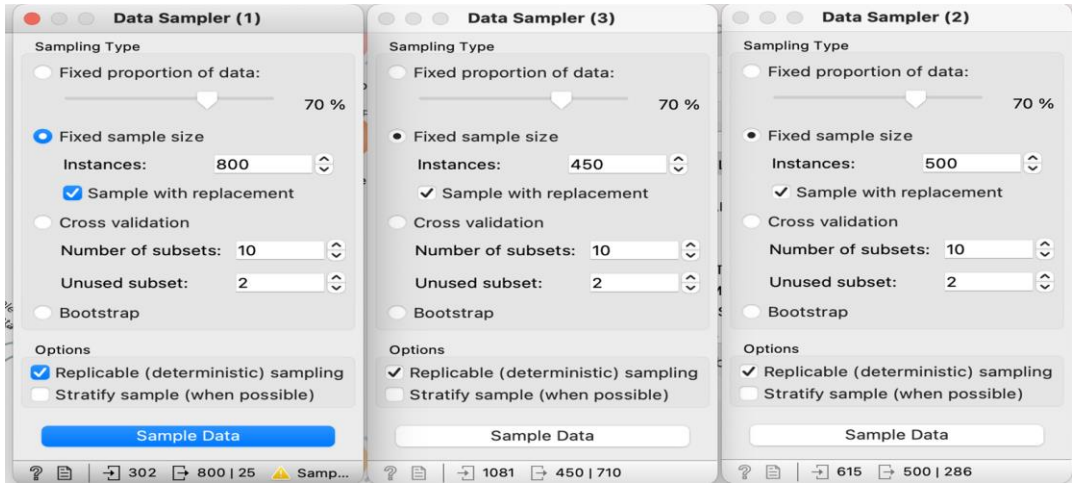
A - Head-On
B - Sideswipe
C - Rear End
D - Broadside
E - Hit Object
F - Overturned
G - Vehicle/Pedestrian
H - Other

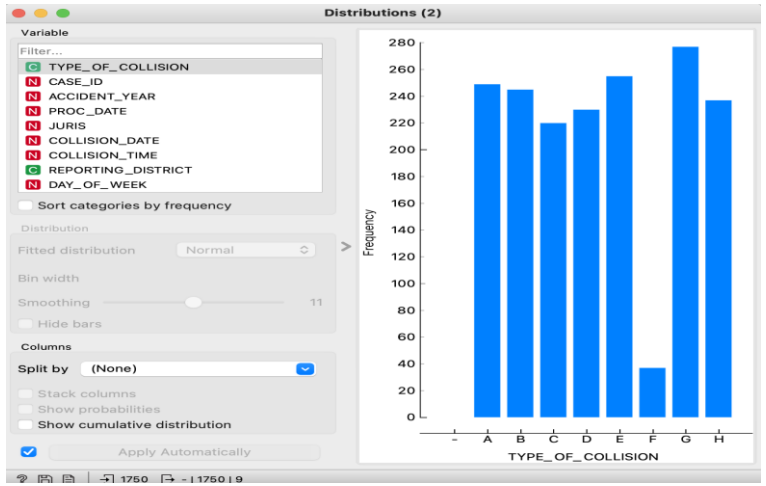


3.3. Classification balancing

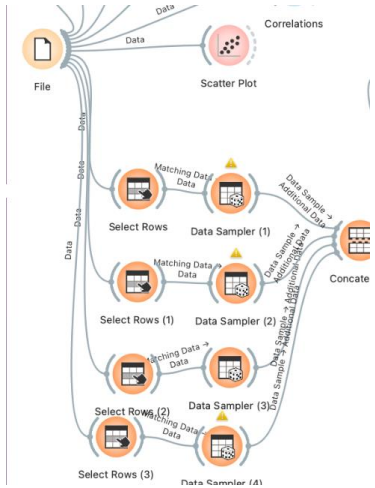
3.3.1. First balancing

1st balancing improved, but ‘F’ still imbalance.



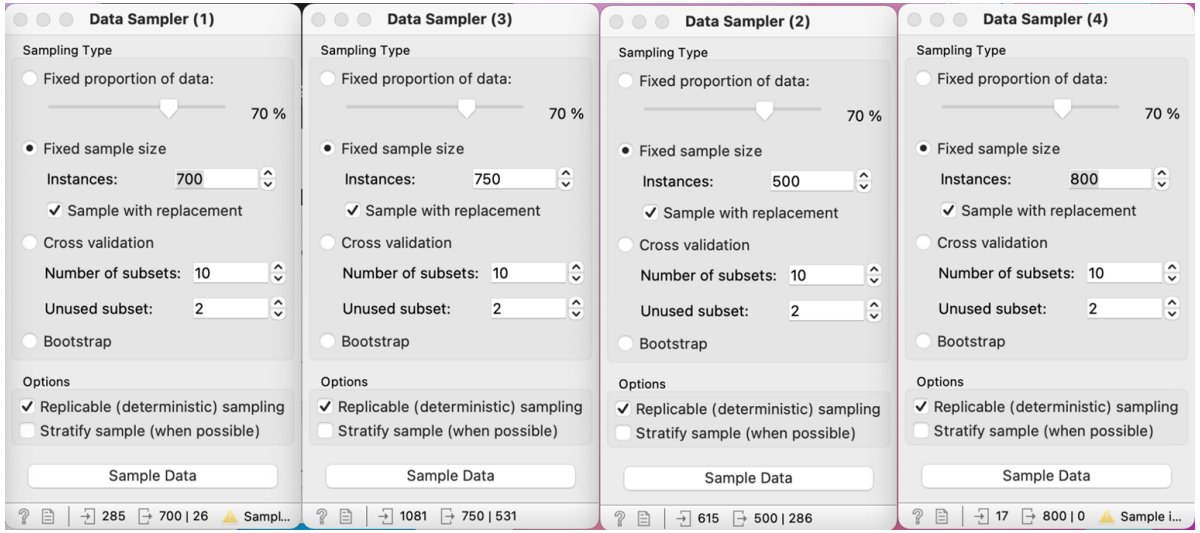


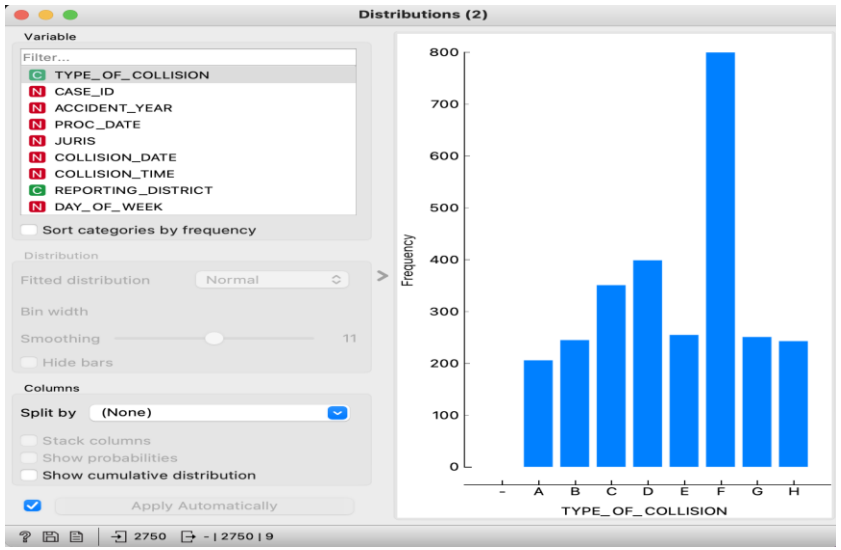
After run multiples data sample, we decide to split ‘F’ into its own category for data sampler.



3.3. Classification balancing (cont.)
3.3.2. Second balancing

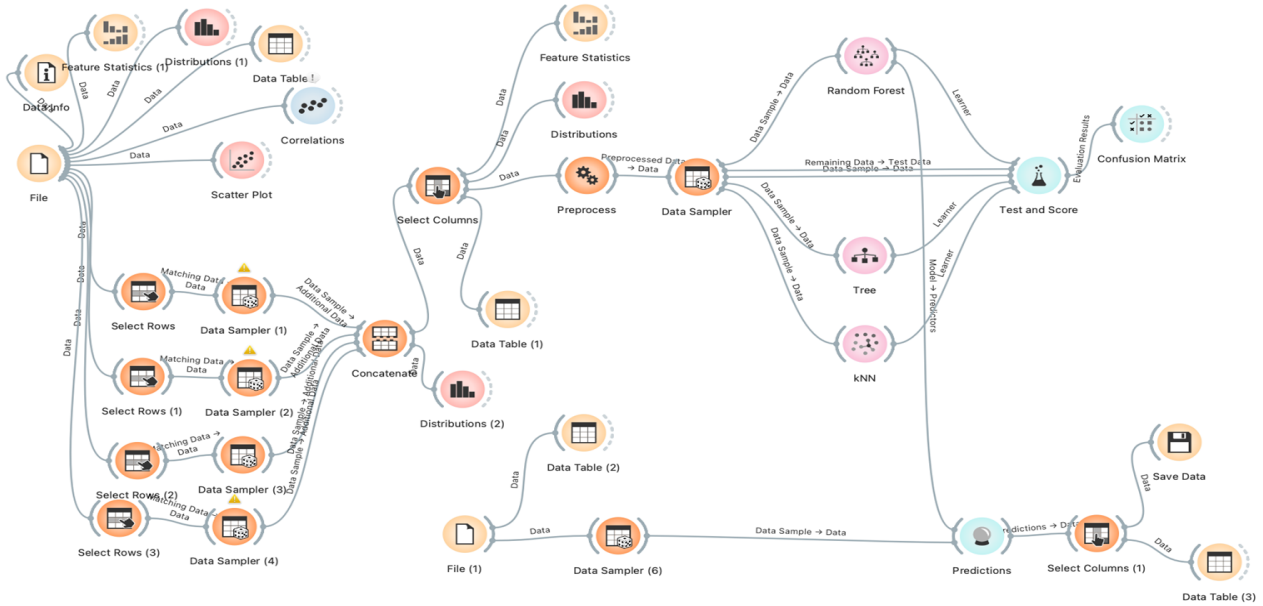
After adjustment, ‘F’ become the highest, classification still imabance, however, we construct a pipeline in orange to test this the classification.

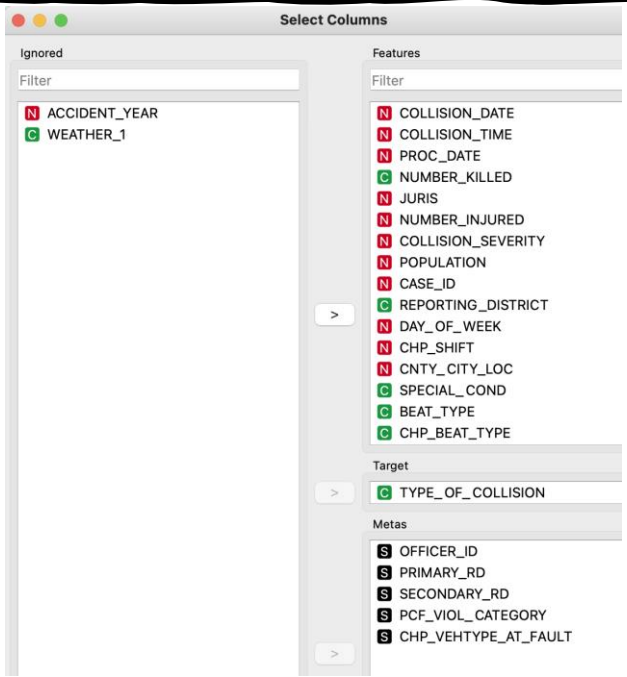


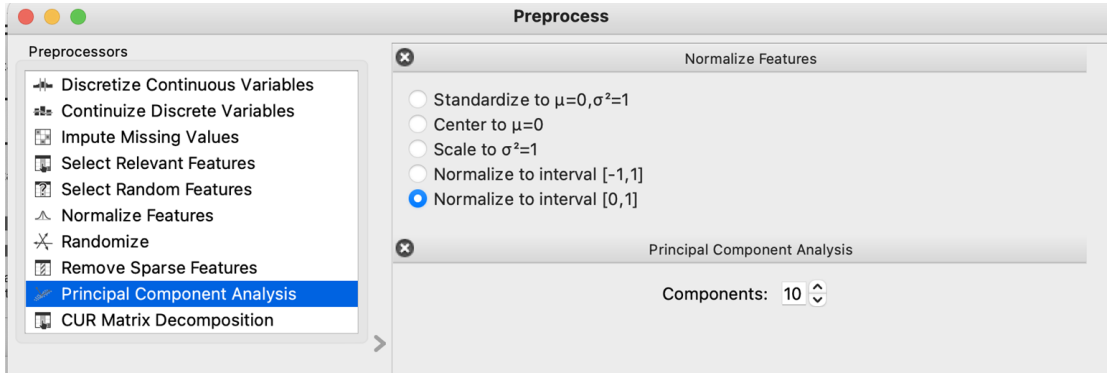


IV. Data Processing and Modeling

from second classification balance, we construct a pipeline with three models: KNN, Tree and Random forest





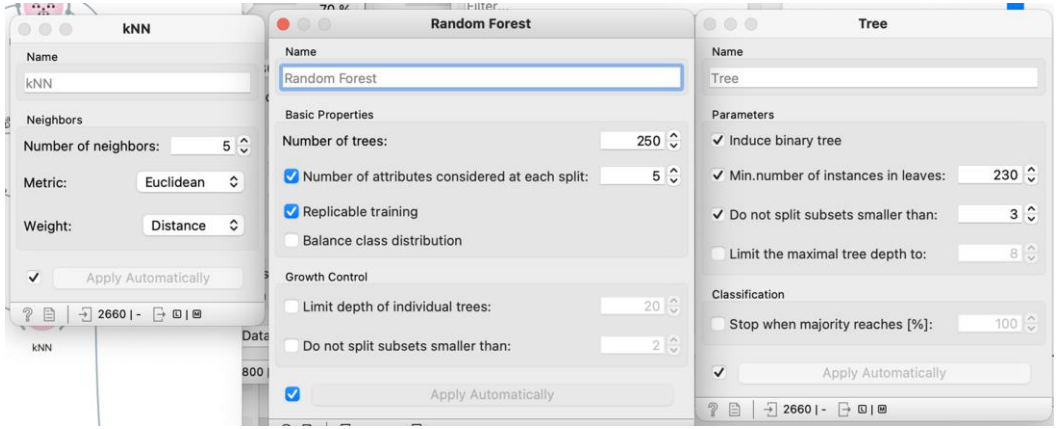


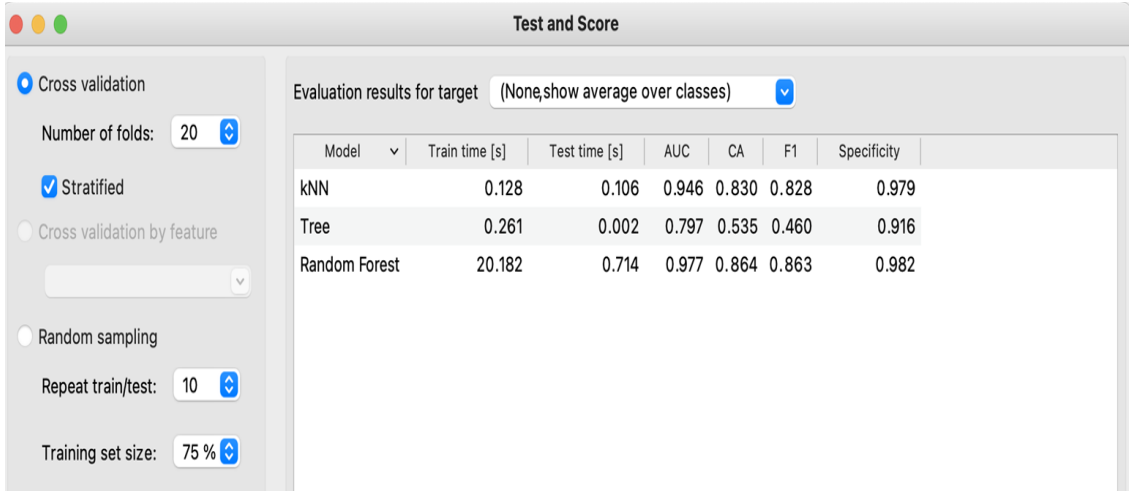
we process without impute the missing value yet seem the missing is small percent.

IV. Data Processing and Modeling (cont.)

KNN and Random forest seem has higher accuracy than tree.

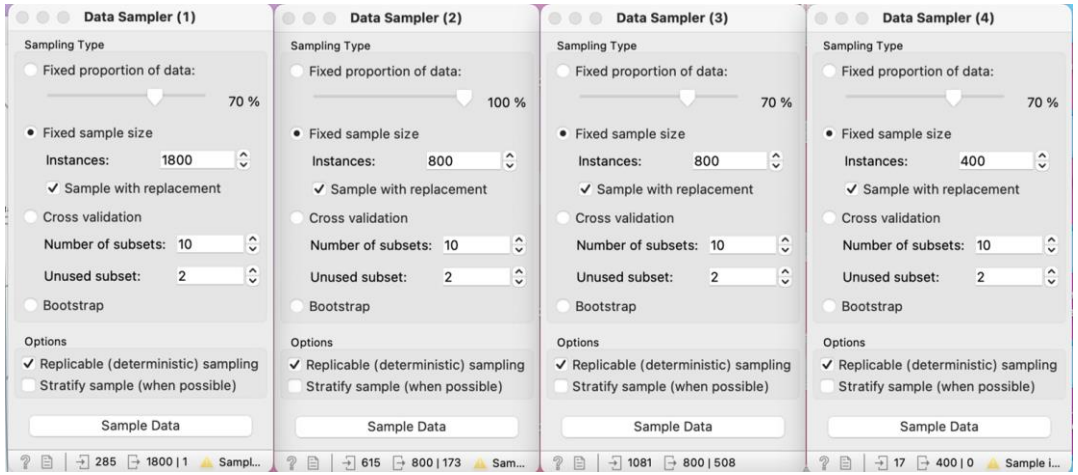
however, we will keep improve some more on classification balance.

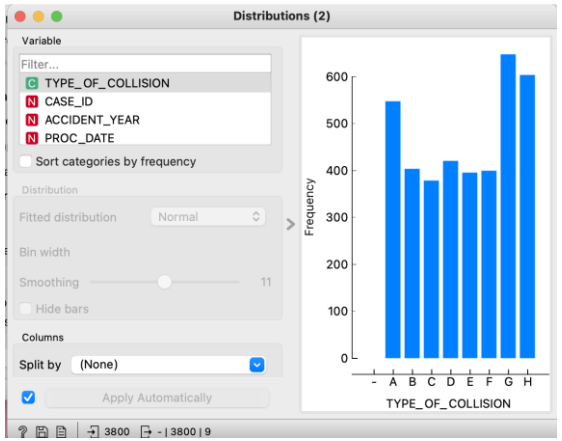


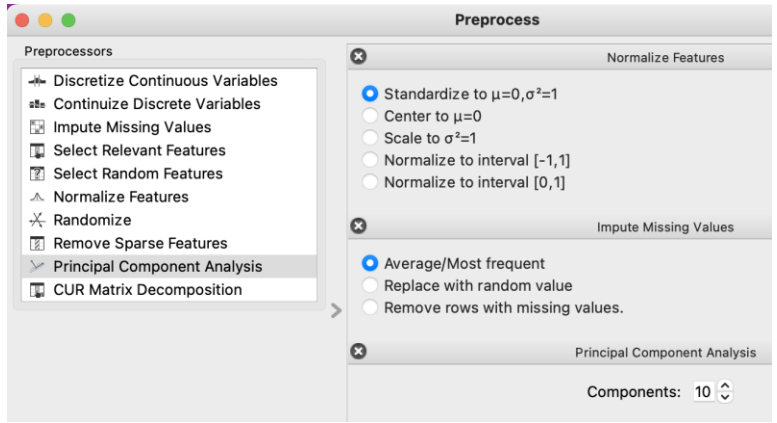


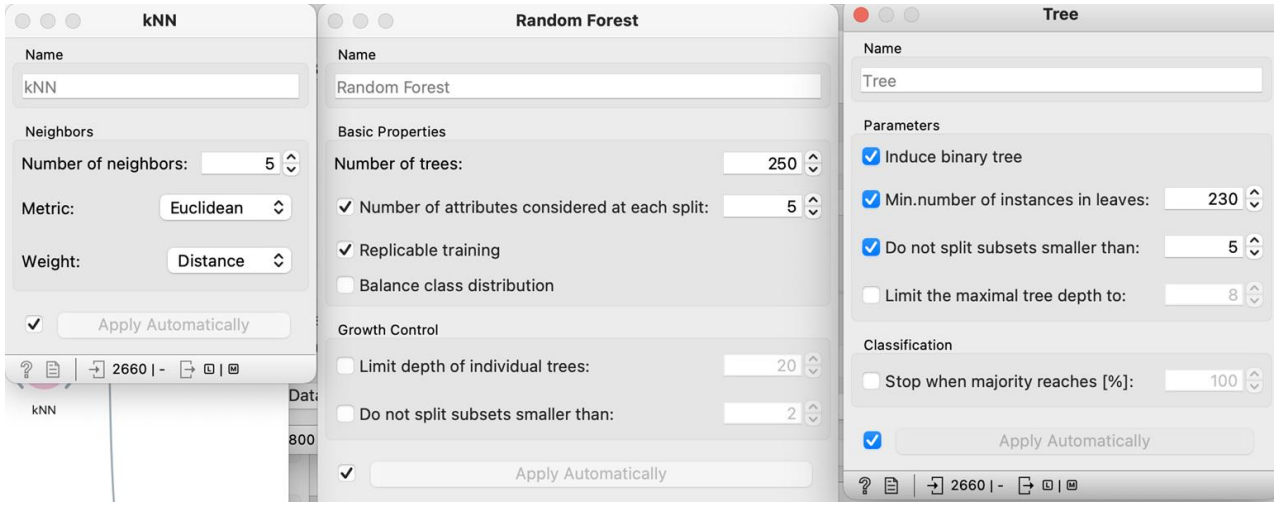
IV. Data Processing and Modeling (cont.)

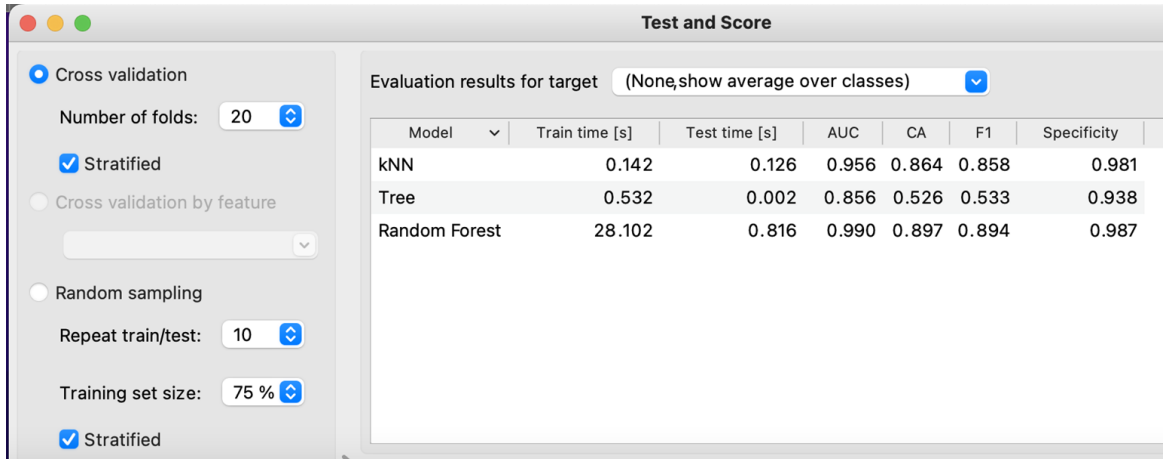
Keep adjust classification balance and add impute missing values. Model is improved and Random forest is still the highest.







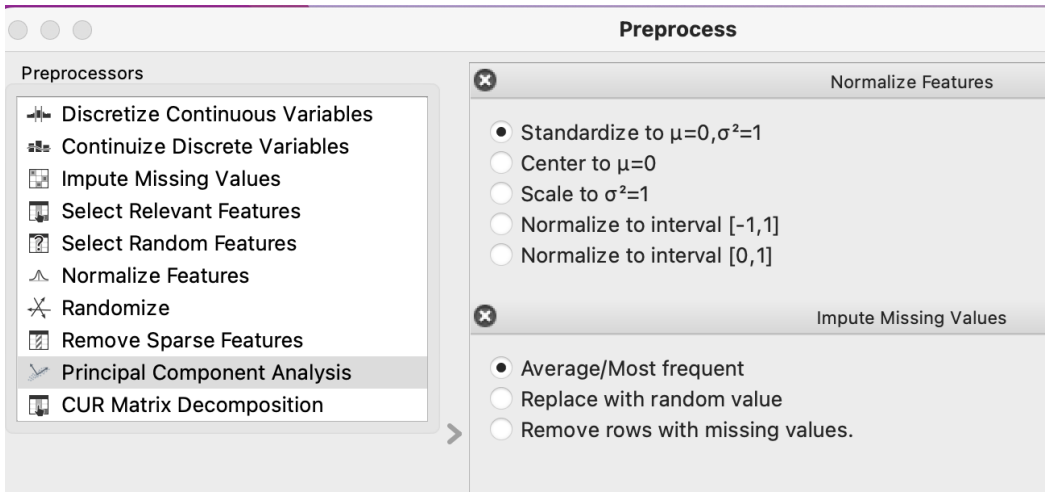


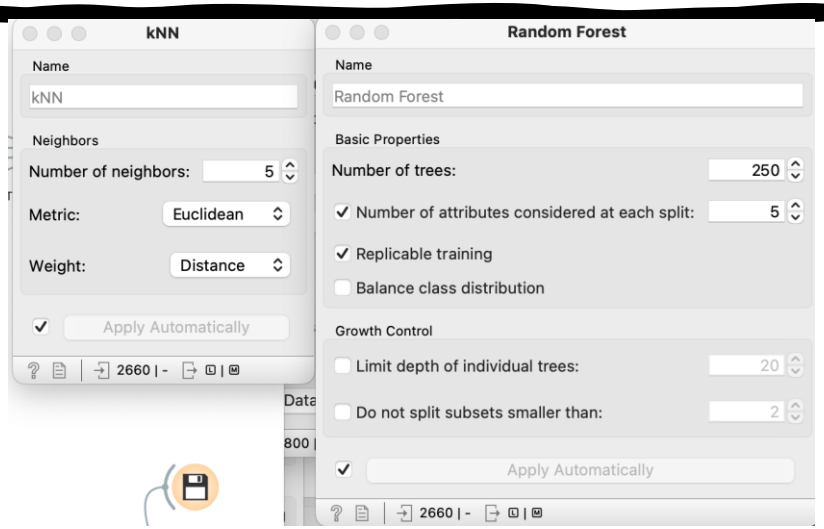


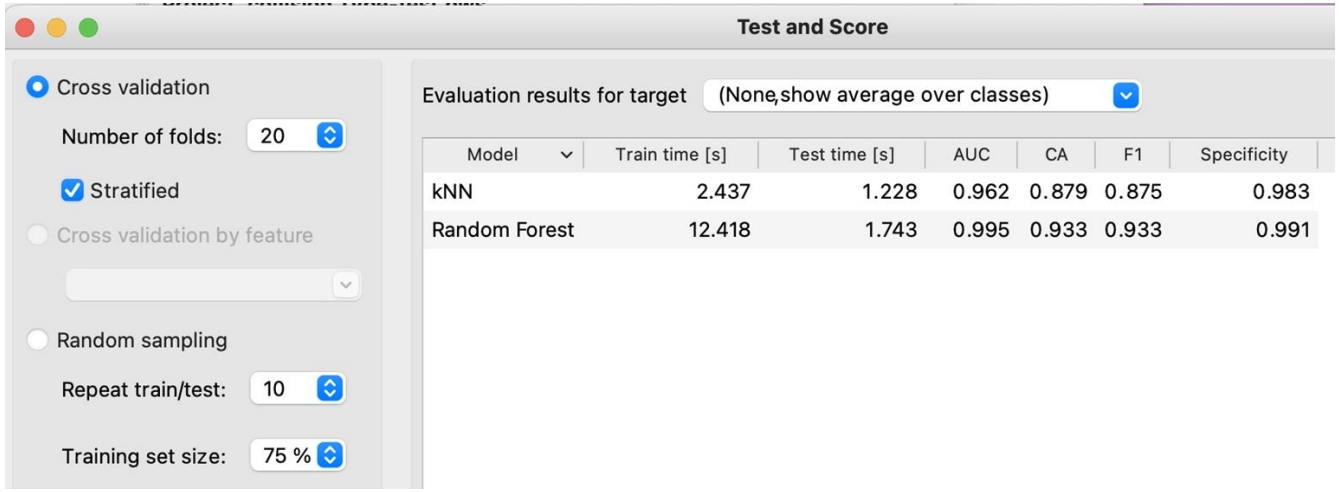
Model	Train time [s]	Test time [s]	AUC	CA	F1	Specificity
kNN	0.142	0.126	0.956	0.864	0.858	0.981
Tree	0.532	0.002	0.856	0.526	0.533	0.938
Random Forest	28.102	0.816	0.990	0.897	0.894	0.987

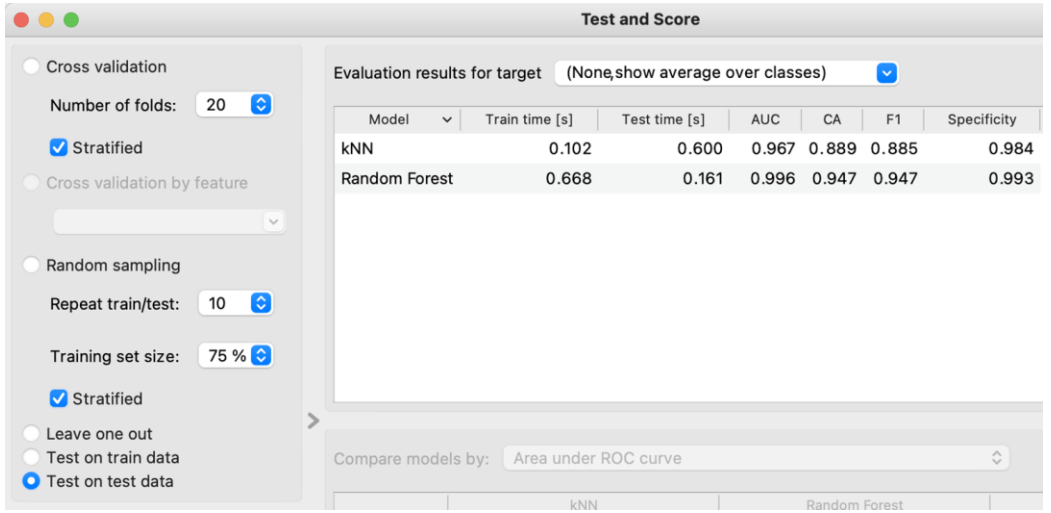
IV. Data Processing and Modeling (cont.)

Remove PCA and keep only KNN and Random forest, seem model improvement on both cross validation and test on test data



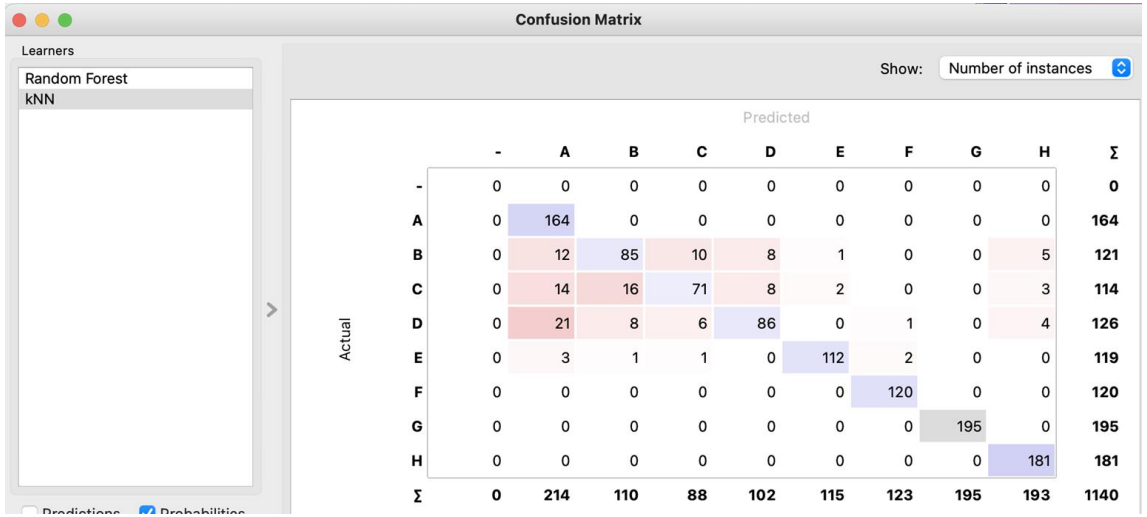


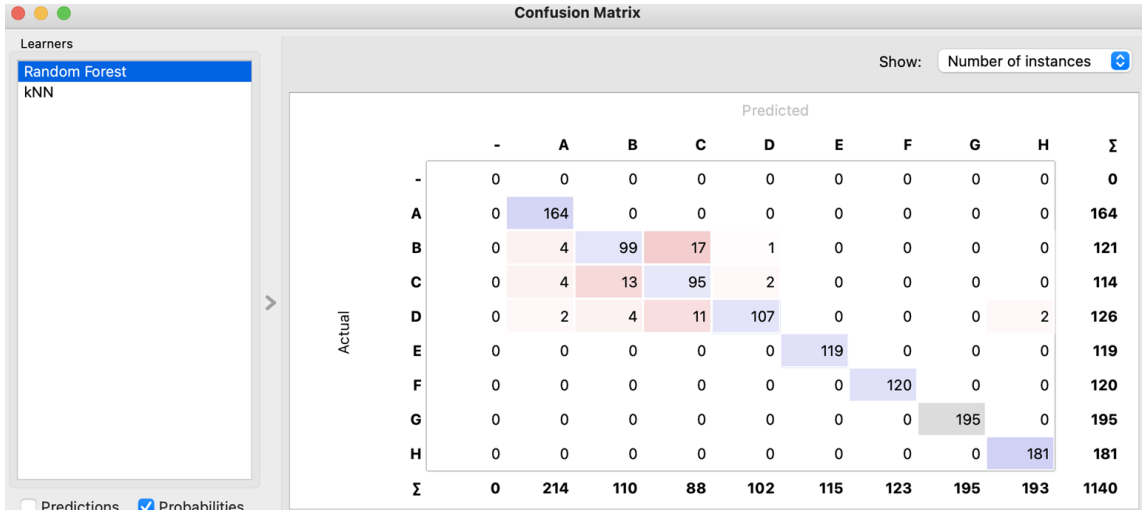




V. Model Evaluation

Random forest takes a bit longer time than KNN, however, look at the confusion matrix, Random forest has more accuracy than KNN. example, collision type classification H,G,F,E are positive and positive while others class has more positive than positive-negative.





VI. Conclusion

- We can define collision type by using most of the feature from dataset, which mean the more features are the better contribution to the model data validation.
- The best model to classify the categorical target is work best with Random forest
- In this dataset, Random forest will be the model put into production to run collision type classification.

VII. Future Plan

- Look into more details of each feature to define best feature that contribute to the collision types
- Working on processing and modeling by try different kind of models.
- Explore more on data if we need to merge other dataset.

Sources:

https://www.kaggle.com/datasets/sonicpsionic/california-switrs-collision-reports

Thank You!