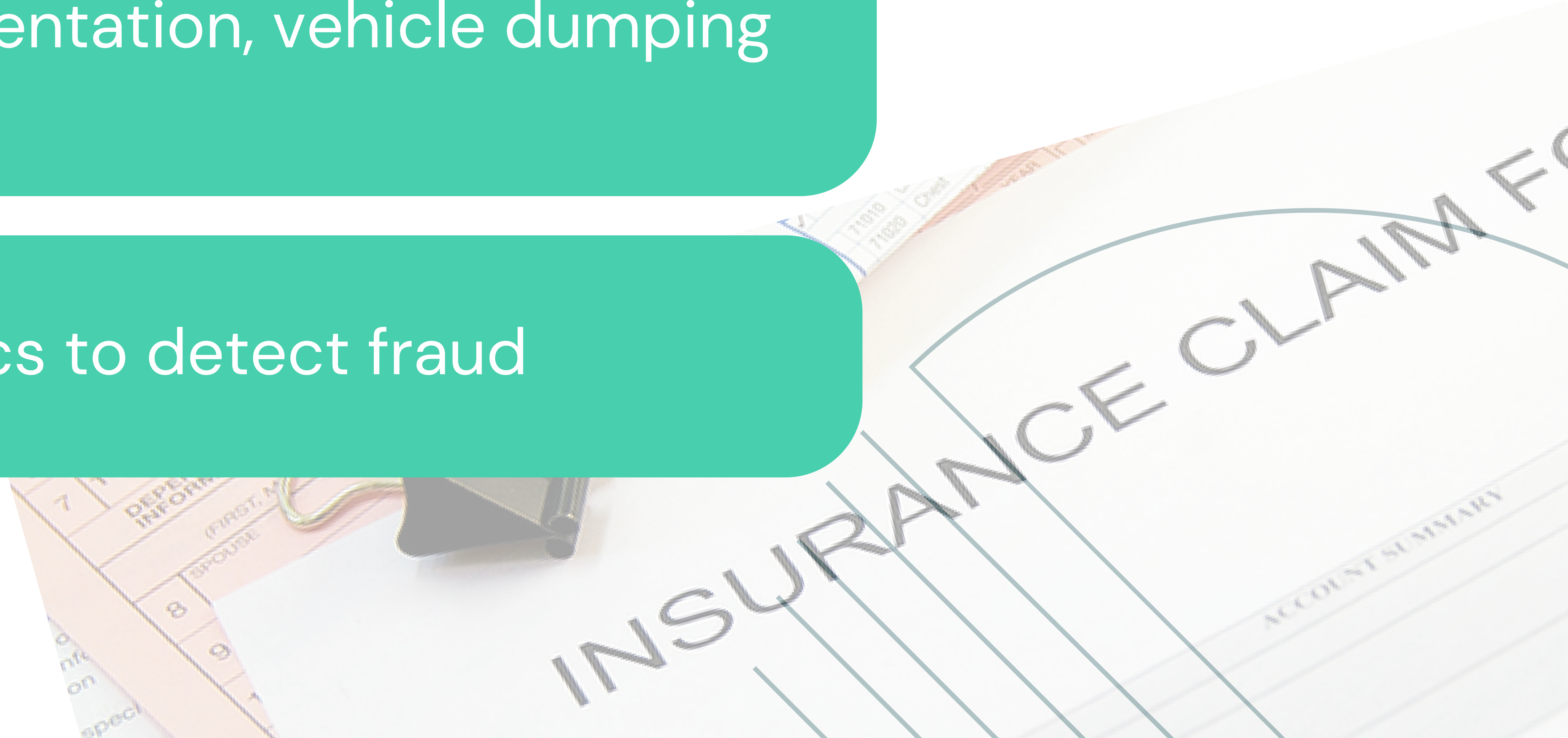
FRAUD DETECTION

FINAL PROJECT

Sepi & Yana

CAR INSURANCE FRAUD

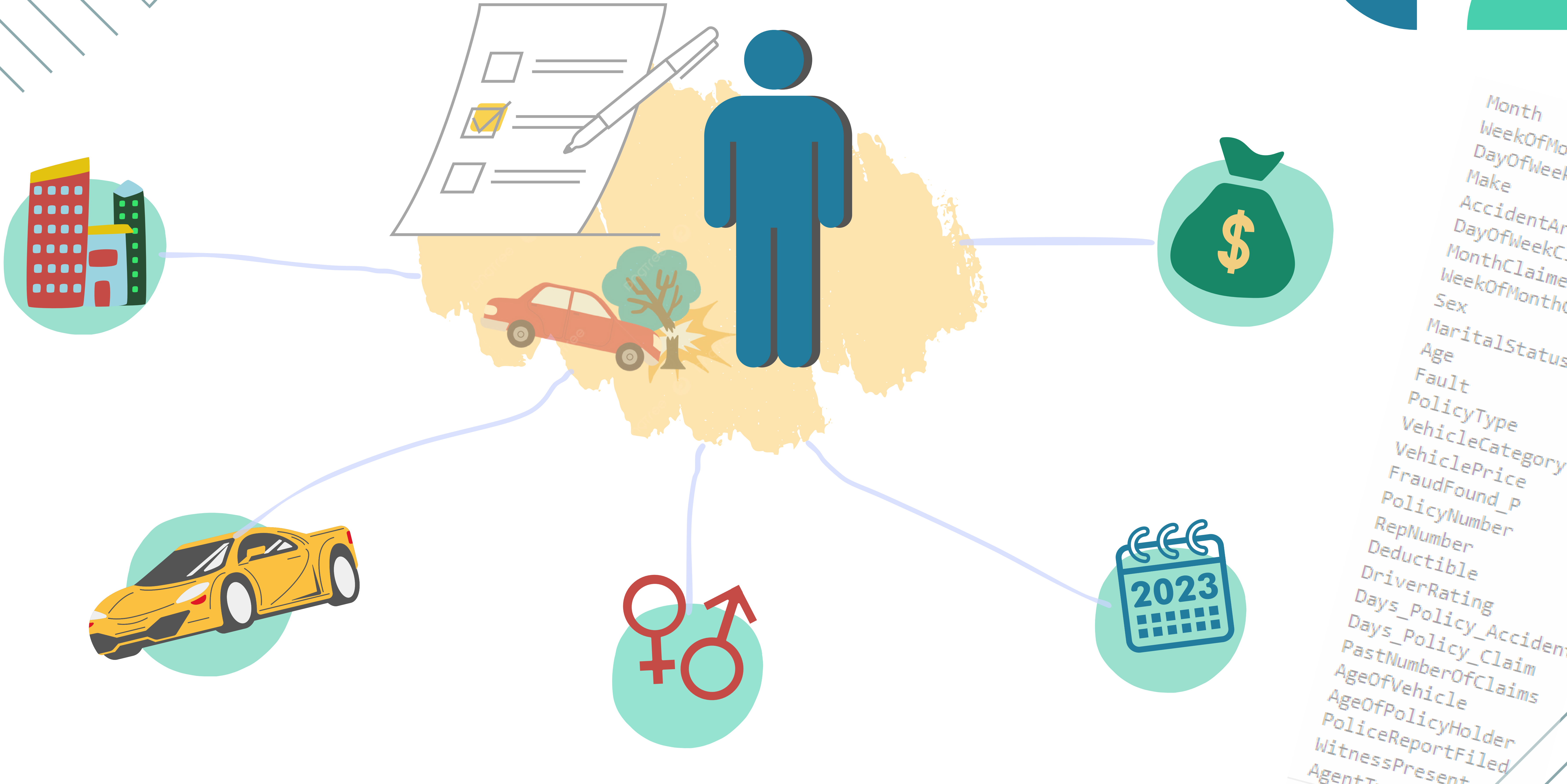
- Lying to the insurance company for financial gain
- Staged accidents, exaggerated claims, false documentation, vehicle dumping etc
- Data analytics to detect fraud



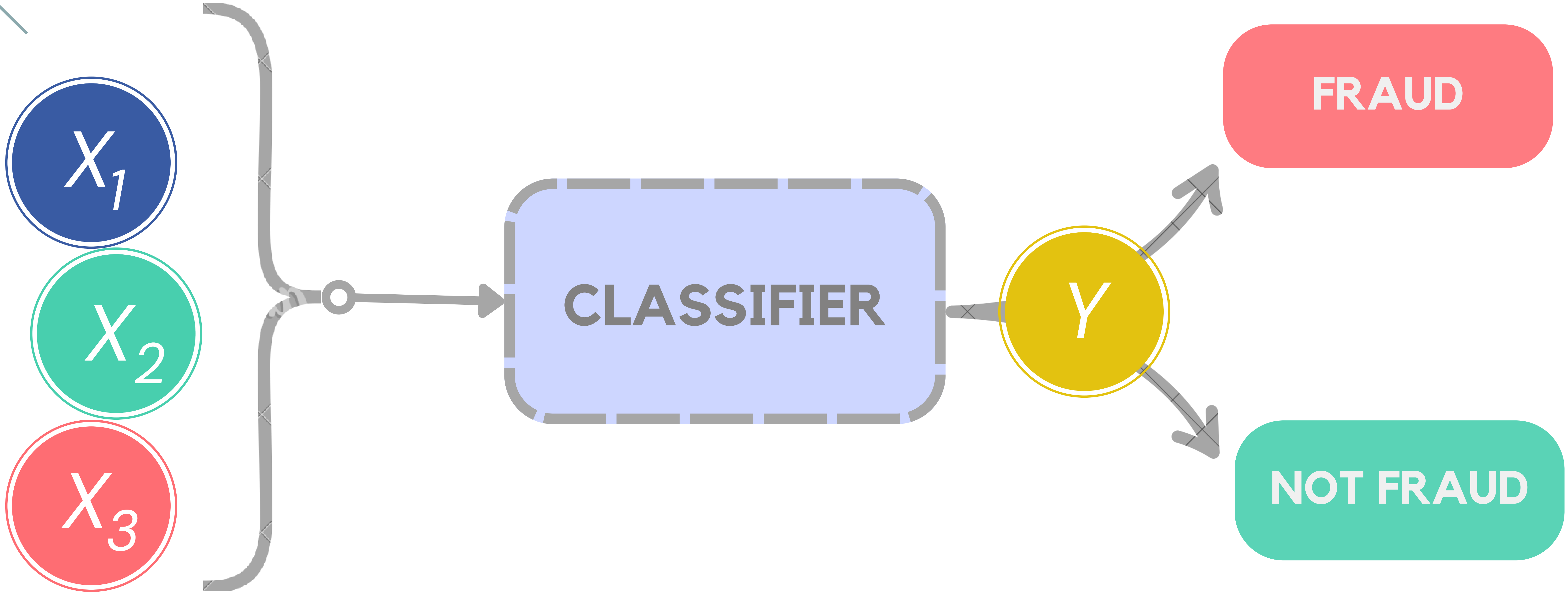
BUILDING A FRAUD DETECTION MODEL

- Exploratory Data Analysis
- Feature Engineering
- Resampling the Dataset
- Model Training & Evaluation

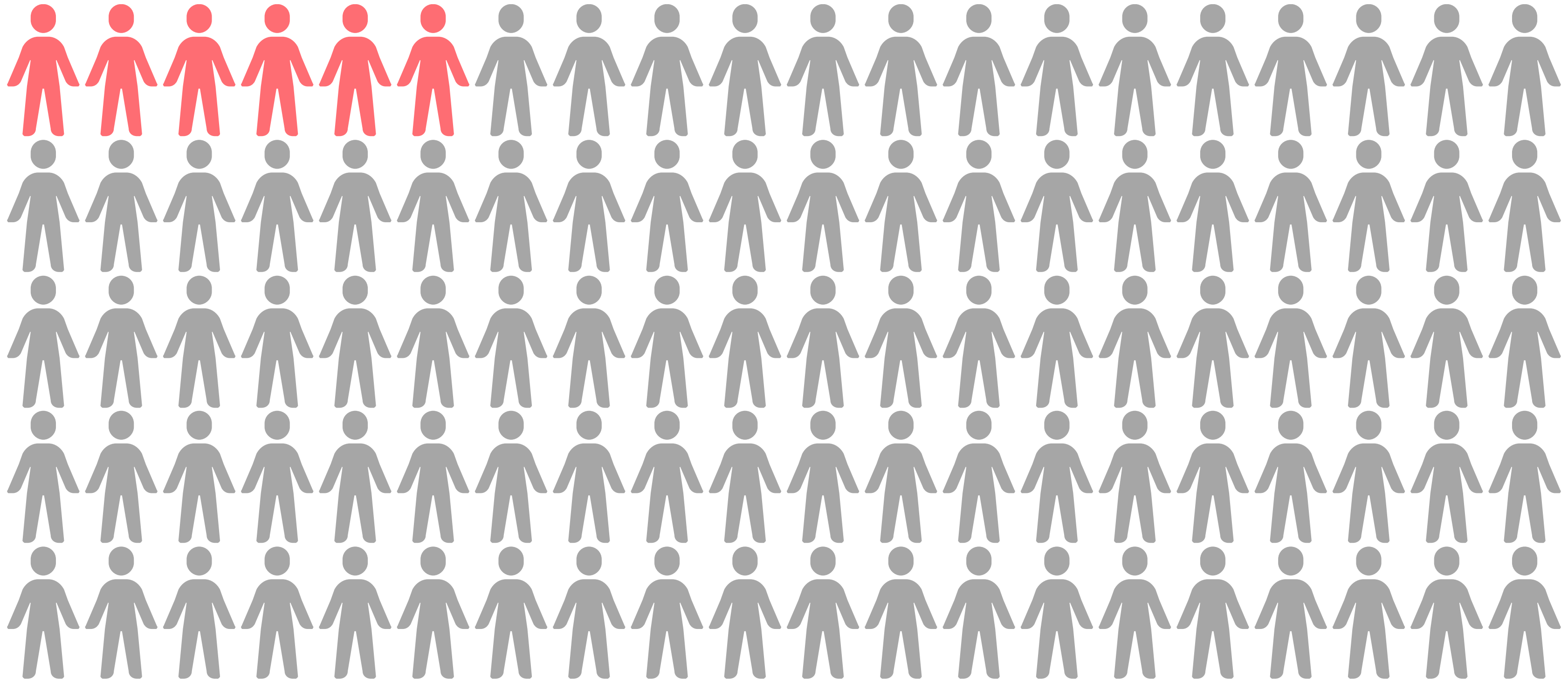
CAR INSURANCE CLAIM



BUILDING A FRAUD DETECTION MODEL



FRAUD IS RARE

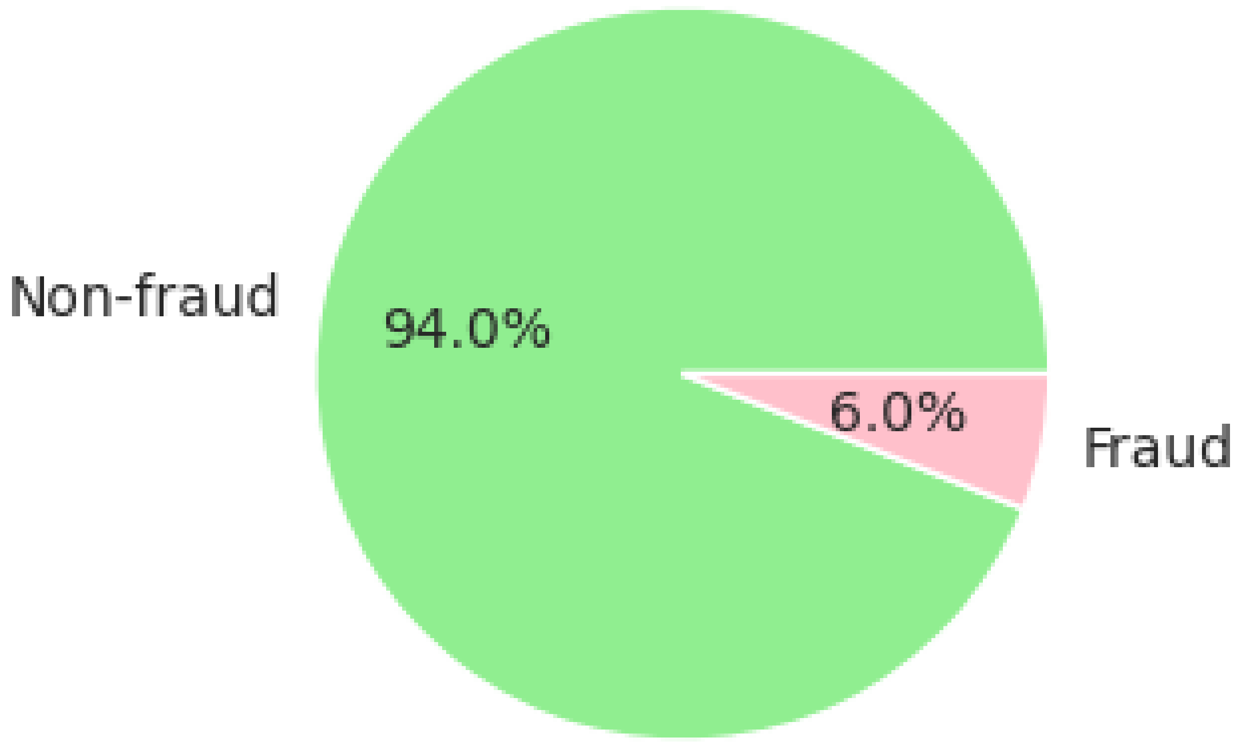


6%

Car Insurance claims are fraud cases

Claims	Percentage
14497	94.01
923	5.99

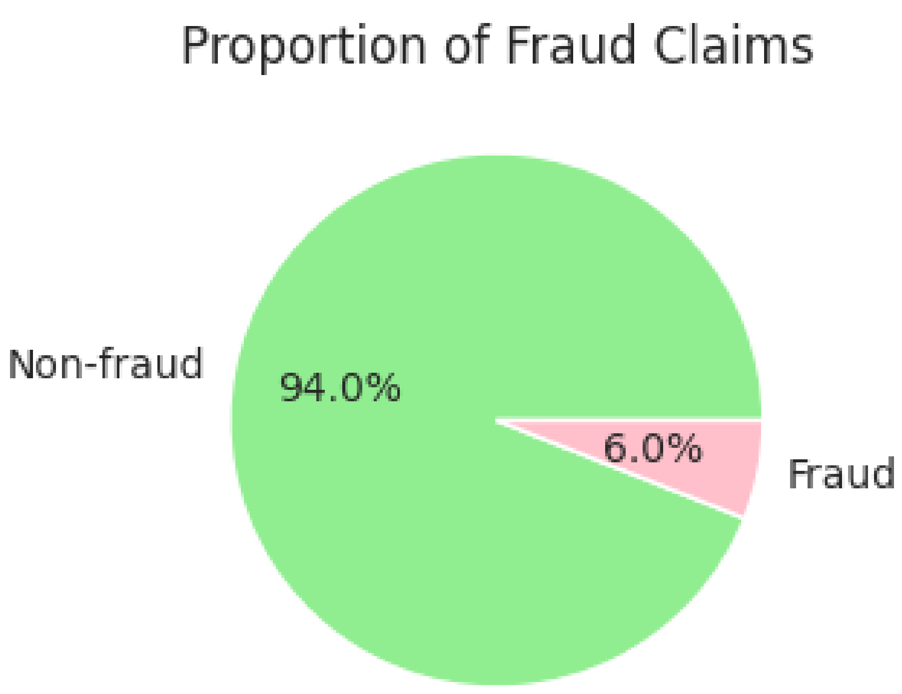
Proportion of Fraud Claims



CHI-SQUARE TEST

ASSOCIATION BETWEEN TWO VARIABLES

AccidentArea $\longleftrightarrow$ Fraud



Expected

	Fraud	Not Fraud
Rural	6%	94%
Urban	6%	94%

Variable	Ch
PolicyType	437.4913
VehicleCategory_BasePolicy	437.4904
BasePolicy	402.9472
VehicleCategory	290.98089
Fault	264.98455
Age	106.14445
AddressChange_Claim	104.72269
Deductible	72.406255
VehiclePrice	67.836116
Make	59.815292
PastNumberOfClaims	53.541755
MonthClaimed	42.200514
AgeOfPolicyHolder	33.104861
Age_Bracket	29.809235
Month	29.771469
AgeOfVehicle	21.995137
NumberOfSuppliments	18.155527
AccidentArea	16.901858
Sex	13.495678
RepNumber	11.799995
Days_Policy_Accident	11.569842
DayOfWeek	[illegible]

CHI-SQUARE TEST

ASSOCIATION BETWEEN TWO VARIABLES

AccidentArea ⟷ Fraud

Expected

	Fraud	Not Fraud
Rural	6%	94%
Urban	6%	94%

Observed

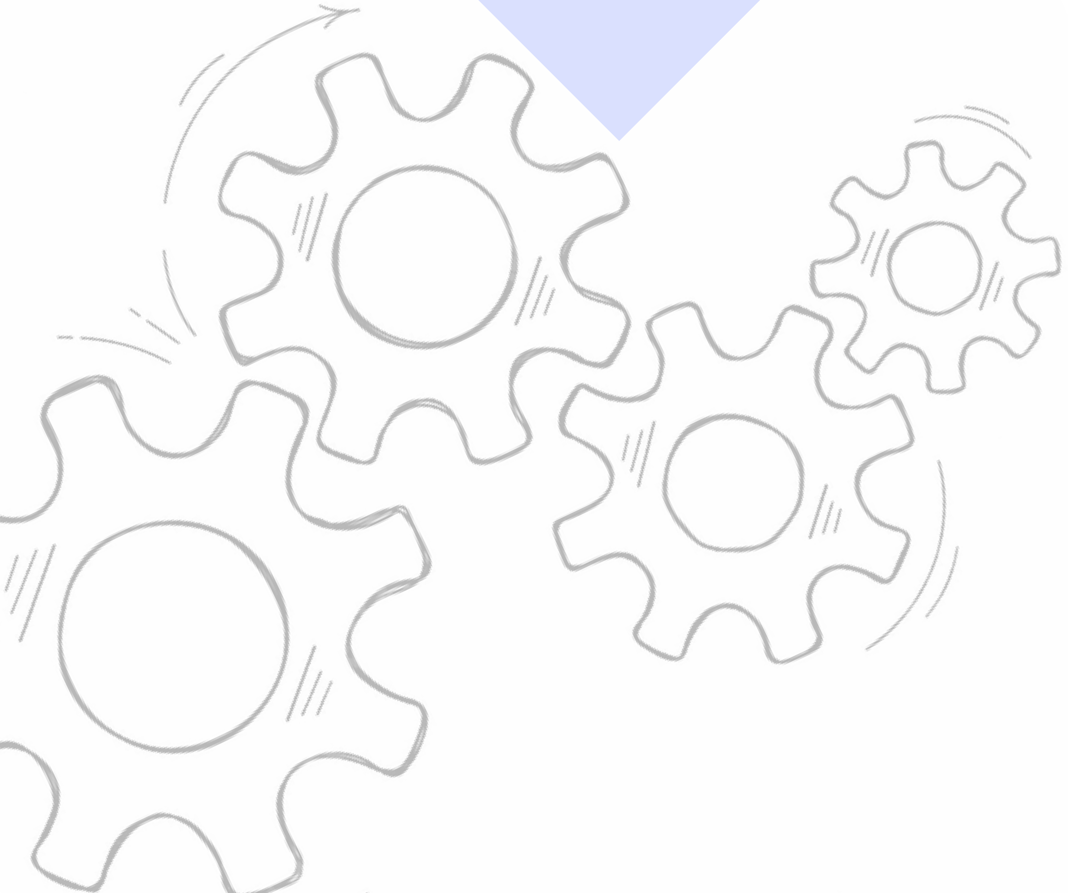
	Fraud	Not Fraud
Rural	8.32%	91.68%
Urban	5.72%	94.28%

Variable	Chi2	P-value
PolicyType	437.491381	1.768441e-89
VehicleCategory_BasePolicy	437.490455	2.154735e-90
BasePolicy	402.947238	3.170436e-88
VehicleCategory	290.980893	6.520817e-64
Fault	264.984556	1.406180e-59
Age	106.144451	7.331495e-04
AddressChange_Claim	104.722693	9.704718e-22
Deductible	72.406255	1.302831e-15
VehiclePrice	67.836116	2.888324e-13
Make	59.815292	2.191573e-06
PastNumberOfClaims	53.541755	1.405198e-11
MonthClaimed	42.200514	1.495245e-05
AgeOfPolicyHolder	33.104861	5.896560e-05
Age_Bracket	29.809235	2.284330e-04
Month	29.771469	1.720902e-03
AgeOfVehicle	21.995137	2.545[illegible]
NumberOfSuppliment[illegible]	[illegible]	[illegible]

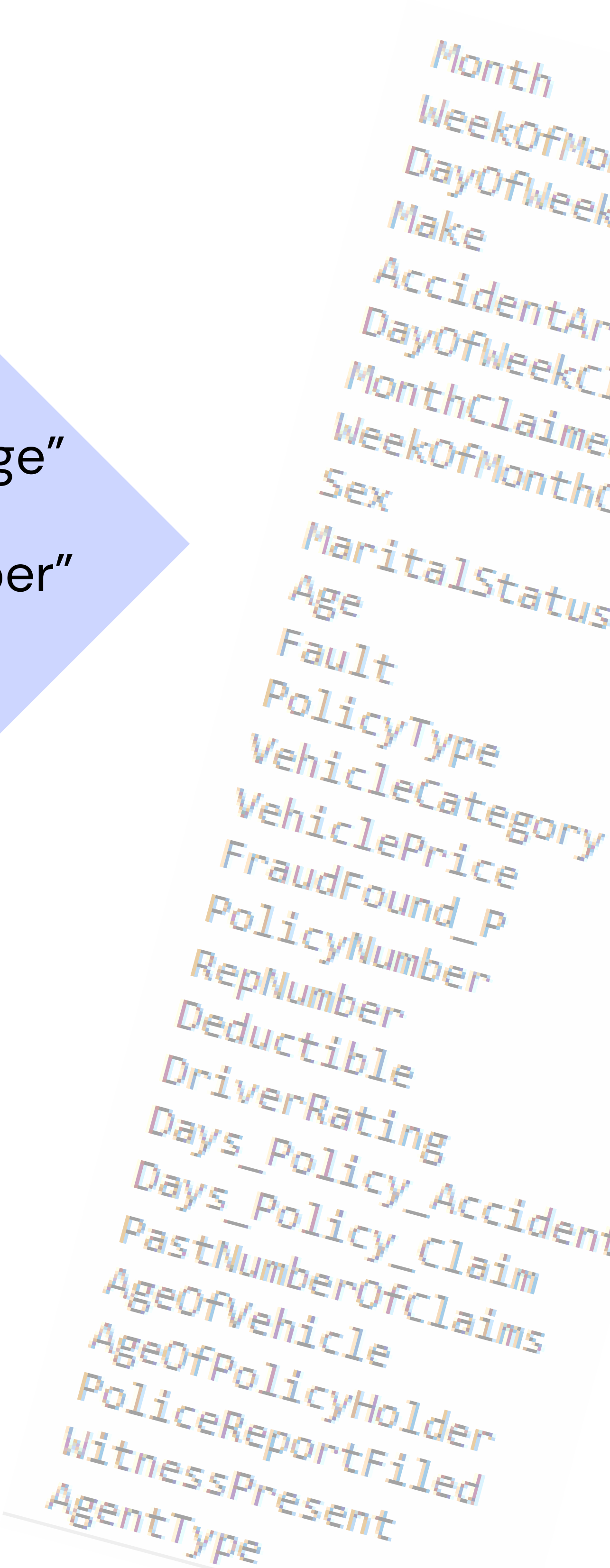
SIMPLIFYING OUR DATA: PREPROCESSING

33 variables
24 categorical
9 numerical

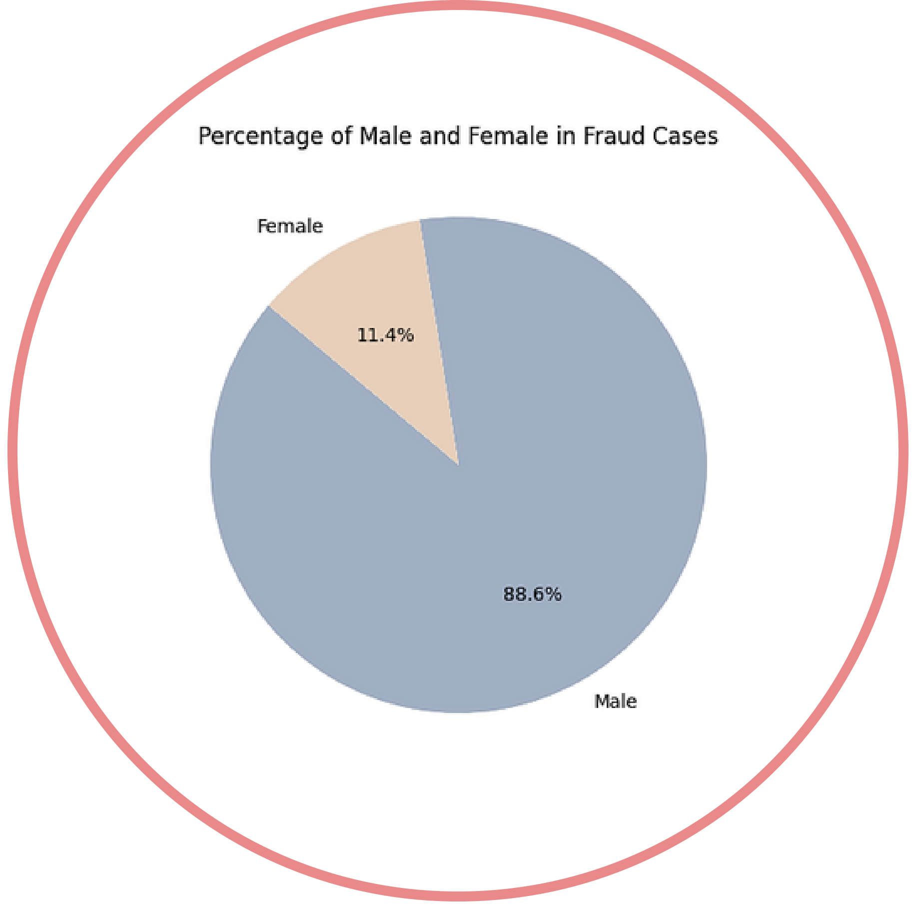
"Age": Imputed -0- values with mean → Dropped rows with -0- values in dates → Checked feature importance on the target variable using Chi-square test → Dropped "Age" and "PolicyNumber" columns

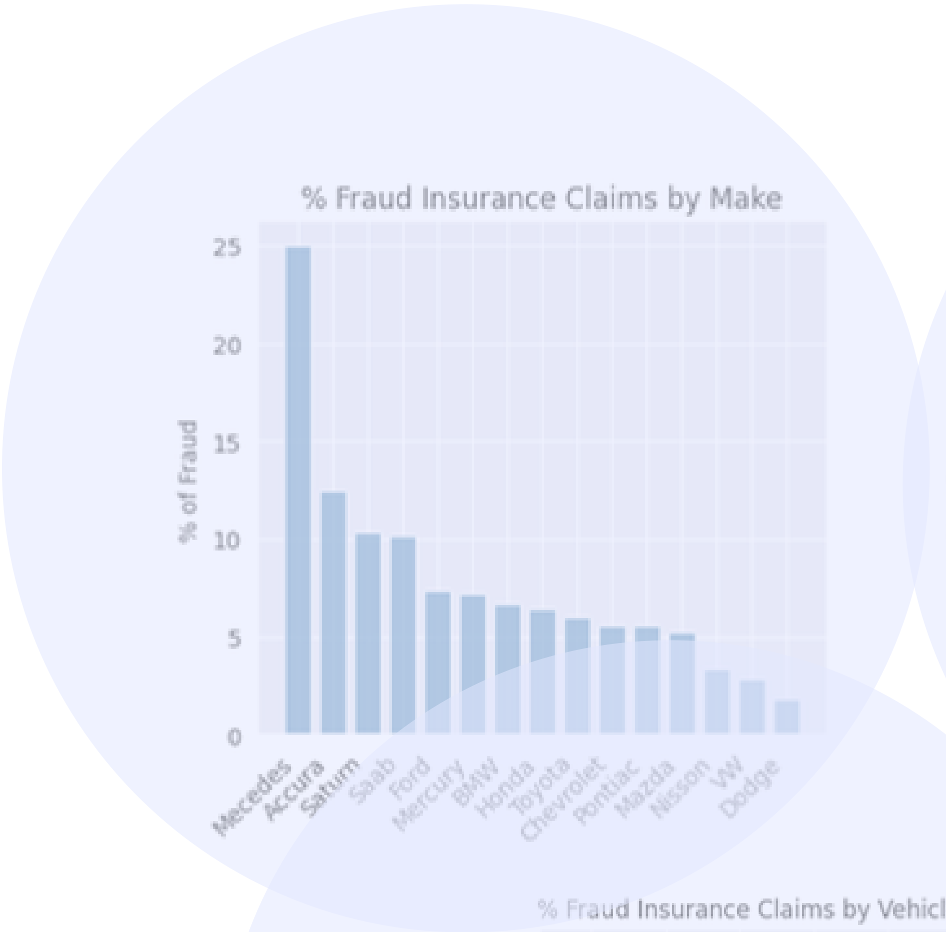


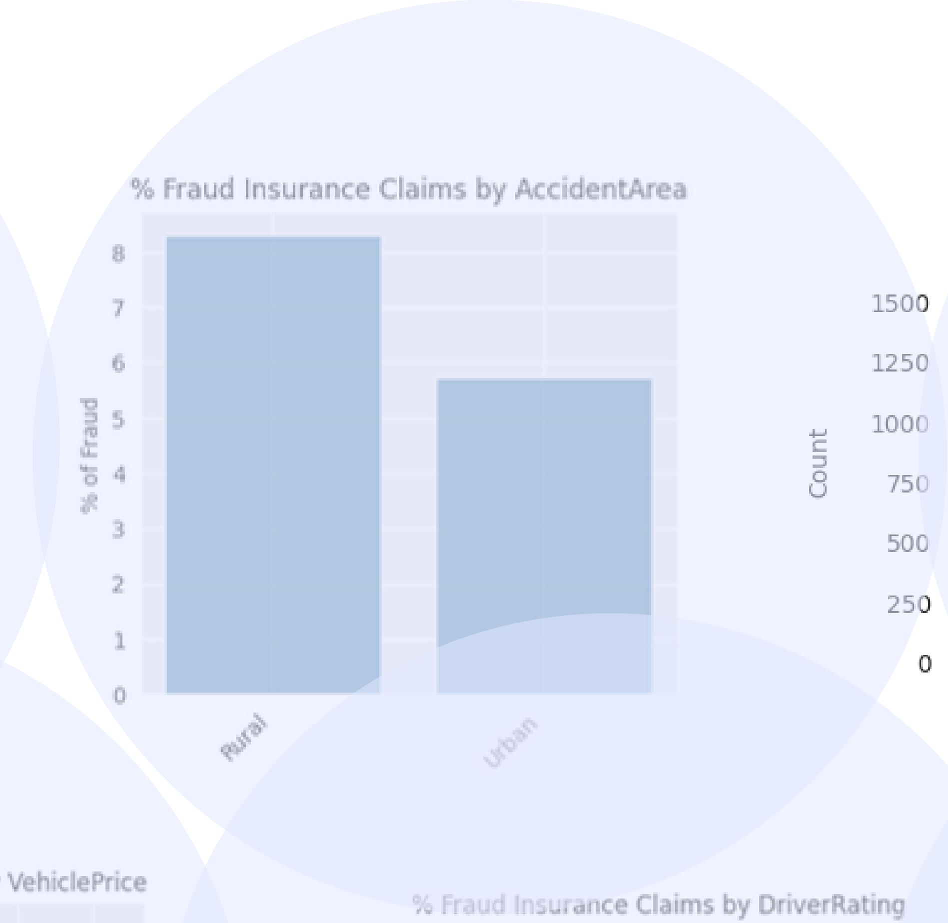
After a few back and forth with the dataset, we decided to keep most of the columns and just dropped "Age" and "Policy Number"

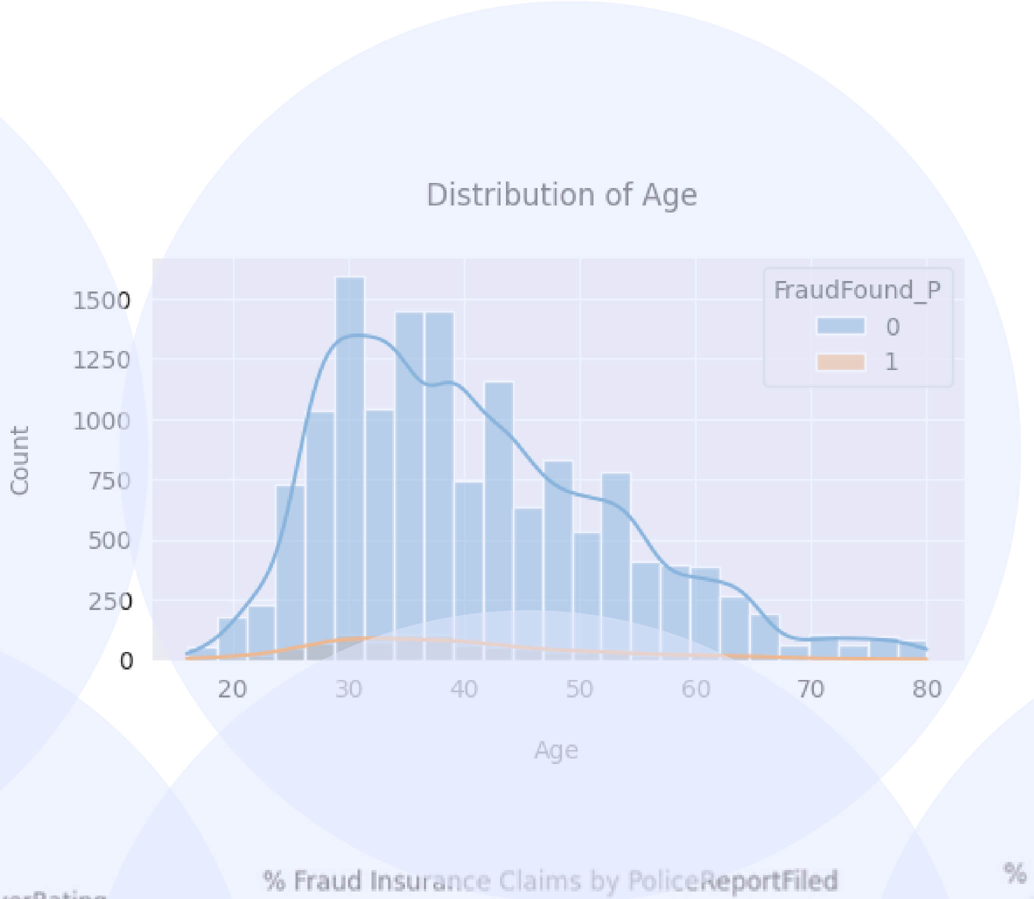


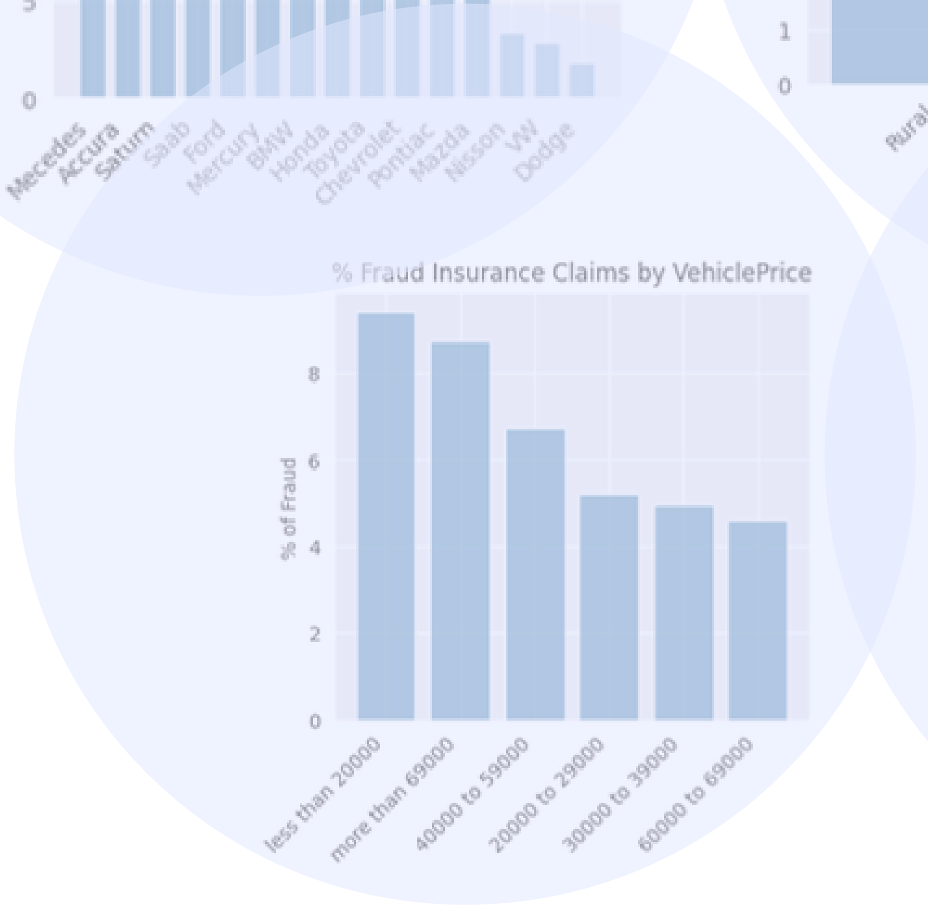
LETS MAKE IT VISUAL: EXPLORATORY DATA ANALYSIS

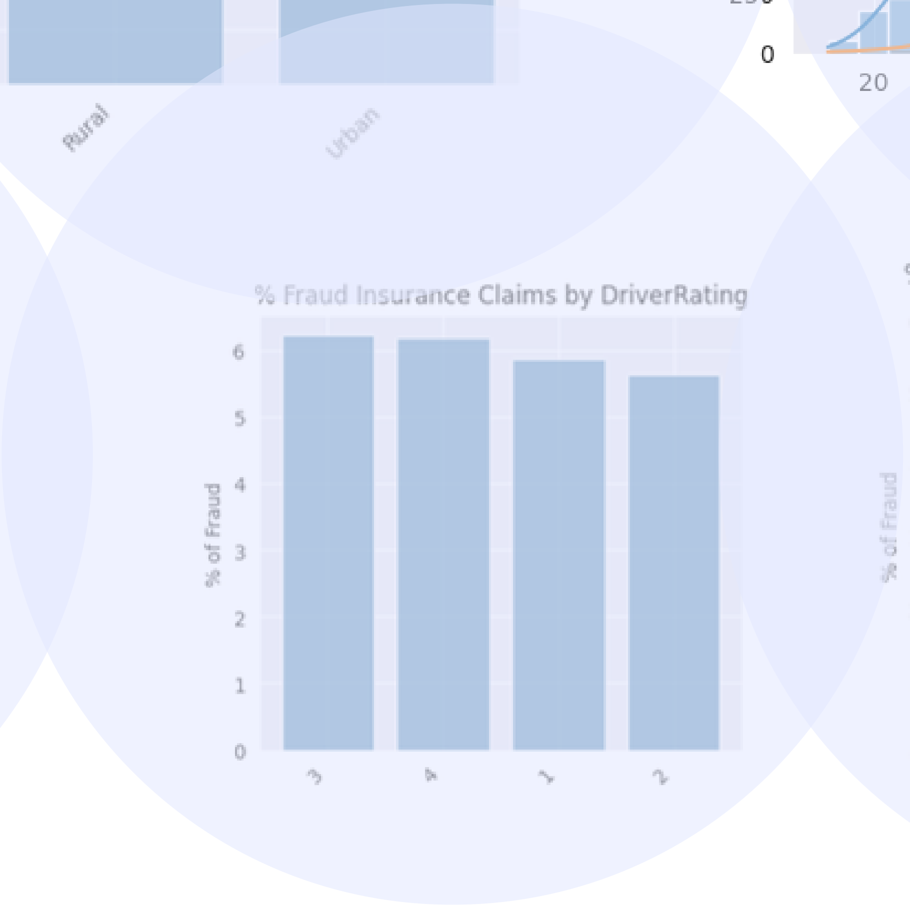


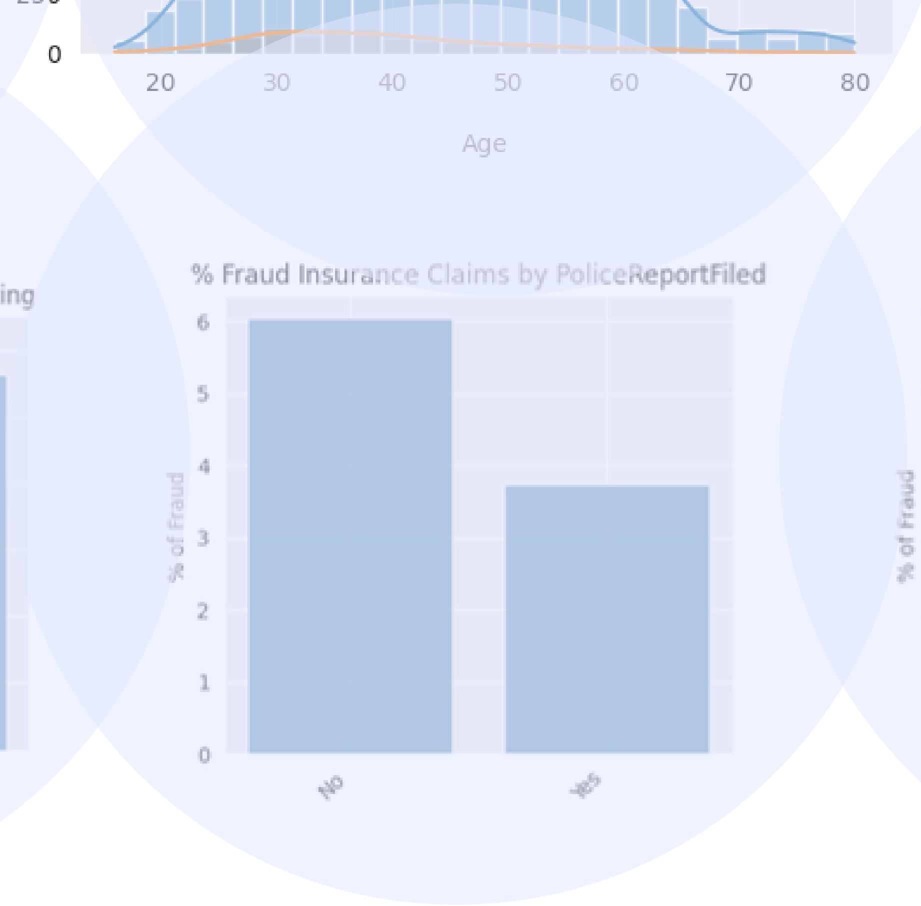




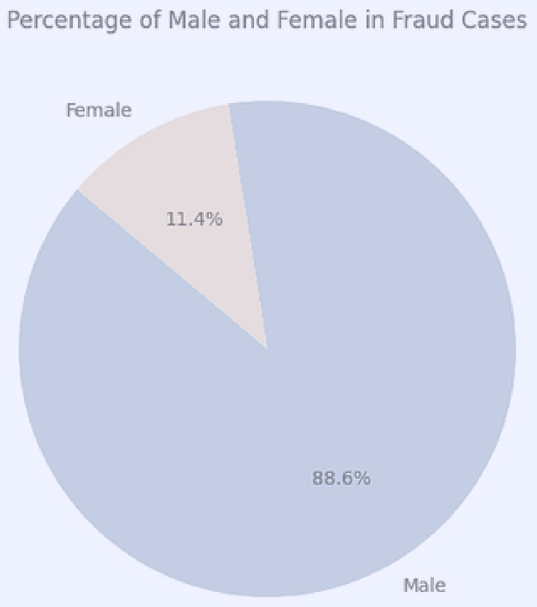




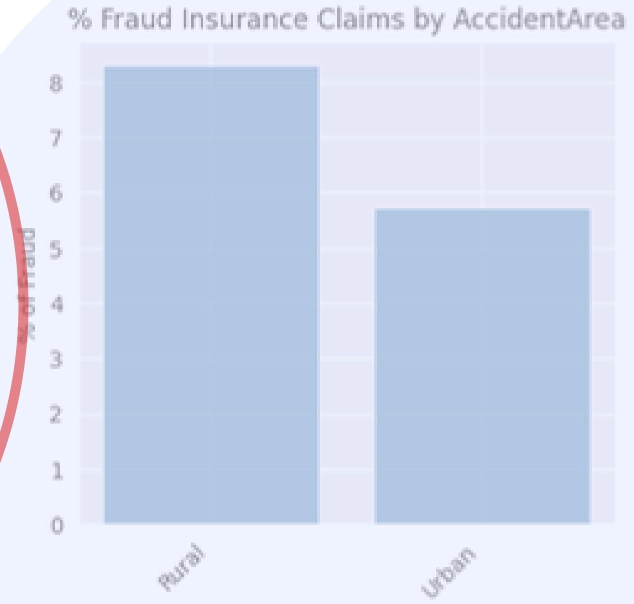


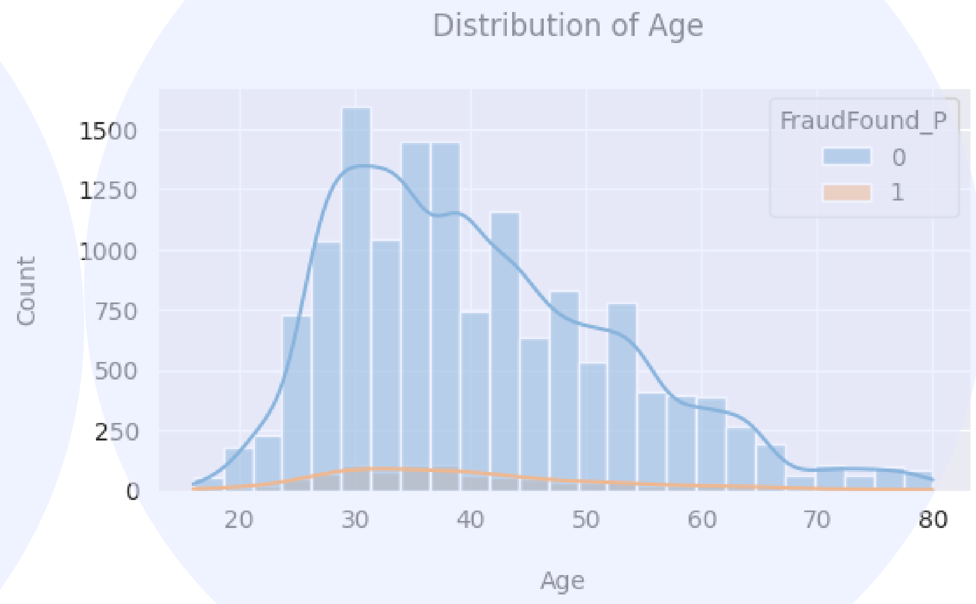


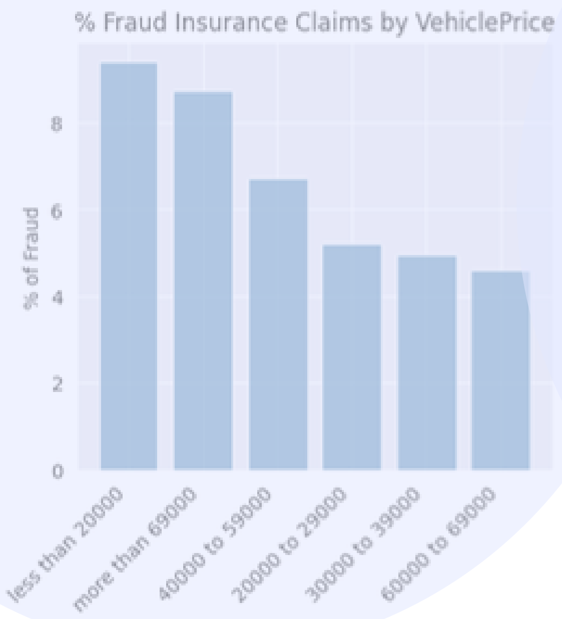
EDA

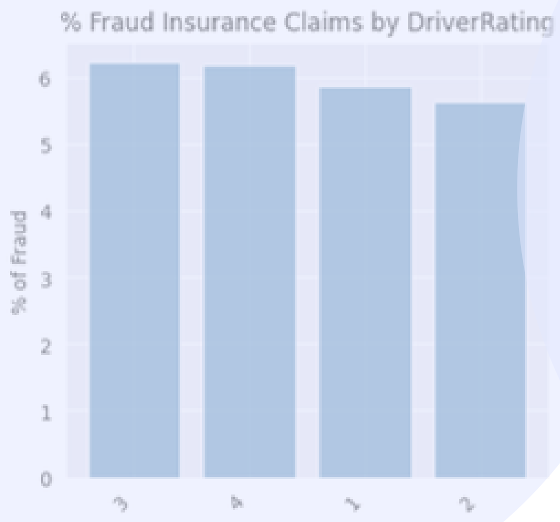

Percentage of Male and Female in Fraud Cases
Female
11.4%
88.6%
Male

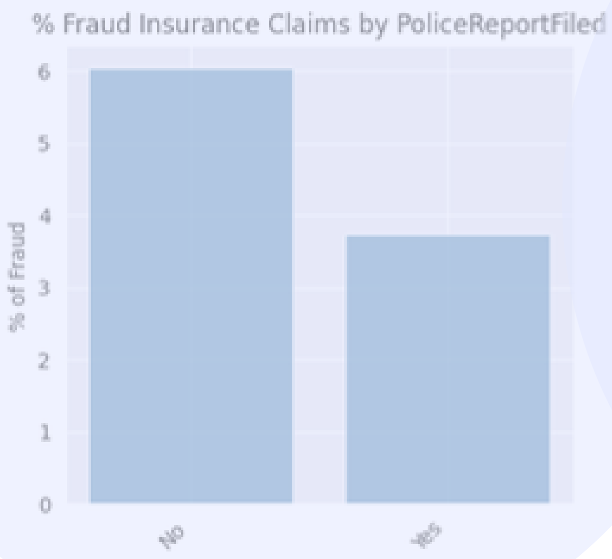

% Fraud Insurance Claims by Make
% of Fraud
Mecedes
Accura
Saturn
Saab
Ford
Mercury
BMW
Honda
Toyota
Chevrolet
Pontiac
Mazda
Nisson
VW
Dodge

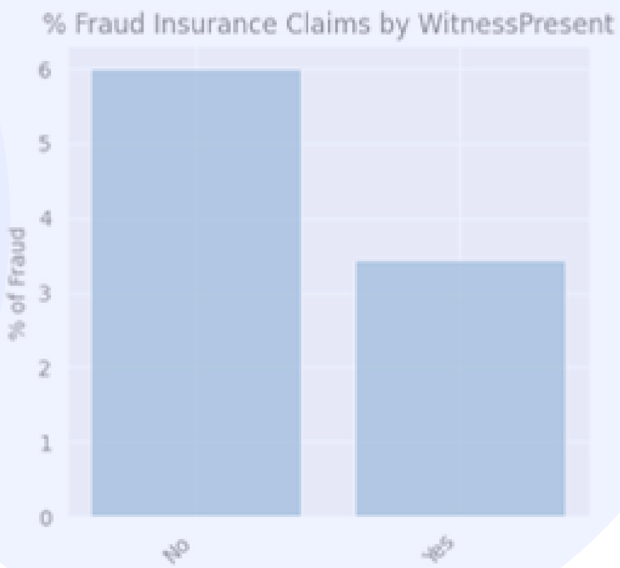

% Fraud Insurance Claims by AccidentArea
Rural
Urban


Distribution of Age
FraudFound_P
0
1
Count
Age

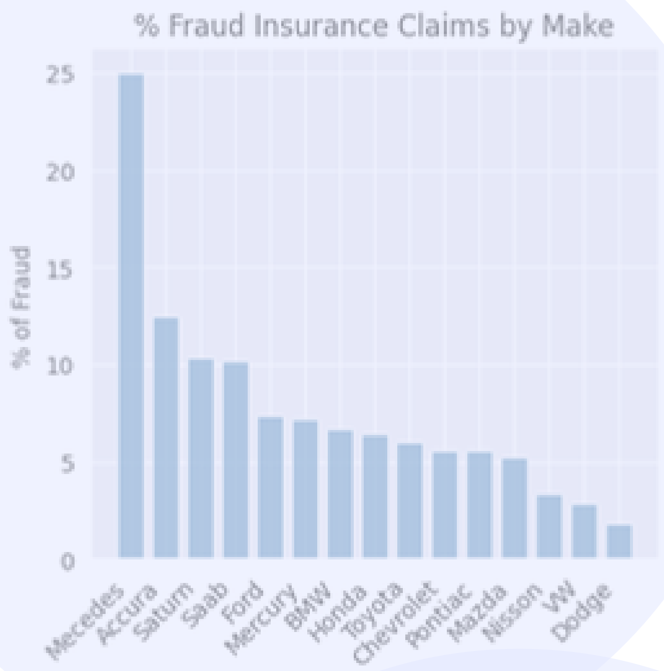

% Fraud Insurance Claims by VehiclePrice
% of Fraud
less than 20000
more than 69000
40000 to 59000
20000 to 29000
30000 to 39000
60000 to 69000


% Fraud Insurance Claims by DriverRating
% of Fraud

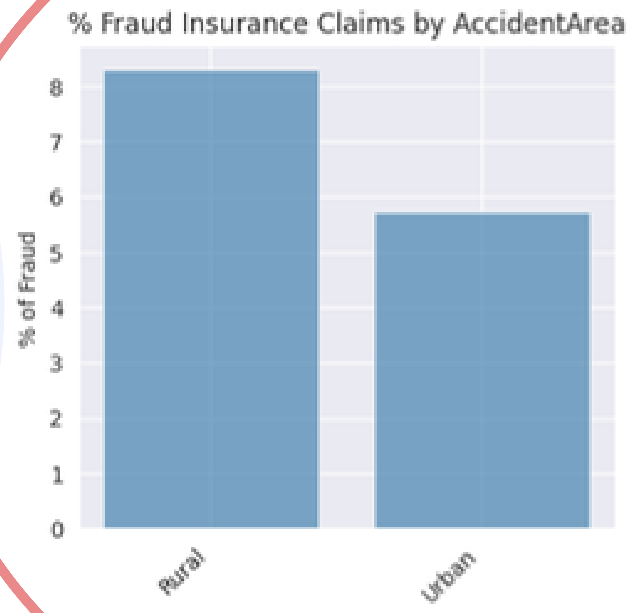

% Fraud Insurance Claims by PoliceReportFiled
% of Fraud
No
Yes


% Fraud Insurance Claims by WitnessPresent
% of Fraud
No
Yes

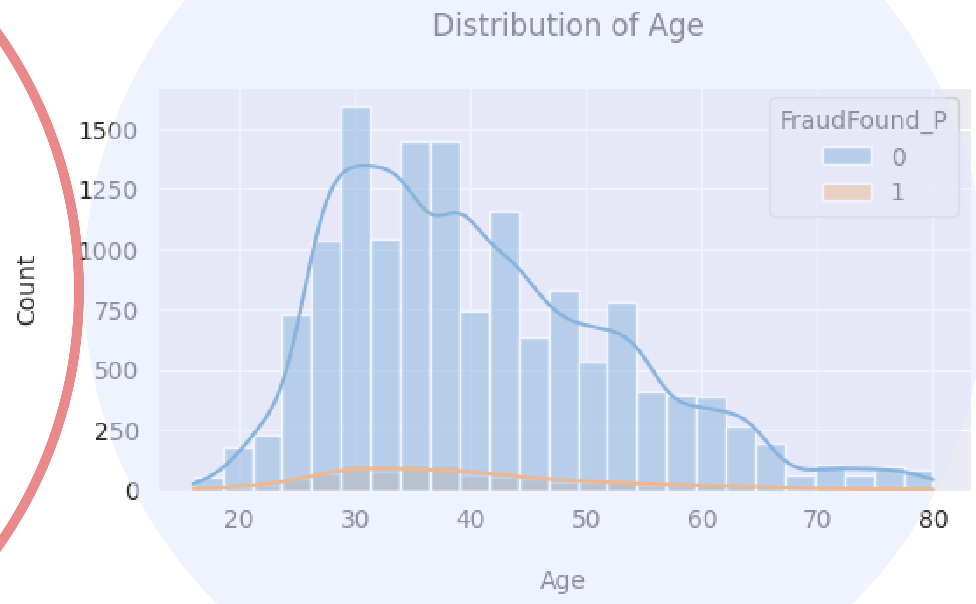
EDA



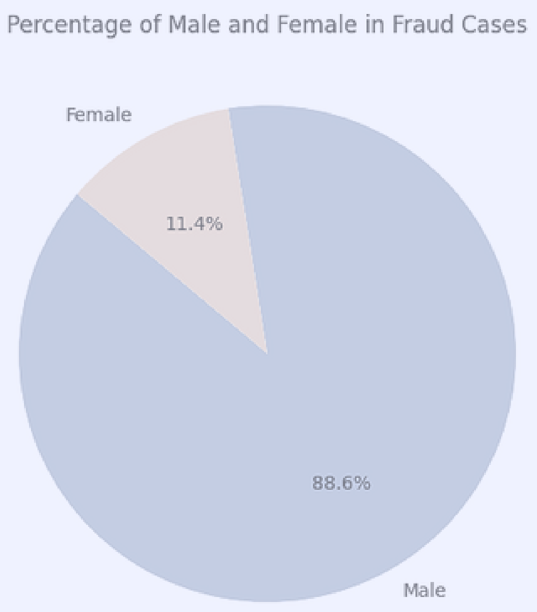
% Fraud Insurance Claims by Make
% of Fraud
Mecedes
Accura
Saturn
Saab
Ford
Mercury
BMW
Honda
Toyota
Chevrolet
Pontiac
Mazda
Nisson
VW
Dodge



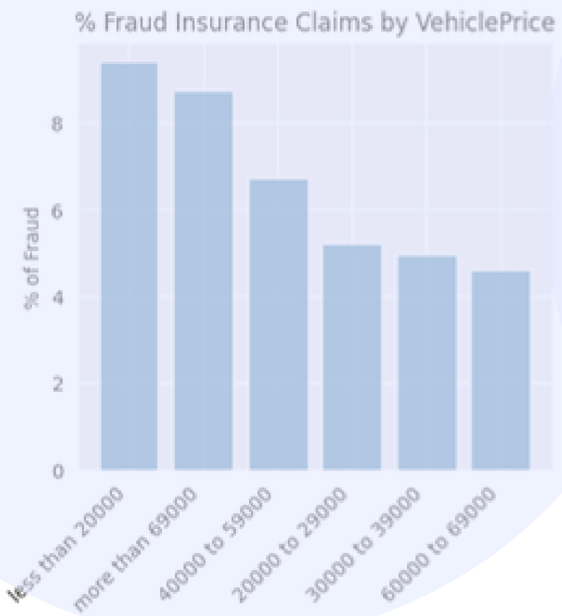
% Fraud Insurance Claims by AccidentArea
% of Fraud
Rural
Urban



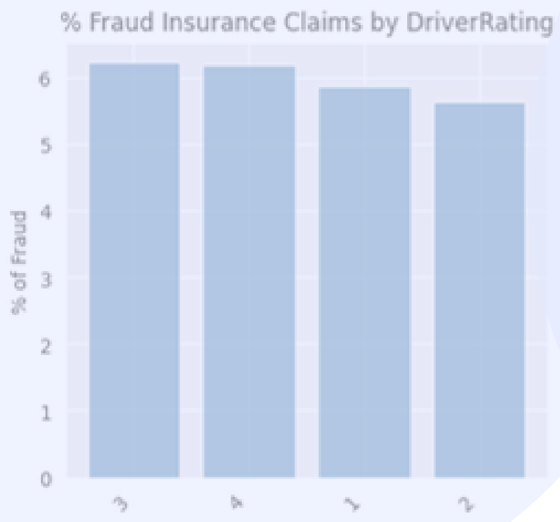
Distribution of Age
FraudFound_P
0
1
Count
Age



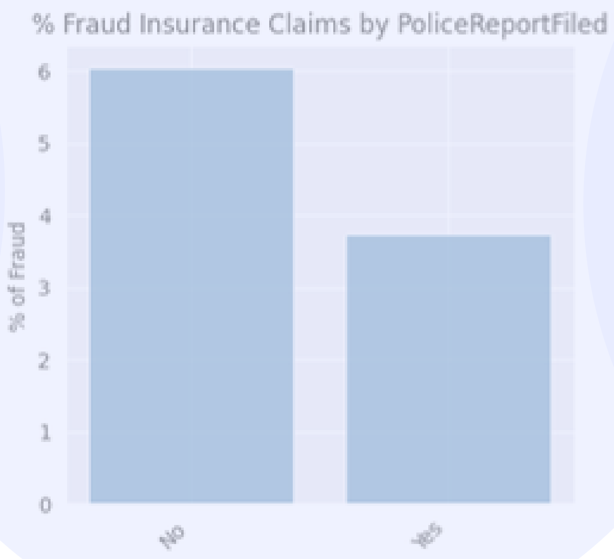
Percentage of Male and Female in Fraud Cases
Female
11.4%
88.6%
Male



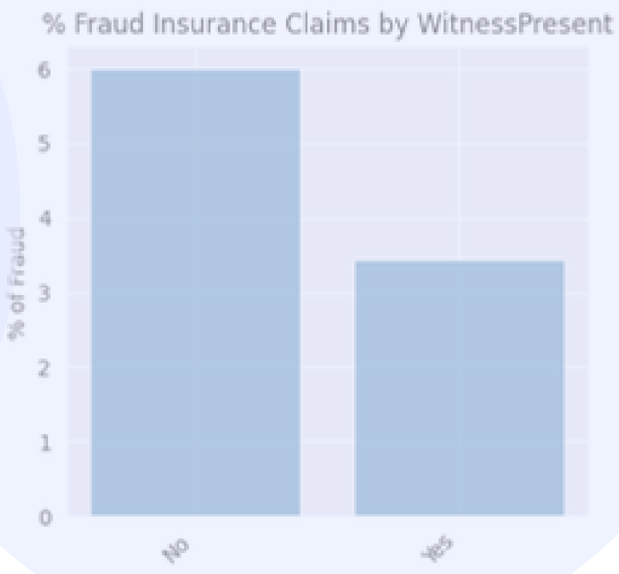
% Fraud Insurance Claims by VehiclePrice
% of Fraud
less than 20000
more than 69000
40000 to 59000
20000 to 29000
30000 to 39000
60000 to 69000



% Fraud Insurance Claims by DriverRating
% of Fraud

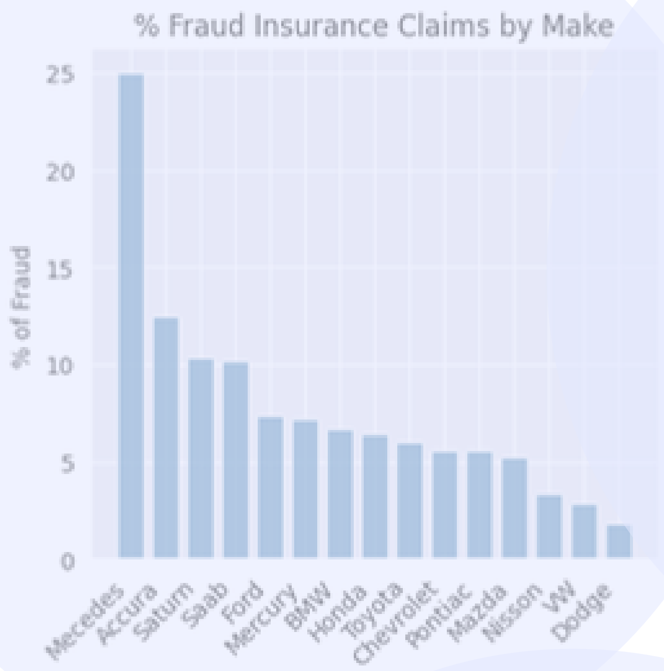


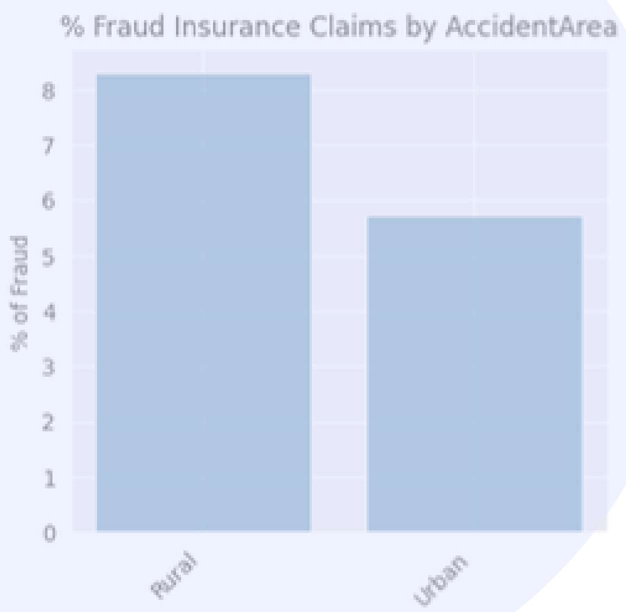
% Fraud Insurance Claims by PoliceReportFiled
% of Fraud
No
Yes

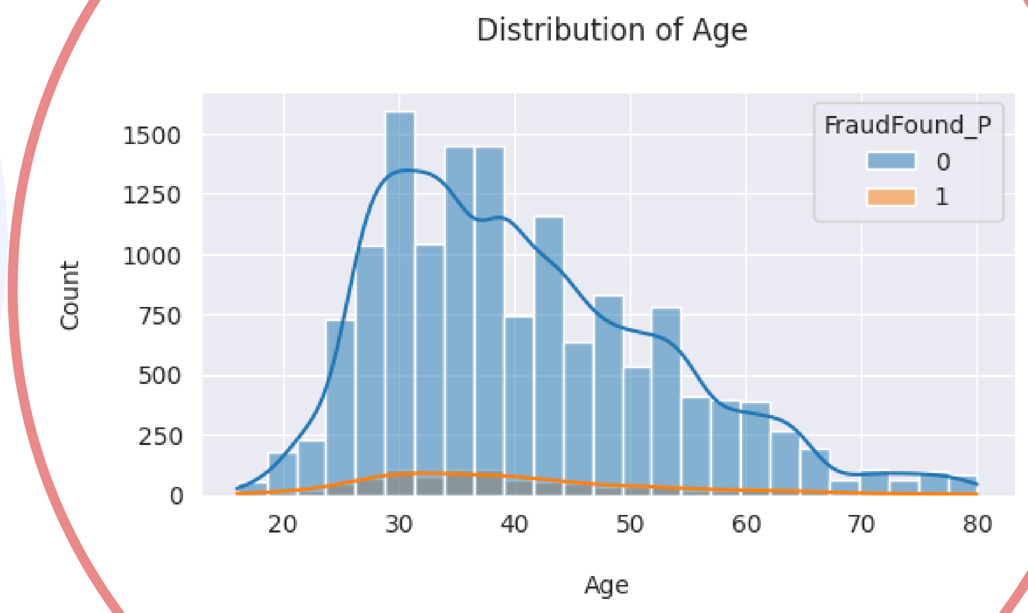


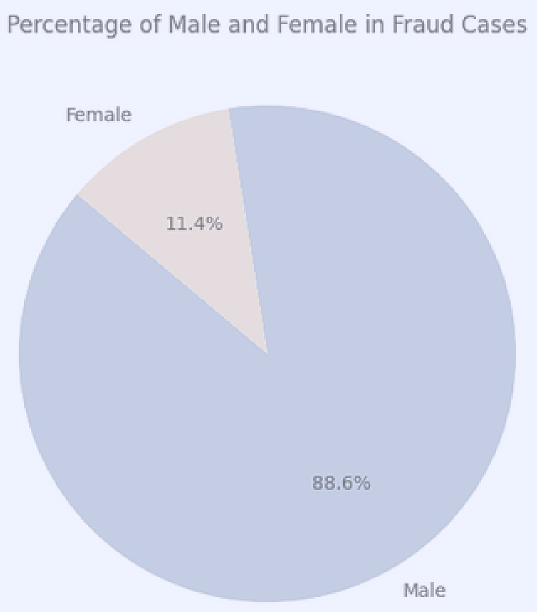
% Fraud Insurance Claims by WitnessPresent
% of Fraud
No
Yes

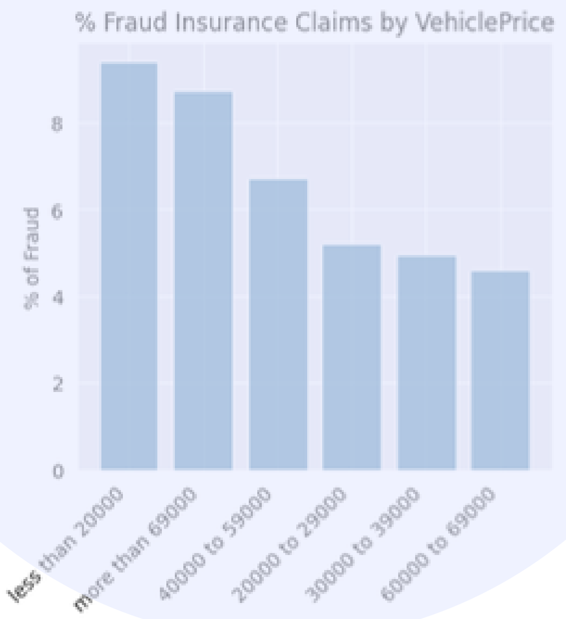
EDA

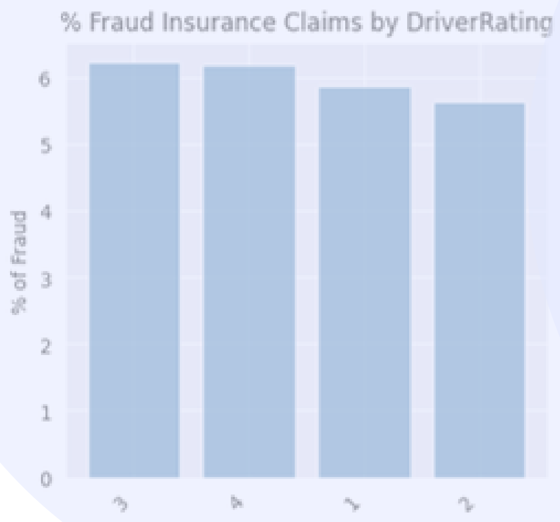

% Fraud Insurance Claims by Make
% of Fraud
Mecedes
Accura
Saturn
Saab
Ford
Mercury
BMW
Honda
Toyota
Chevrolet
Pontiac
Mazda
Nisson
VW
Dodge

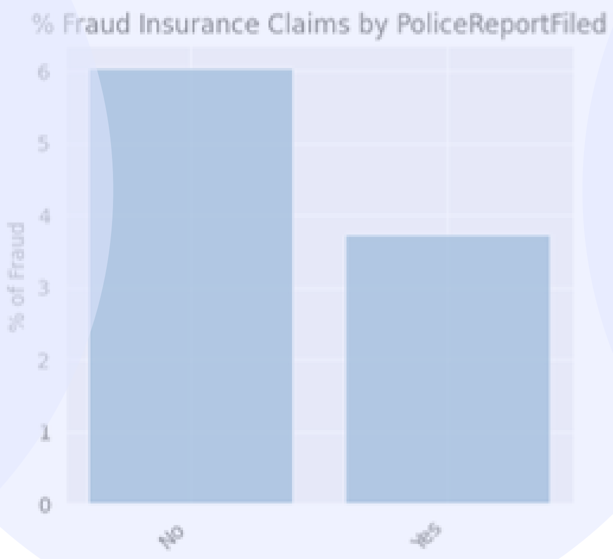

% Fraud Insurance Claims by AccidentArea
% of Fraud
Rural
Urban

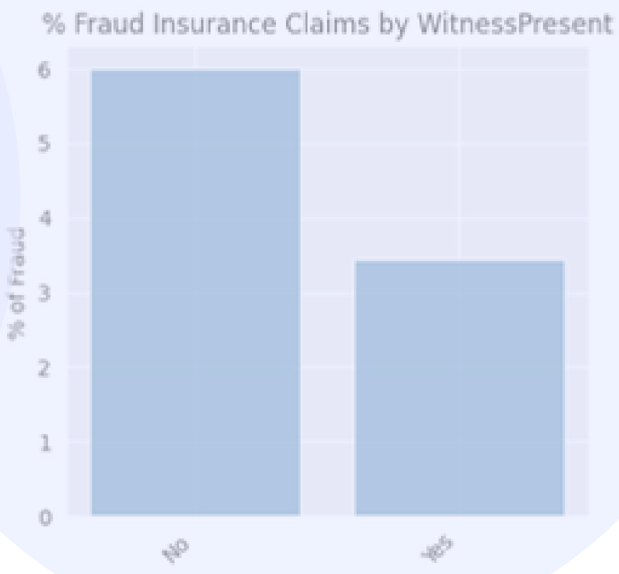

Distribution of Age
FraudFound_P
0
1
Count
Age


Percentage of Male and Female in Fraud Cases
Female
11.4%
88.6%
Male

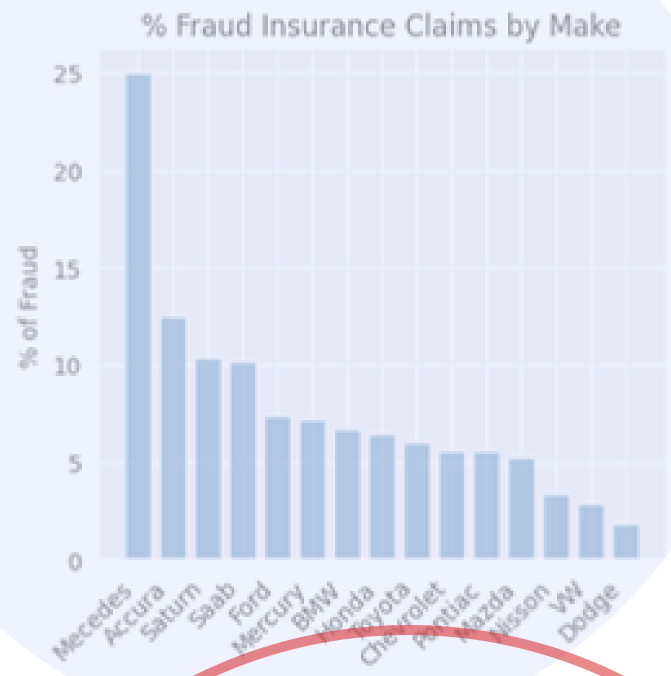

% Fraud Insurance Claims by VehiclePrice
% of Fraud
less than 20000
more than 69000
40000 to 59000
20000 to 29000
30000 to 39000
60000 to 69000

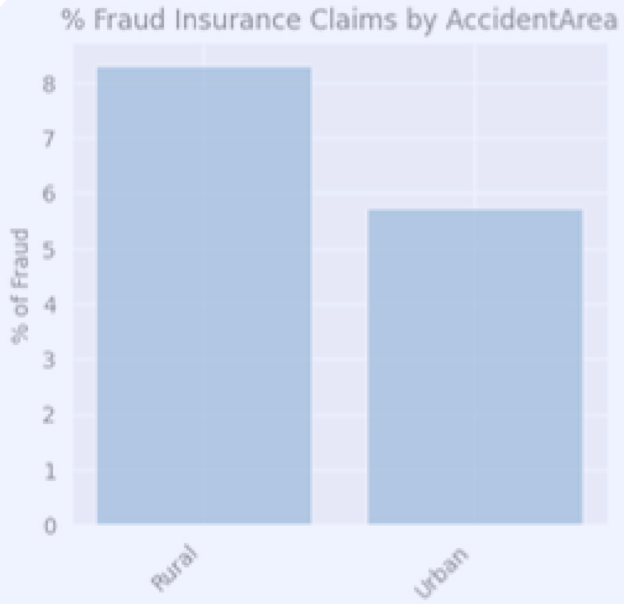

% Fraud Insurance Claims by DriverRating
% of Fraud

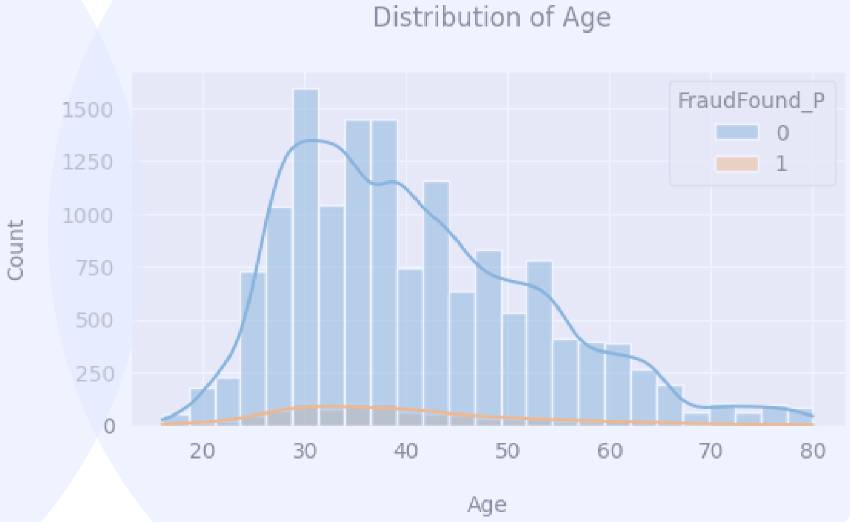

% Fraud Insurance Claims by PoliceReportFiled
No
Yes

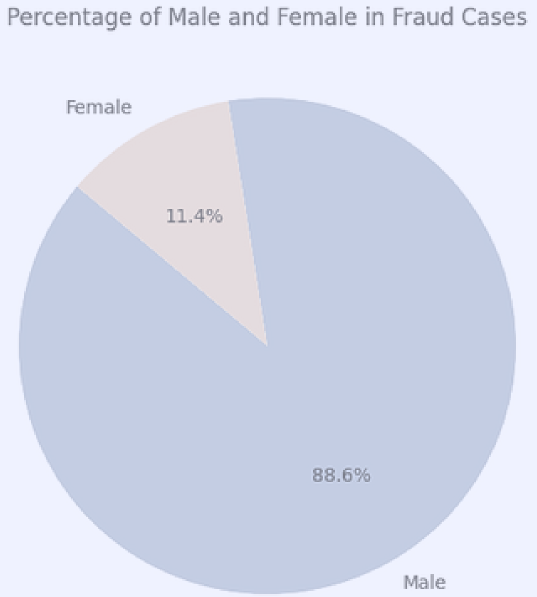

% Fraud Insurance Claims by WitnessPresent
% of Fraud
No
Yes

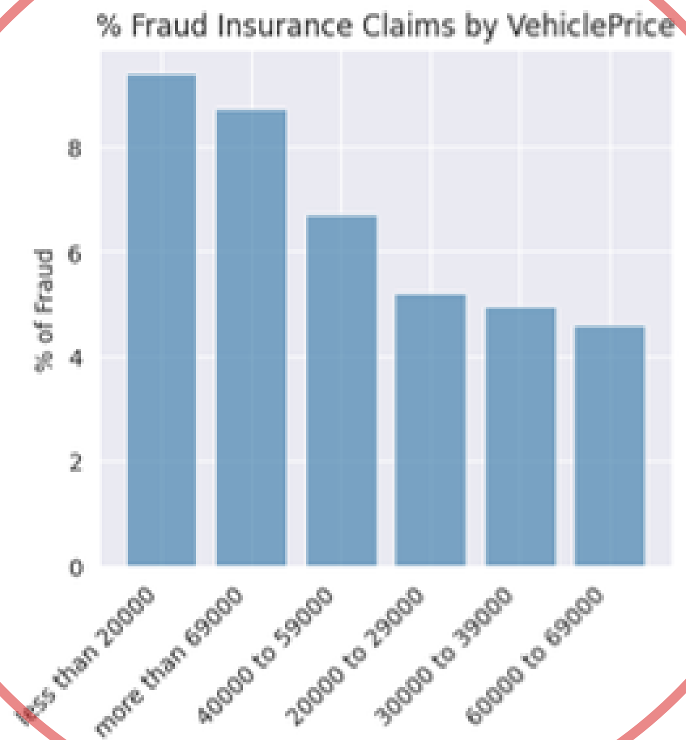
EDA


% Fraud Insurance Claims by Make
% of Fraud
Mecedes
Accura
Saturn
Saab
Ford
Mercury
BMW
Honda
Toyota
Chevrolet
Pontiac
Mazda
Nisson
VW
Dodge

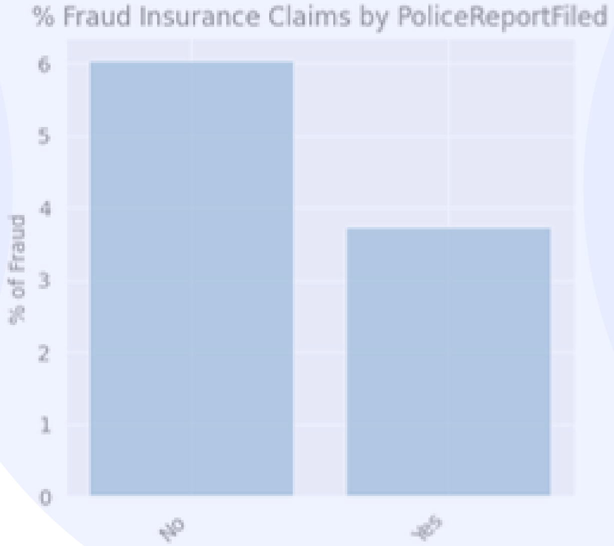

% Fraud Insurance Claims by AccidentArea
% of Fraud
Rural
Urban

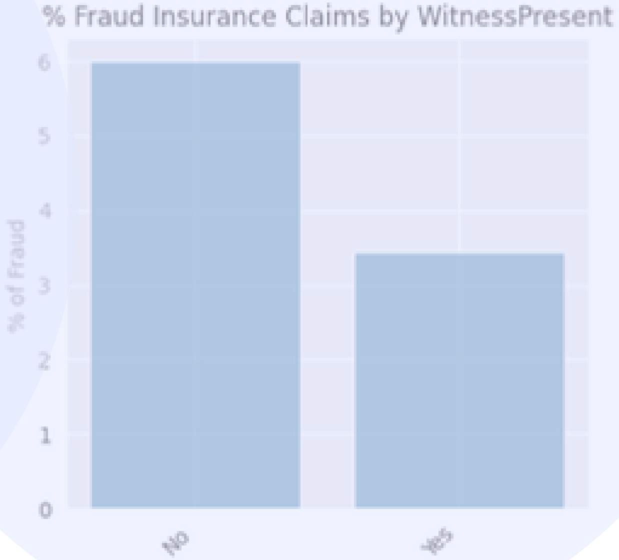

Distribution of Age
Count
Age
FraudFound_P
0
1


Percentage of Male and Female in Fraud Cases
Female
11.4%
88.6%
Male

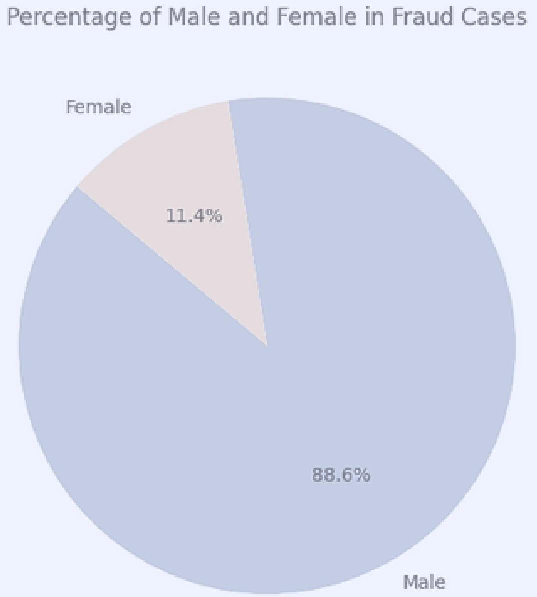

% Fraud Insurance Claims by VehiclePrice
% of Fraud
less than 20000
more than 69000
40000 to 59000
20000 to 29000
30000 to 39000
60000 to 69000


% Fraud Insurance Claims by DriverRating
% of Fraud

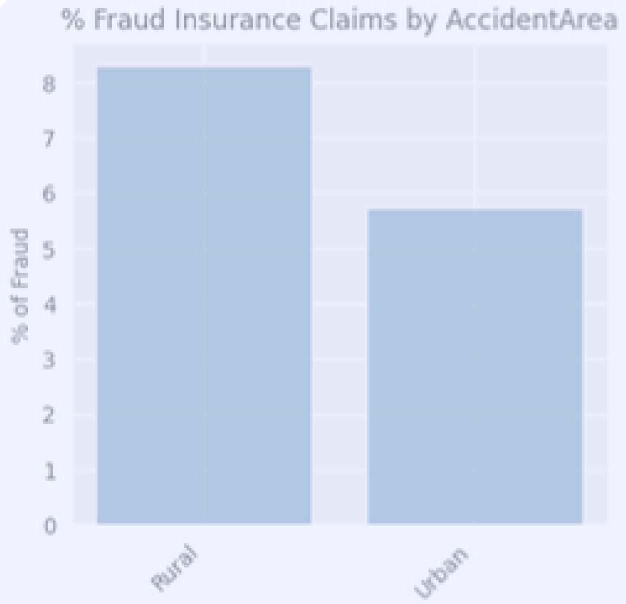

% Fraud Insurance Claims by PoliceReportFiled
% of Fraud
No
Yes

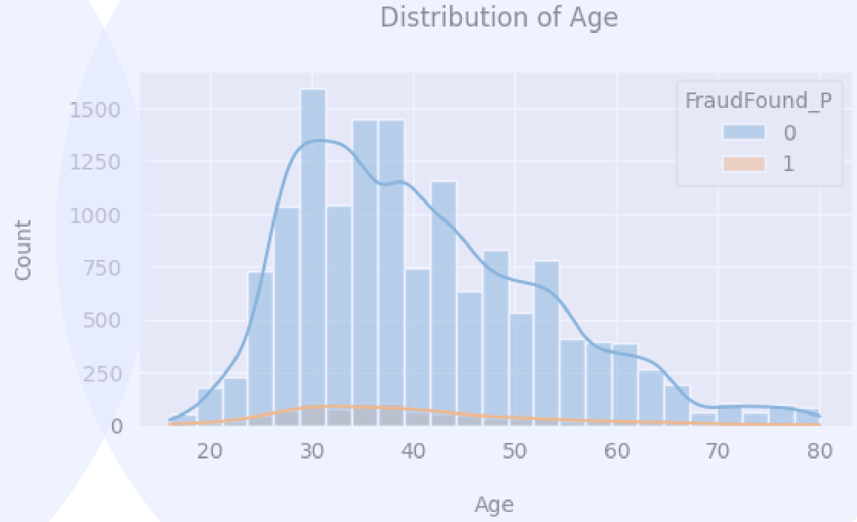

% Fraud Insurance Claims by WitnessPresent
% of Fraud
No
Yes

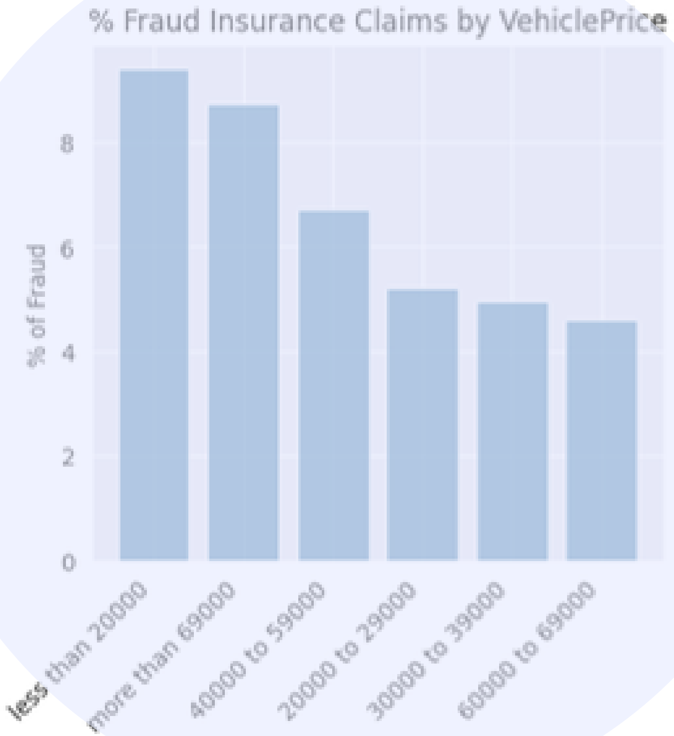
EDA

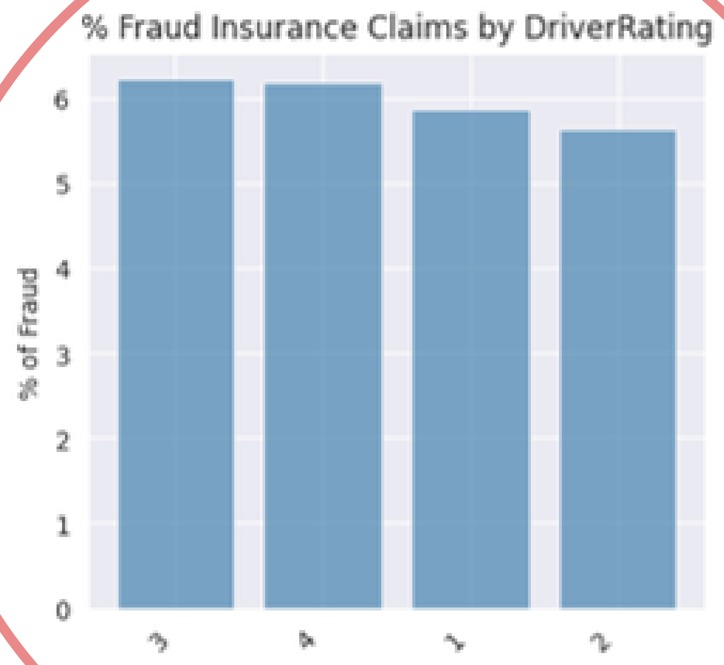

Percentage of Male and Female in Fraud Cases
Female
11.4%
88.6%
Male

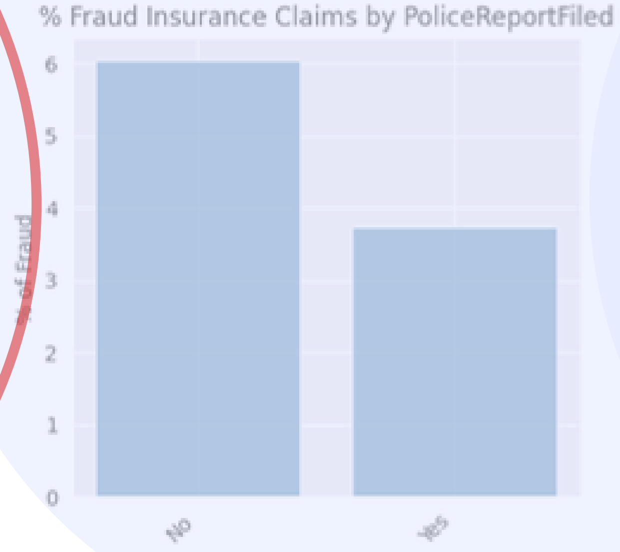

% Fraud Insurance Claims by Make
% of Fraud
Mecedes
Accura
Saturn
Saab
Ford
Mercury
BMW
Honda
Toyota
Chevrolet
Pontiac
Mazda
Nisson
VW
Dodge

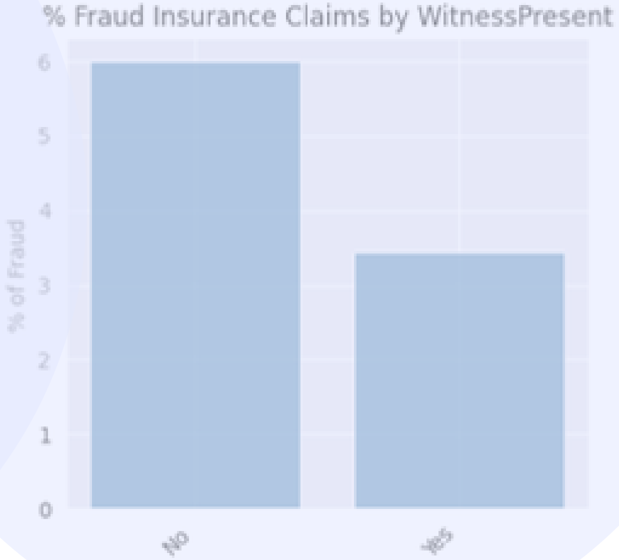

% Fraud Insurance Claims by AccidentArea
% of Fraud
Rural
Urban


Distribution of Age
FraudFound_P
0
1
Count
Age


% Fraud Insurance Claims by VehiclePrice
% of Fraud
less than 20000
more than 69000
40000 to 59000
20000 to 29000
30000 to 39000
60000 to 69000

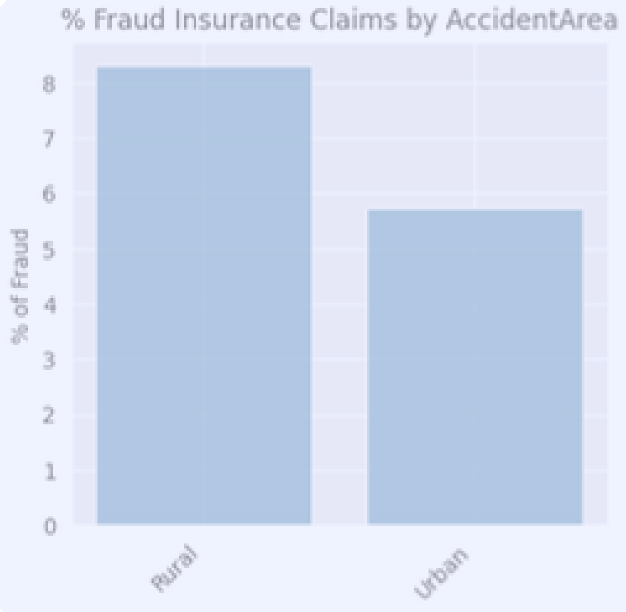

% Fraud Insurance Claims by DriverRating
% of Fraud
3
4
1
2

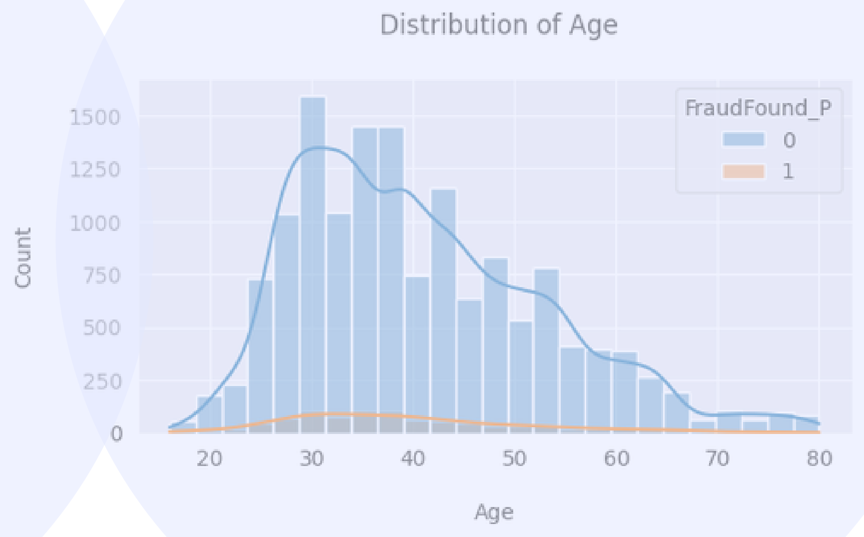

% Fraud Insurance Claims by PoliceReportFiled
No
Yes

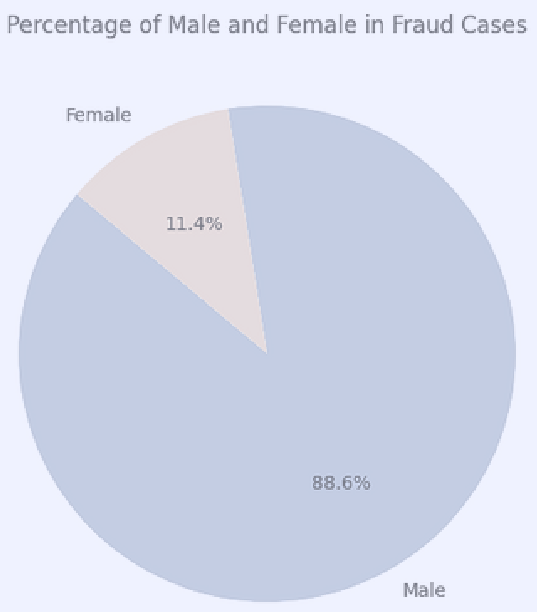

% Fraud Insurance Claims by WitnessPresent
No
Yes

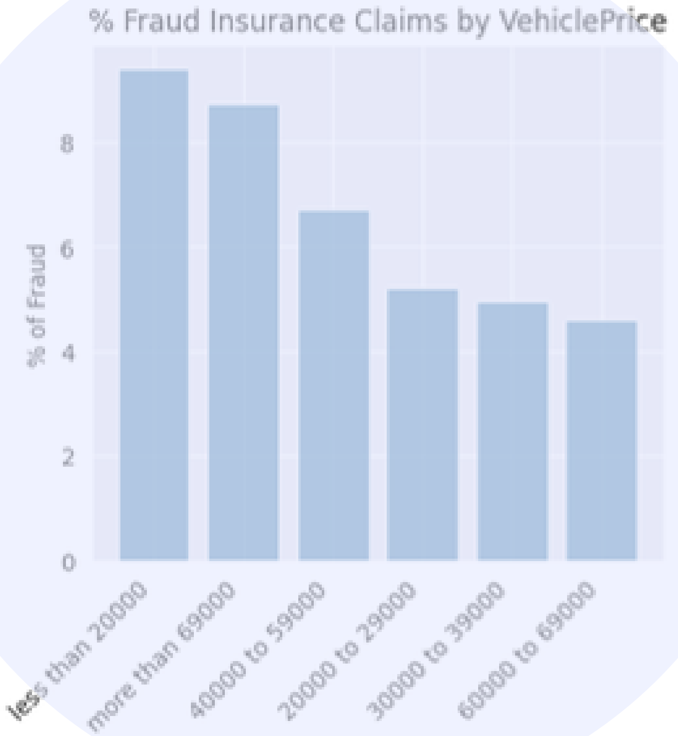
EDA

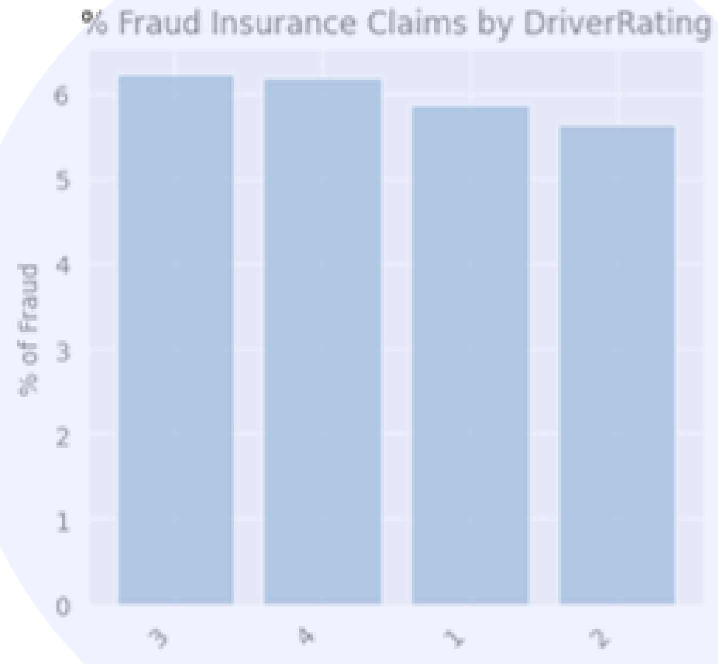

% Fraud Insurance Claims by Make
% of Fraud
Mercedes
Accura
Saturn
Saab
Ford
Mercury
BMW
Honda
Toyota
Chevrolet
Pontiac
Mazda
Nisson
VW
Dodge

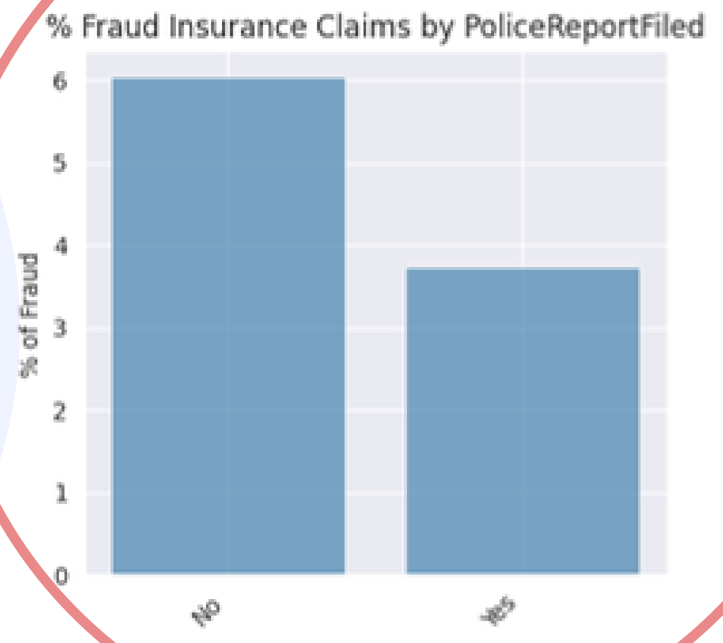

% Fraud Insurance Claims by AccidentArea
% of Fraud
Rural
Urban


Distribution of Age
Count
Age
FraudFound_P
0
1


Percentage of Male and Female in Fraud Cases
Female
11.4%
88.6%
Male

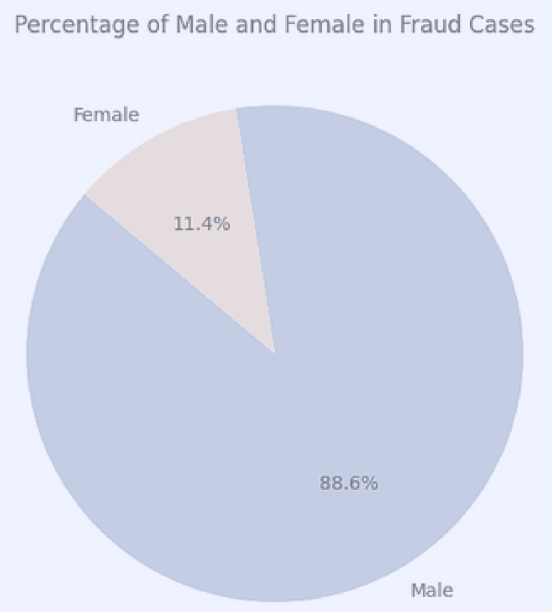

% Fraud Insurance Claims by VehiclePrice
% of Fraud
less than 20000
more than 69000
40000 to 59000
20000 to 29000
30000 to 39000
60000 to 69000


% Fraud Insurance Claims by DriverRating
% of Fraud

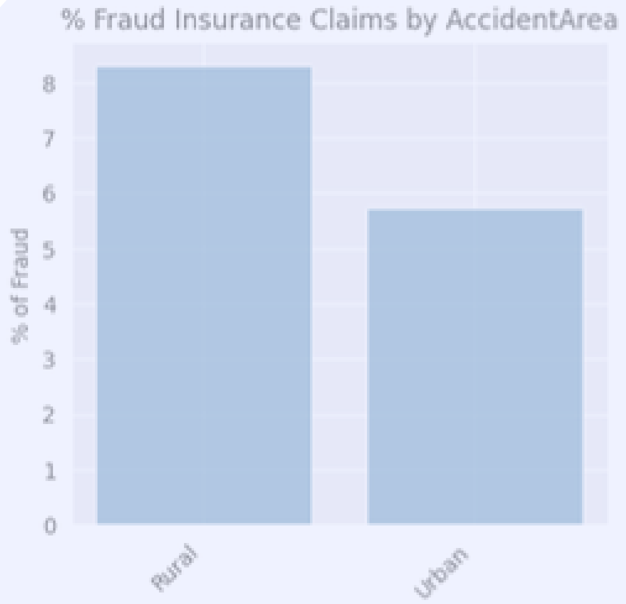

% Fraud Insurance Claims by PoliceReportFiled
% of Fraud
No
Yes

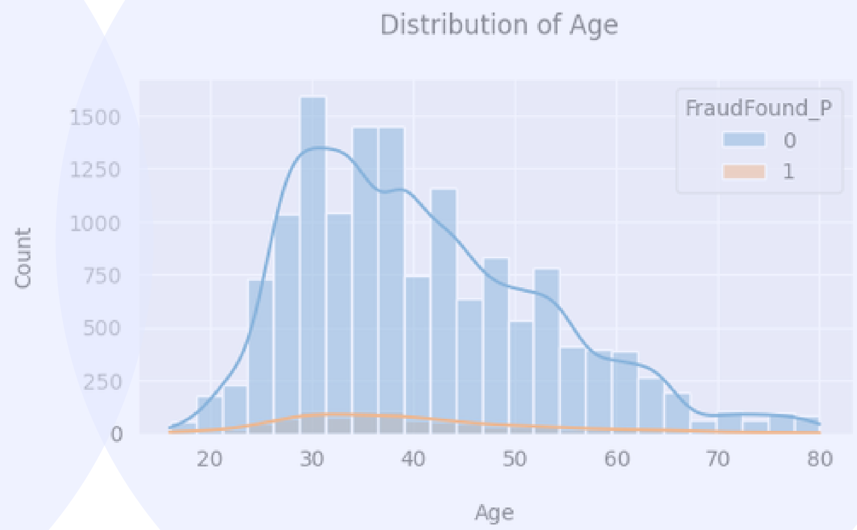

% Fraud Insurance Claims by WitnessPresent
% of Fraud
No
Yes

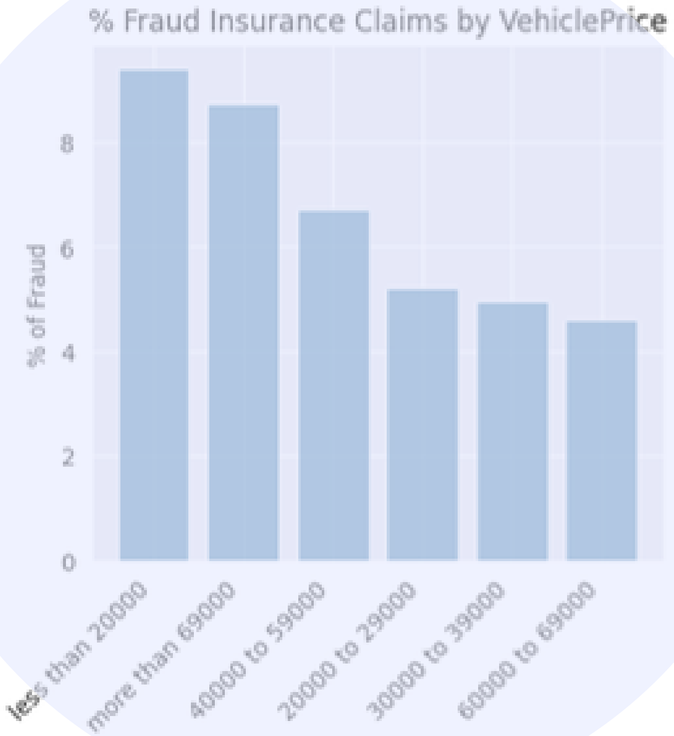
EDA

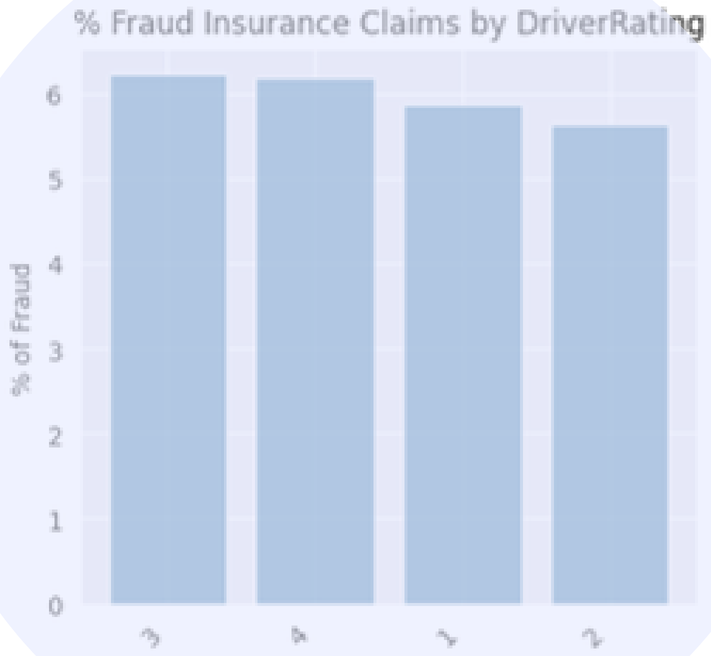

Percentage of Male and Female in Fraud Cases
Female
11.4%
88.6%
Male

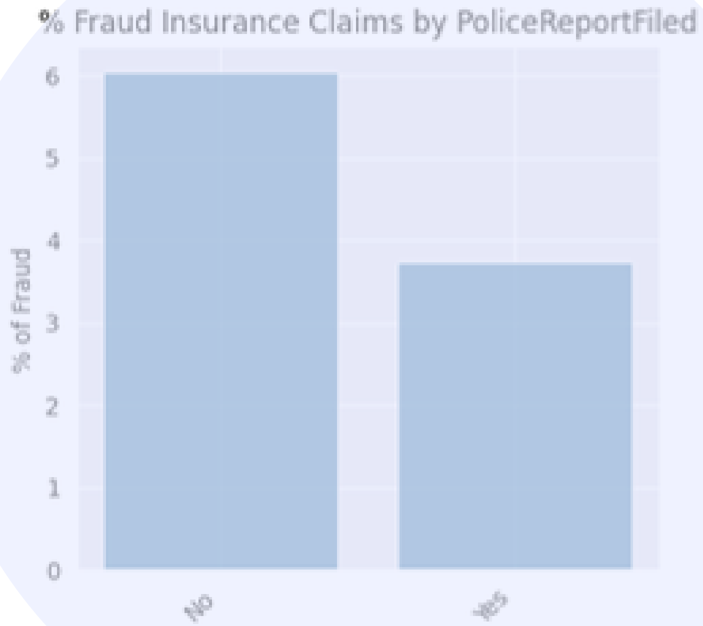

% Fraud Insurance Claims by Make
% of Fraud
Mecedes
Accura
Saturn
Saab
Ford
Mercury
BMW
Honda
Toyota
Chevrolet
Pontiac
Mazda
Nisson
VW
Dodge


% Fraud Insurance Claims by AccidentArea
% of Fraud
Rural
Urban


Distribution of Age
FraudFound_P
0
1
Count
Age


% Fraud Insurance Claims by VehiclePrice
% of Fraud
less than 20000
more than 69000
40000 to 59000
20000 to 29000
30000 to 39000
60000 to 69000


% Fraud Insurance Claims by DriverRating
% of Fraud
3
4
1
2

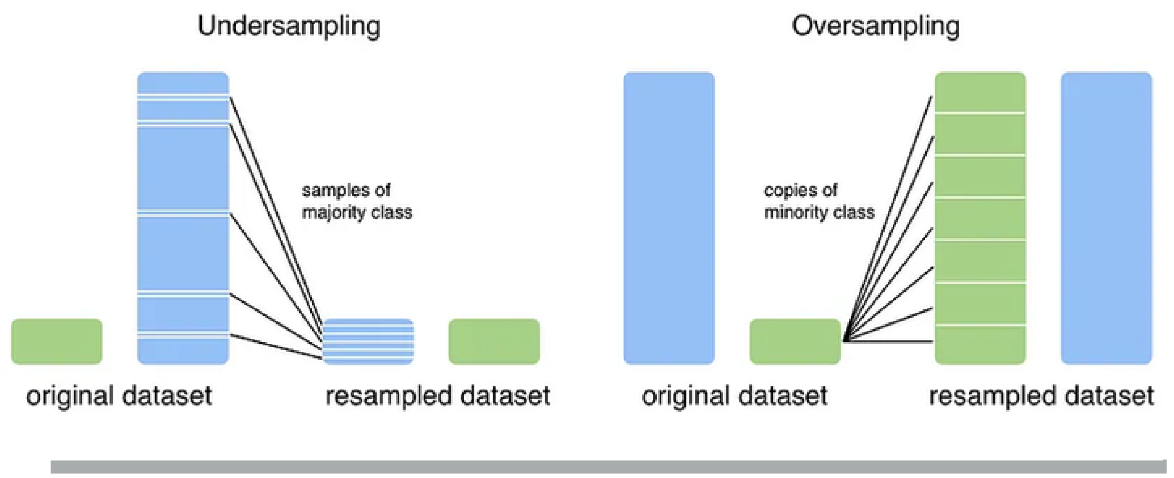

% Fraud Insurance Claims by PoliceReportFiled
% of Fraud
No
Yes


% Fraud Insurance Claims by WitnessPresent
% of Fraud
No
Yes

RESAMPLING THE DATA

General Resampling Methods

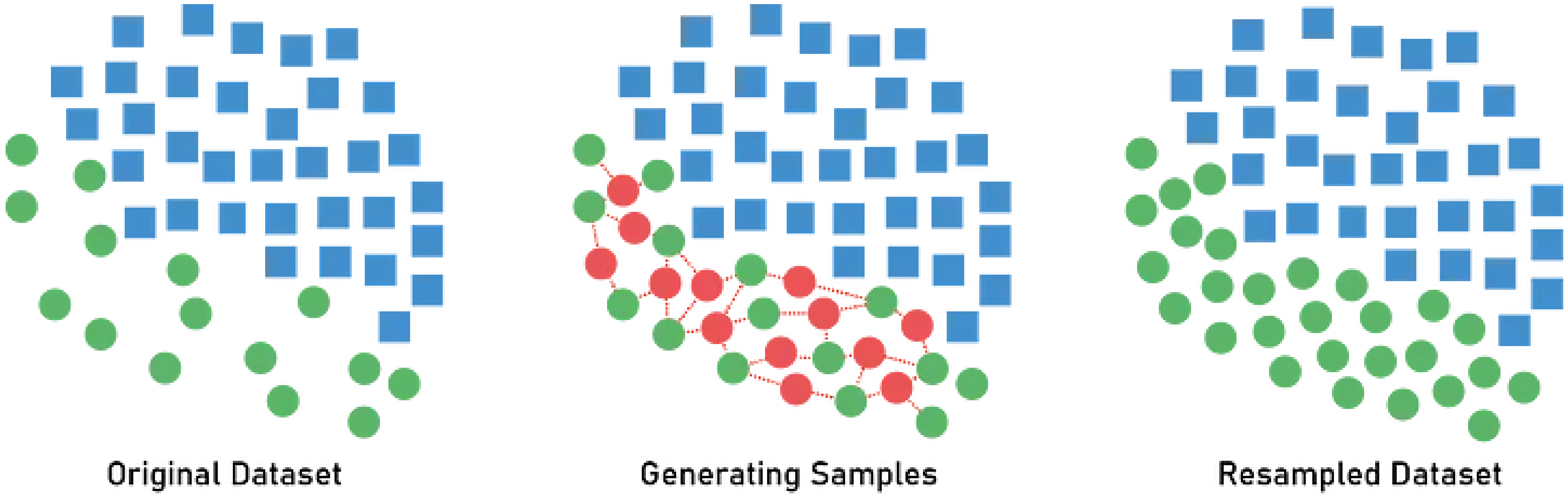
- 01 - OVER SAMPLING
- 02 - UNDER SAMPLING



OVERSAMPLING

SMOTE

Synthetic Minority Oversampling Technique



Adding samples to minority class (fraud cases)

OVERSAMPLING

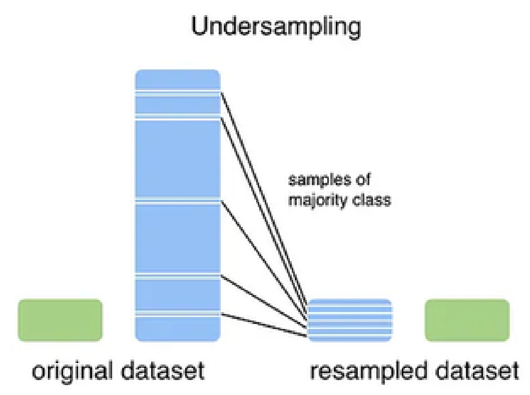
ADVANTAGES

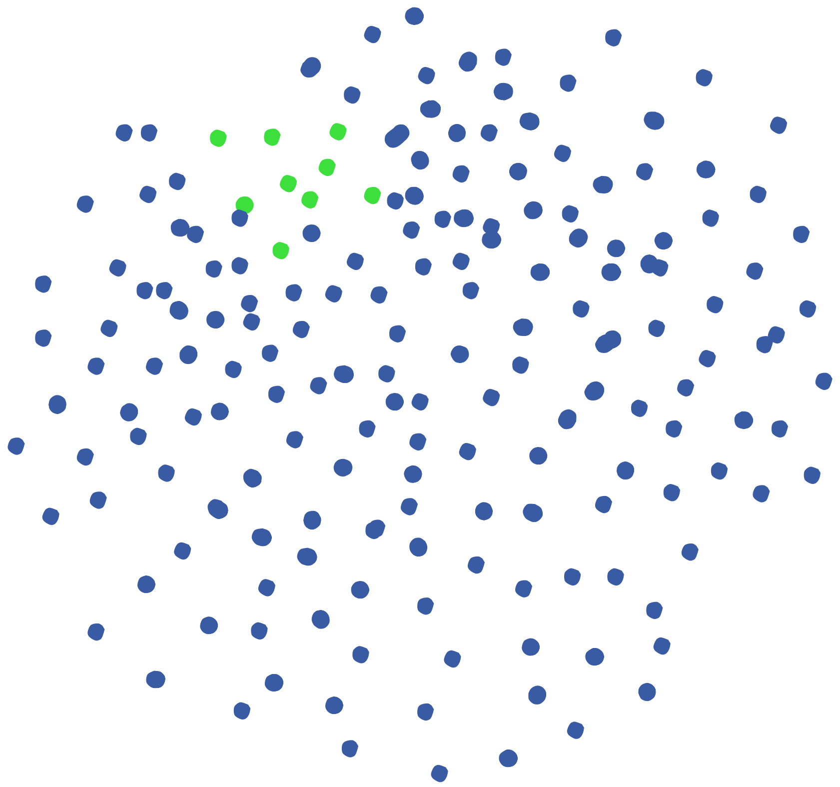
- Can improve the accuracy of classification models on the minority class.
- Can reduce the overfitting of classification models.
- Relatively simple to implement and can be used with a variety of classification algorithms

LIMITATIONS

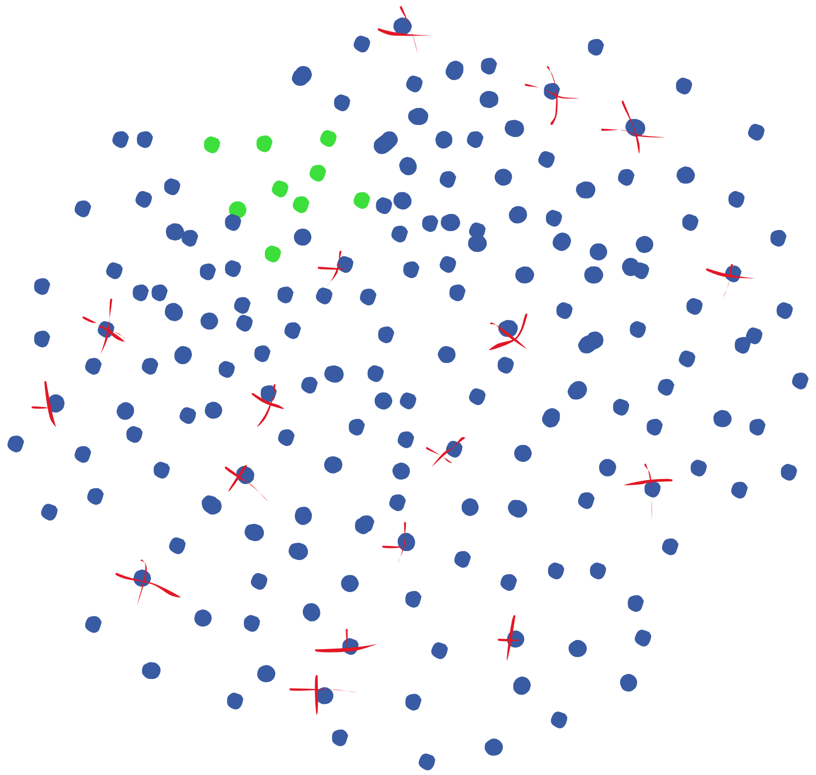
- Can introduce bias into the dataset.
- Can be computationally expensive for large datasets.
- May not be effective for all types of imbalanced datasets.

UNDERSAMPLING

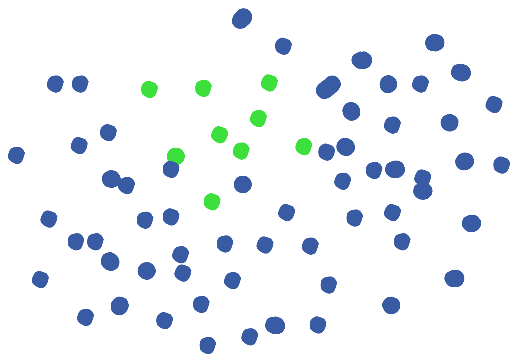




Original dataset



Reducing samples



Resampled dataset

Removing samples from majority class (non- fraud cases)

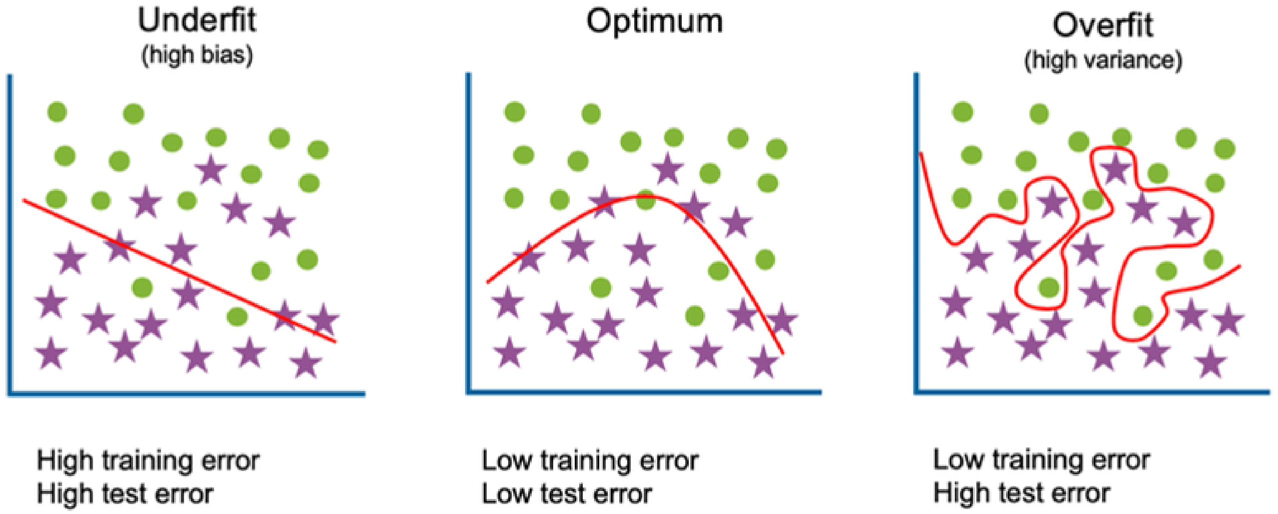
UNDERSAMPLING

ADVANTAGES

- Can significantly decrease the amount of data, which in turn speeds up the training process of machine learning models.
- Can improve the performance of the model on minority class data points by balancing the class distribution.
- Relatively simple to implement and can be used with a variety of classification algorithms.

LIMITATIONS

- Can increase risk of losing important or representative information.
- Not suitable for very small datasets.
- Risk of increased variance and overfitting (because of fewer datapoints).



MODELS USED

01 - LOGISTIC REGRESSION

02 - DECISION TREE

Single Learning Models

03 - RANDOM FOREST

04 - XGBOOST

Ensemble Learning Models

Neural Network Models

05 - ARTIFICIAL NN

LOGISTIC REGRESSION

- Traditional regression formula inside the ***logistic*** function

$$\hat{y} = \boxed{b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a}$$

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \cdot X$$

$$P = \frac{e^{-0.15\times\text{Rural}+0.35\times\text{Collision}+0.6\times\text{All Perils}+\alpha}}{1+e^{-0.15\times\text{Rural}+0.35\times\text{Collision}+0.6\times\text{All Perils}+\alpha}}$$

- Interpretability: *e.g.* log-odds of fraud decrease by 0.15 when the claim is in a rural area.

LOGISTIC REGRESSION

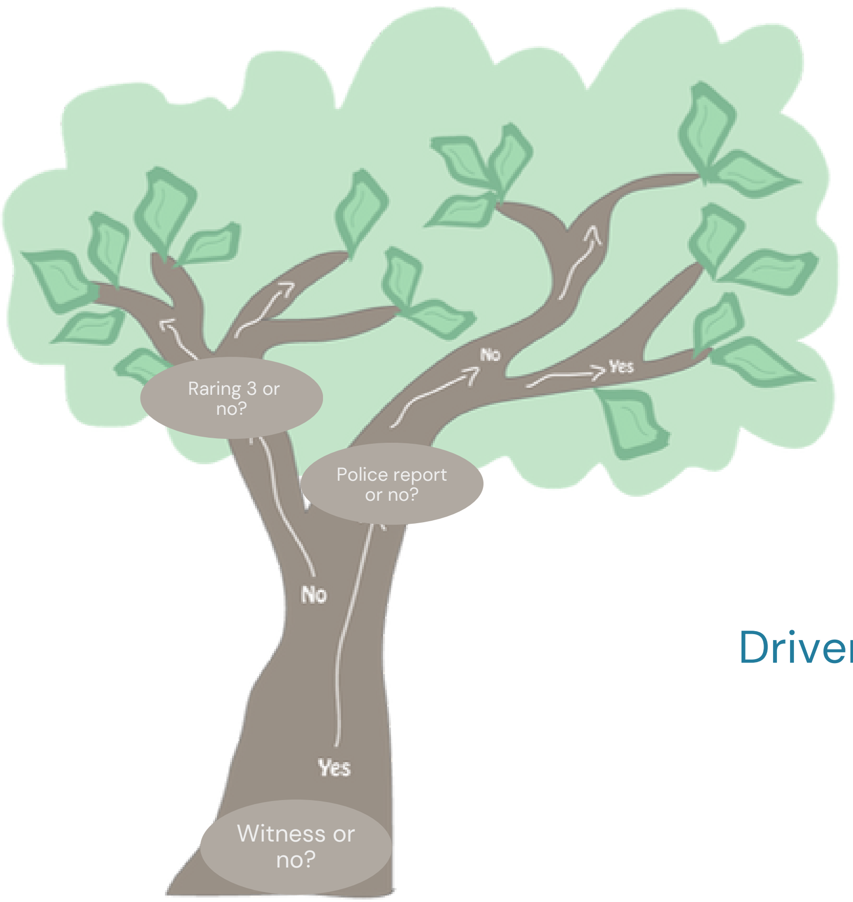
ADVANTAGES

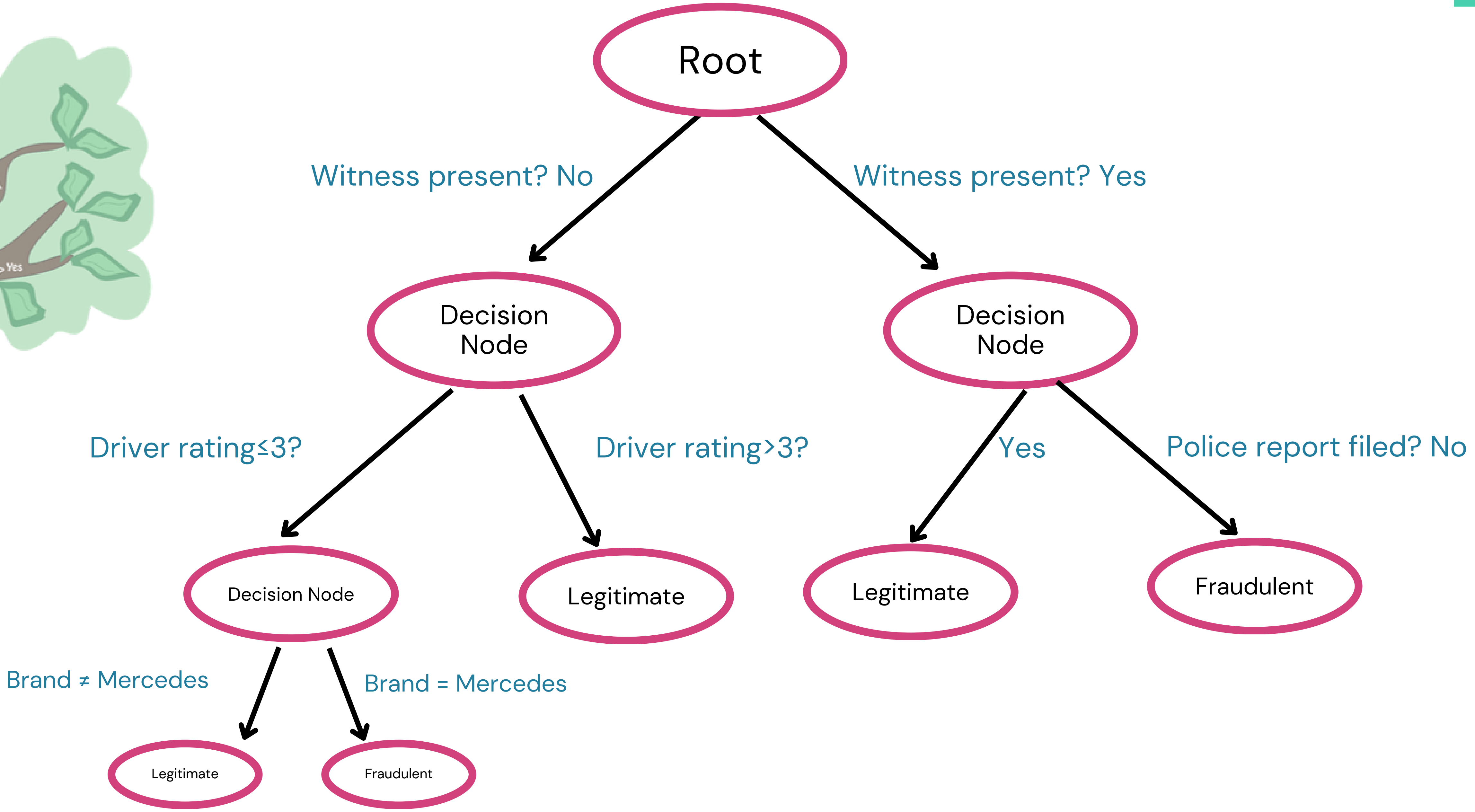
- **Interpretability**: Clear and interpretable results. The coefficients represent the impact of each independent variable on the log-odds of the outcome
- **Probabilistic Predictions**: Models the probability of an event occurring. Valuable when its crucial to understand the likelihood of the outcome
- **Low Variance**: Less prone to overfitting.

LIMITATIONS

- **Assumption of Linearity**: Assumes a linear relationship between independent variables and the log-odds, may fail to capture complex non-linear patterns.
- **Sensitivity to Outliers:** Extreme values can disproportionately impact the model's coefficients and predictions.

DECISION TREE

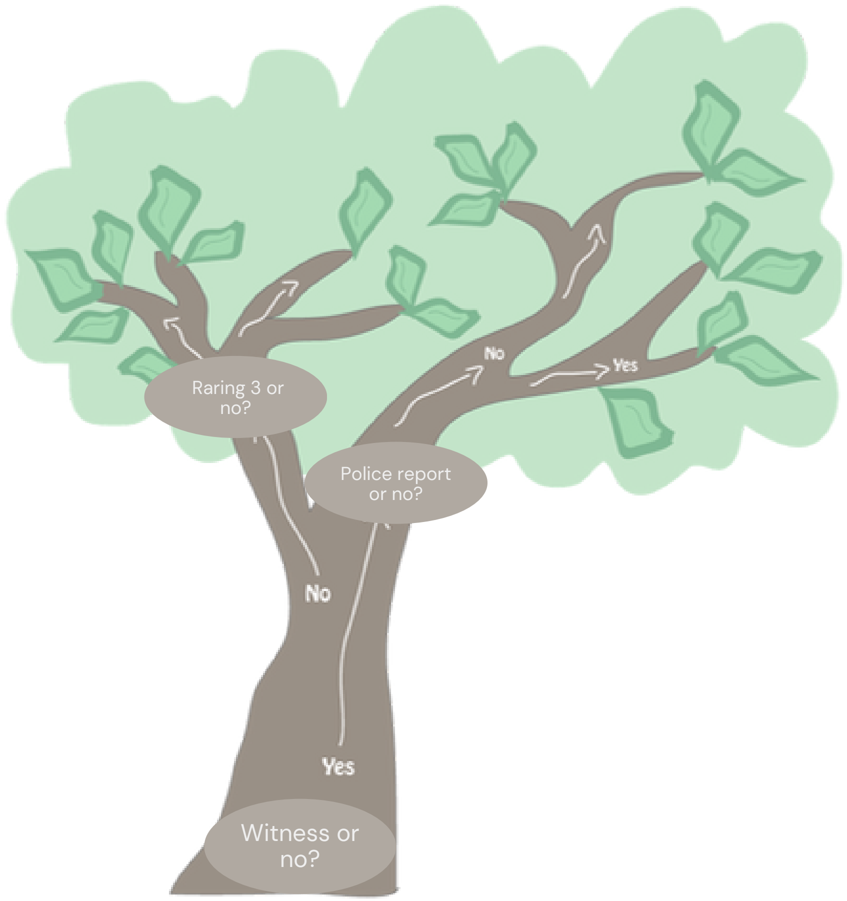




DECISION TREE

- Interpretability and visualisation
- No need for data normalisation

- Prone to overfitting, especially with complex datasets.
- Instability



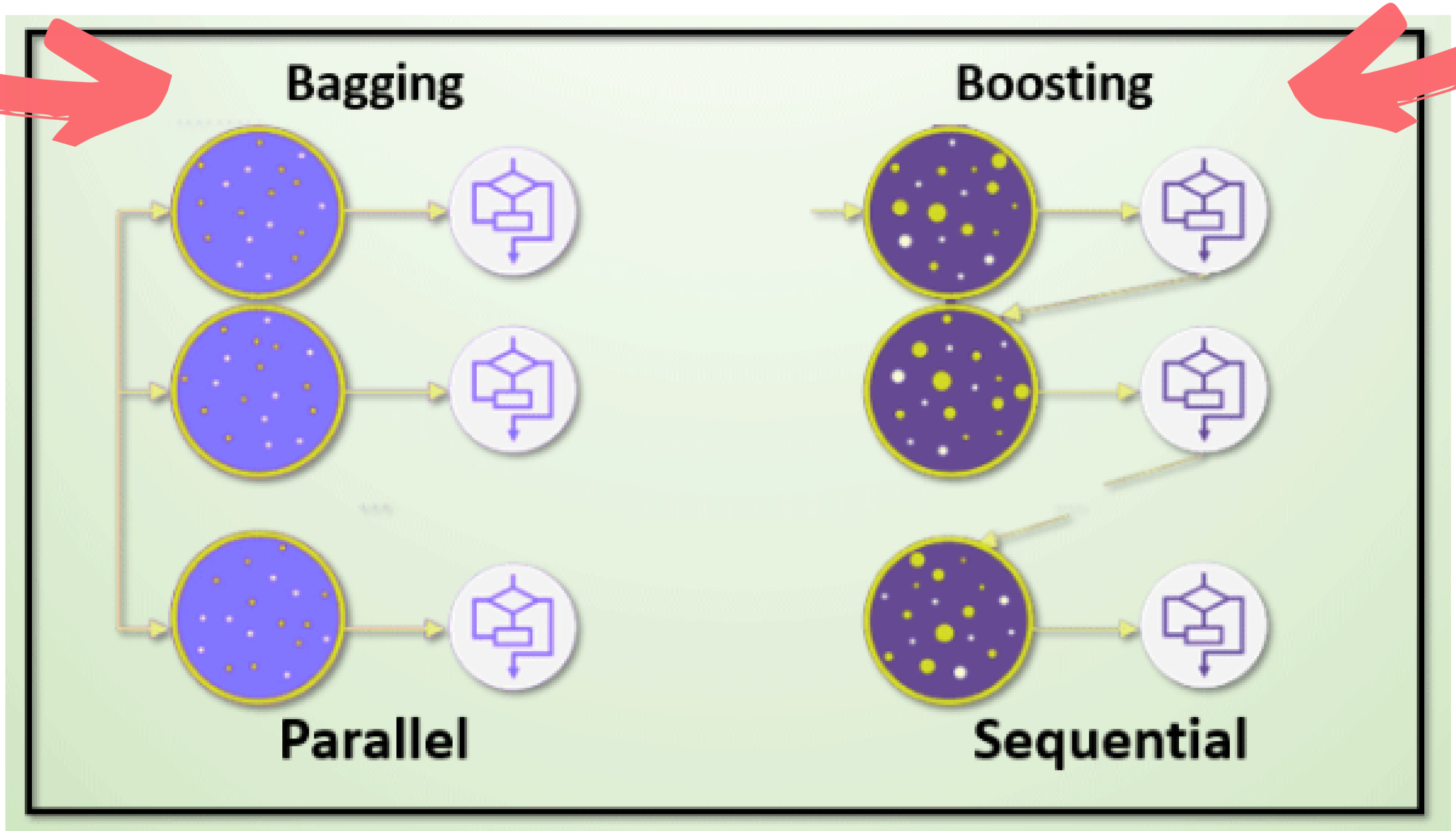
ENSEMBLE METHODS

BAGGING

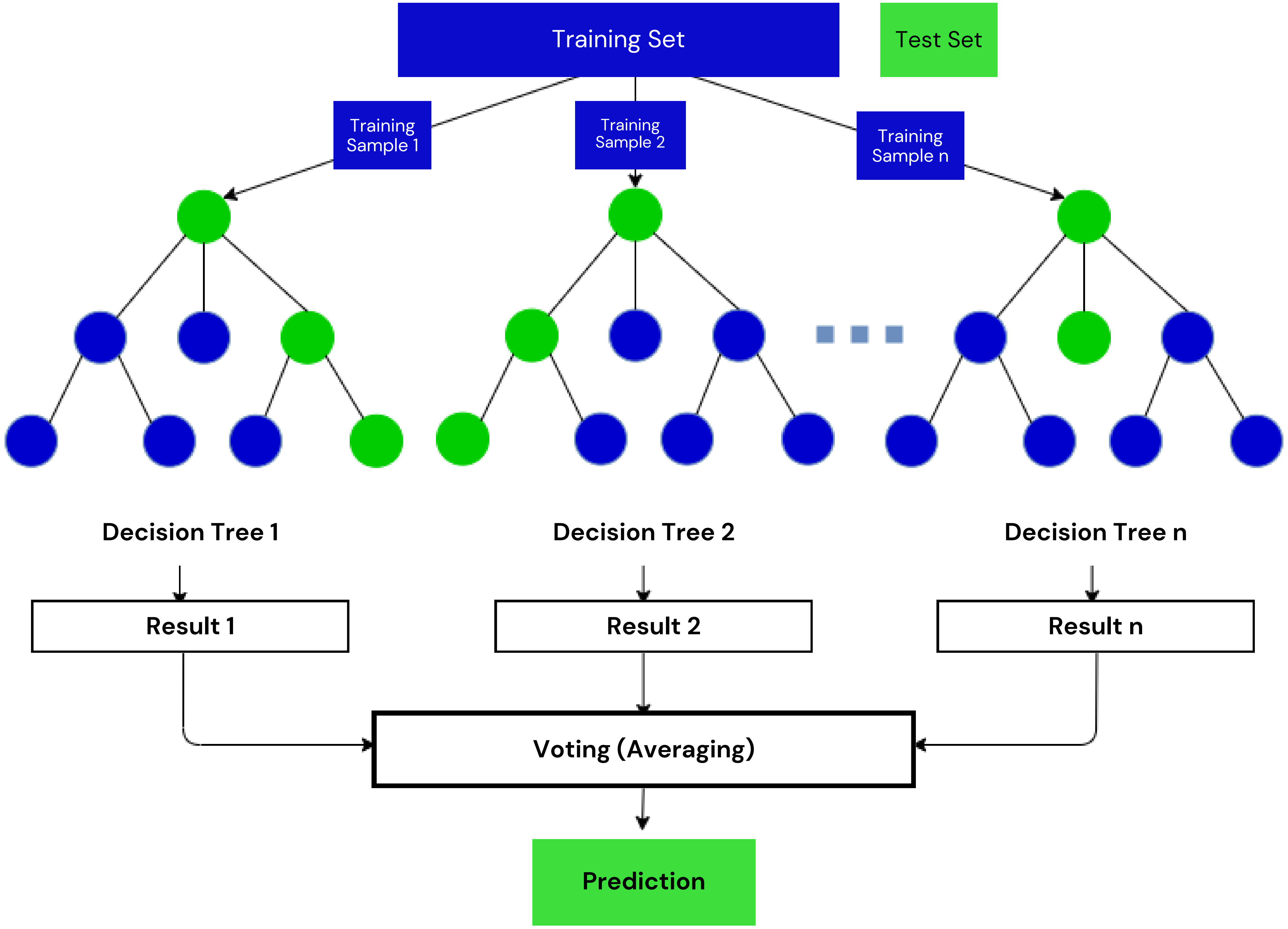
RANDOM FOREST

BOOSTING

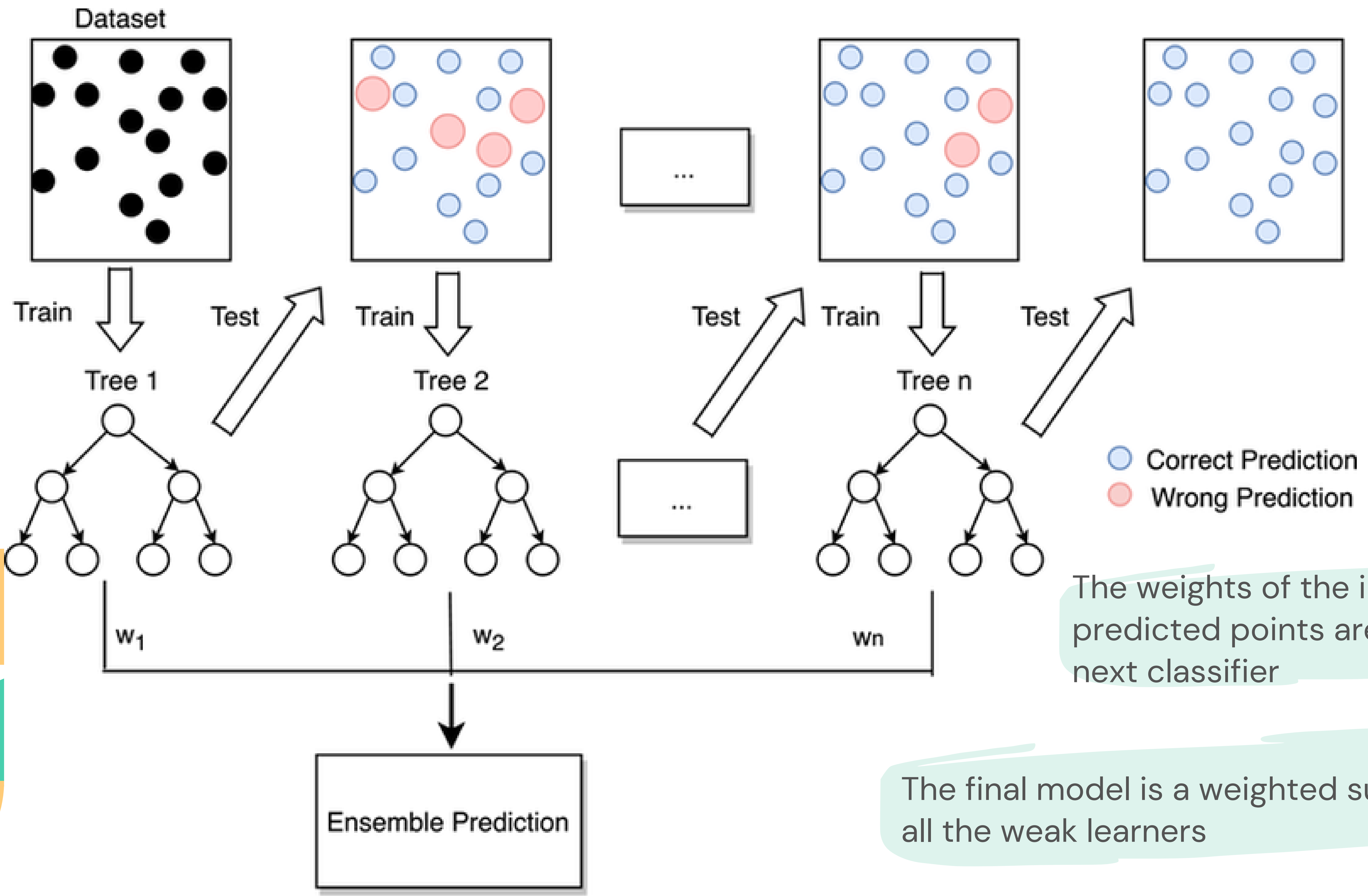
XGBOOST



RANDOM FOREST: AN ENSEMBLE OF DECISION TREES



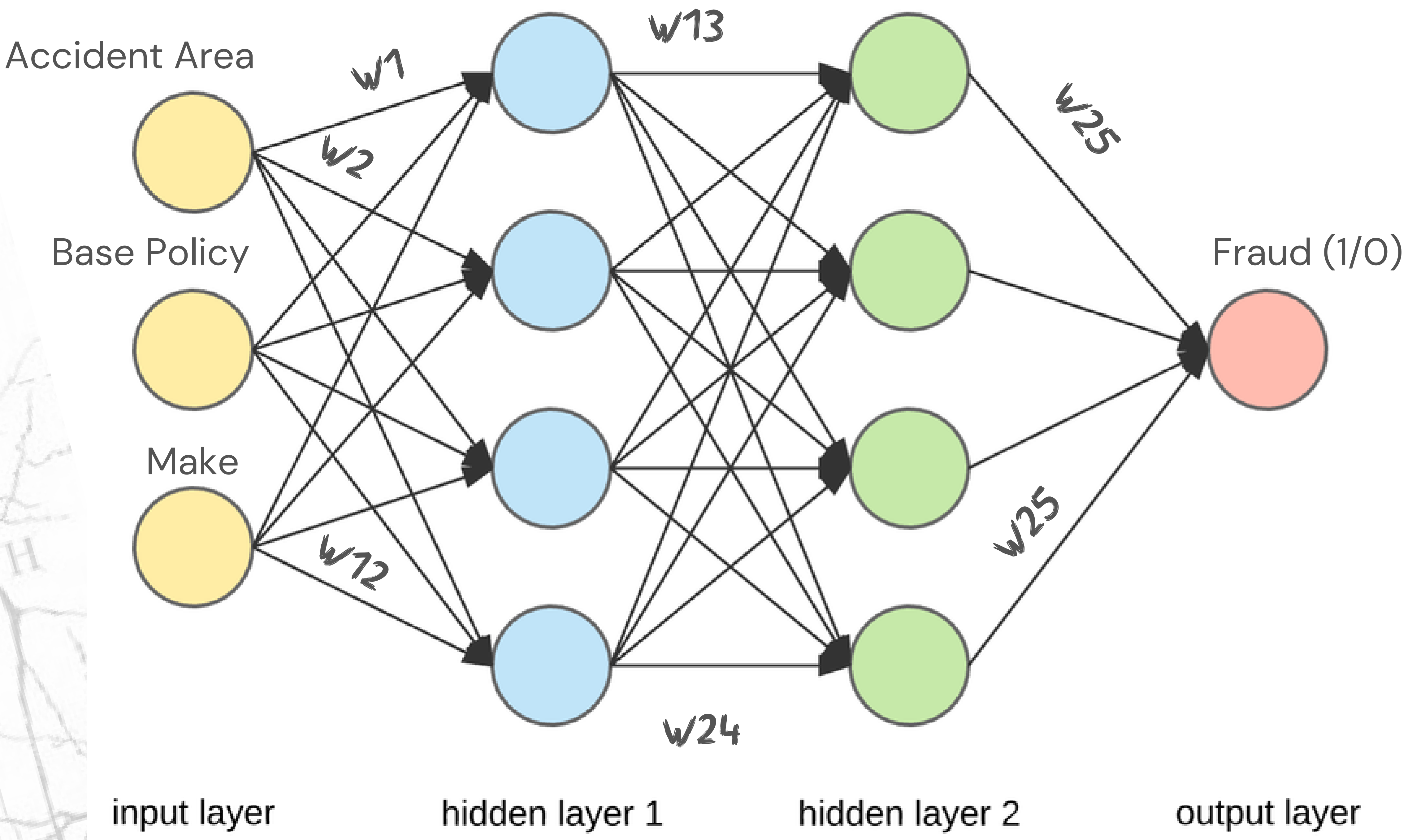
XGBOOST: EXTREME GRADIENT BOOSTING



The weights of the incorrectly predicted points are increased in the next classifier

The final model is a weighted sum of all the weak learners

ARTIFICIAL NEURAL NETWORKS



Forward propagation
Input training data and propagate it forward

Error Calculation: Assess the difference between the predicted output and the actual target values

Learn by adjusting the weights via **backpropagation**.

EVALUATION METRICS

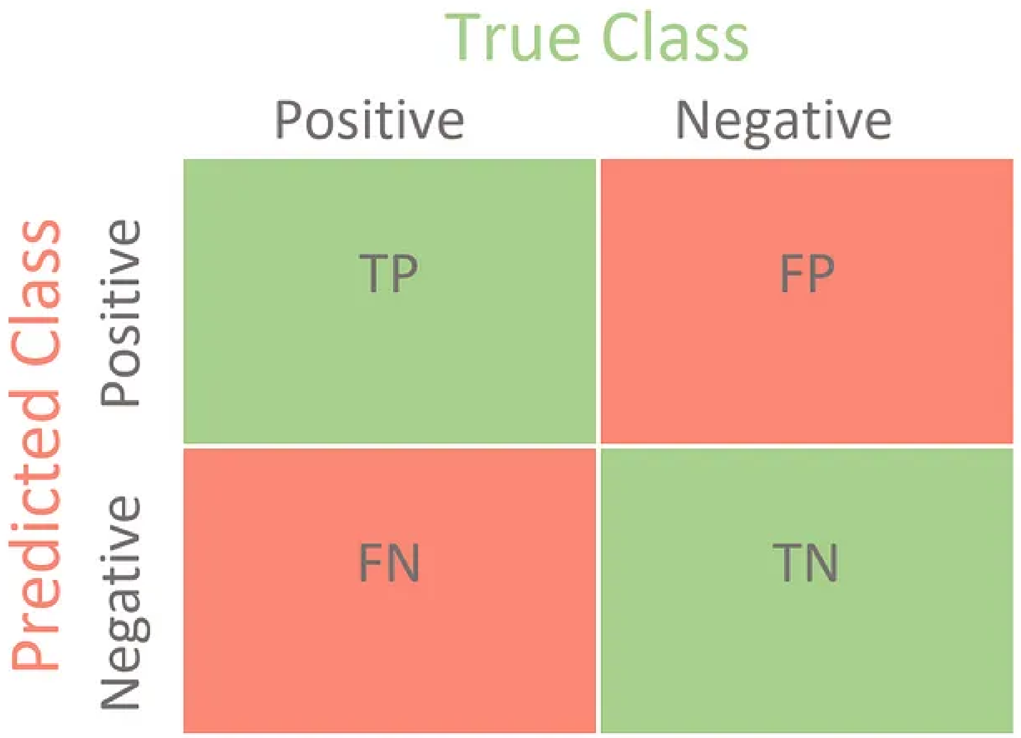
CONFUSION MATRIX

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

CONFUSION MATRIX

Visualises the
actual values in each class
vs.
predicted values by the machine learning model

```
Random Forest Confusion Matrix:
 [[2899    0]
 [ 182    3]]
```



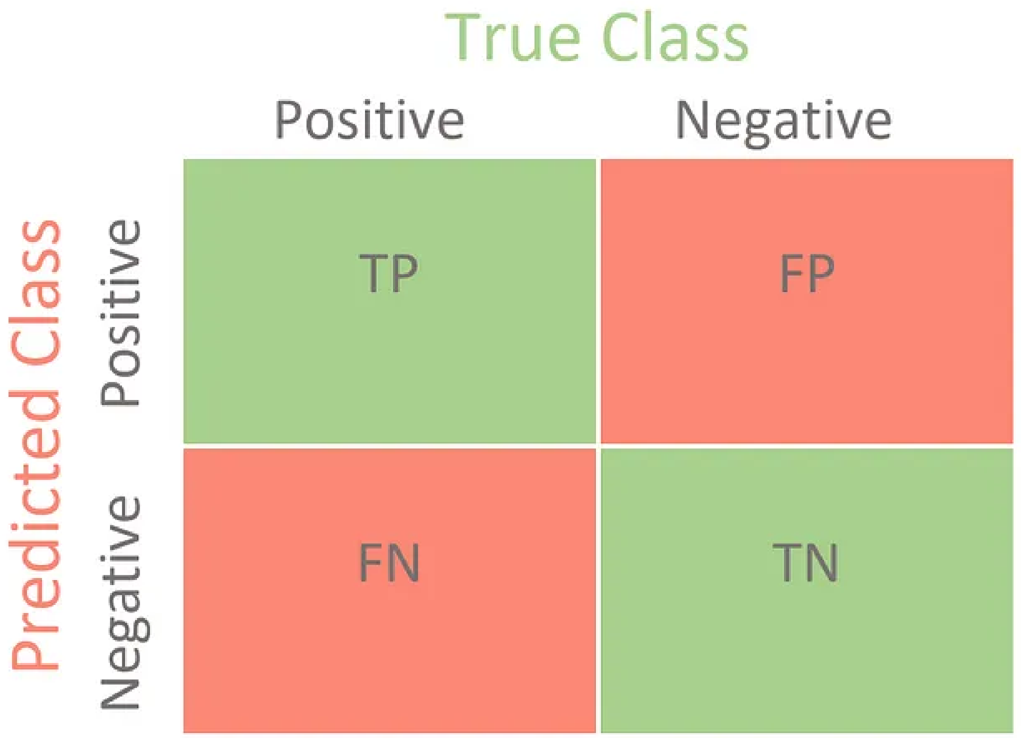
CONFUSION MATRIX

```
Random Forest Confusion Matrix:
 [[2899    0]
 [ 182    3]]
```

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

CONFUSION MATRIX

```
Random Forest Confusion Matrix:
 [[2899    0]
 [ 182    3]]
```



CONFUSION MATRIX

True Negative (TN) = 2899
False Positive (FP) = 0
False Negative (FN) = 182
True Positive (TP) = 3

```
Random Forest Confusion Matrix:
 [[2899    0]
 [ 182    3]]
```

		True Class: Positive	True Class: Negative
Predicted Class	Positive	TP	FP
Predicted Class	Negative	FN	TN

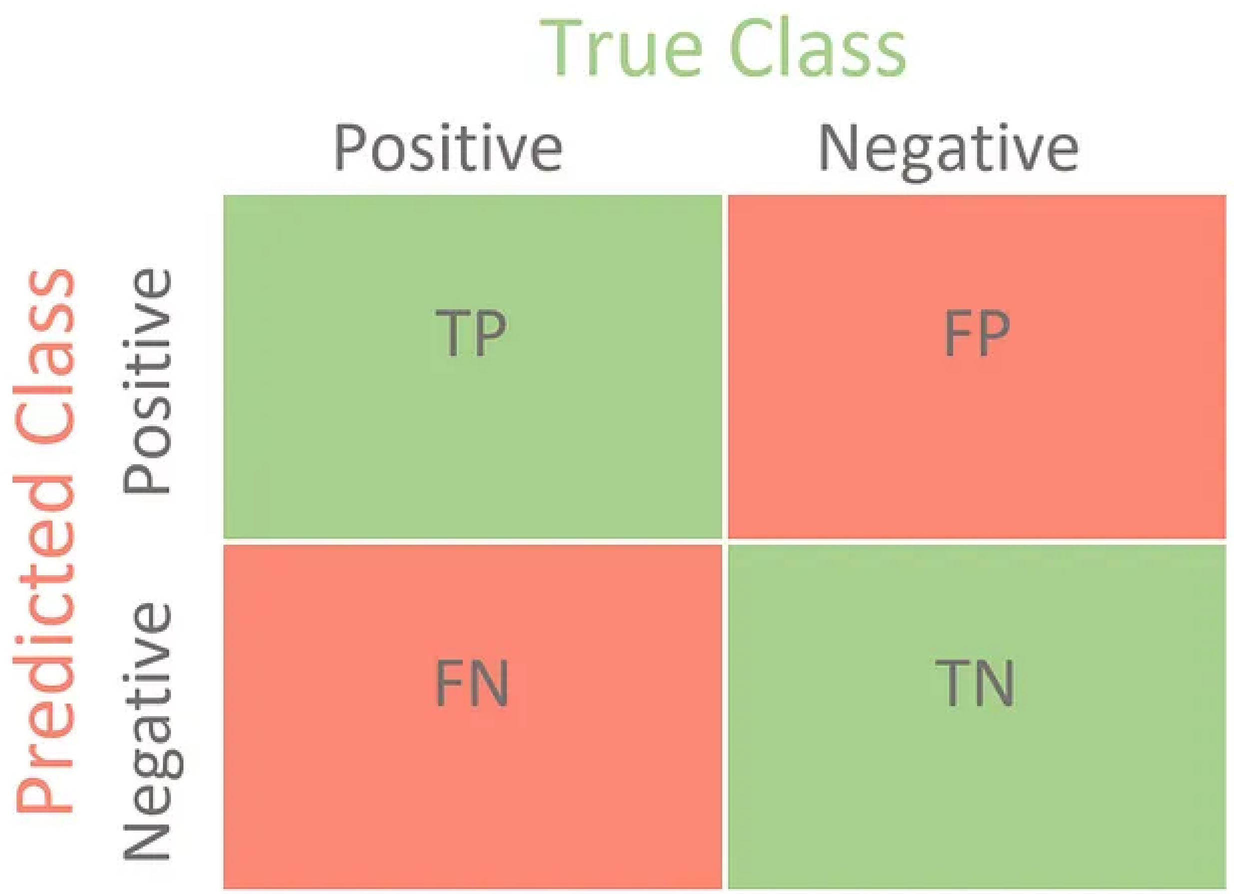
RECALL, PRECISION & F1 SCORE

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision + Recall}}$$

Model	Sampler	Precision	Recall	F1 Score
Decision Tree	RandomUnderSampler	0.127530	0.663158	0.213922
Random Forest	RandomUnderSampler	0.143099	0.891228	0.246602
Logistic Regression	RandomUnderSampler	0.131416	0.817544	0.226433
XGBoost	RandomUnderSampler	0.146563	0.792982	0.247400
Decision Tree	SMOTEENN	0.157985	0.649123	0.254121
Random Forest	SMOTEENN	0.145266	0.785965	0.245211
Logistic Regression	SMOTEENN	0.139406	0.708772	0.232987
XGBoost	SMOTEENN	0.160123	0.729825	0.262626
Decision Tree	RandomOverSampler	0.225352	0.224561	0.224956
Random Forest	RandomOverSampler	0.500000	0.017544	0.033898
Logistic Regression	RandomOverSampler	0.127425	0.852632	0.221715
XGBoost	RandomOverSampler	0.245989	0.322807	0.279211
Decision Tree	SMOTE	0.176316	0.235088	0.201504
Random Forest	SMOTE	0.538462	0.024561	0.046980
Logistic Regression	SMOTE	0.108911	0.038596	0.056995
XGBoost	SMOTE	0.357143	0.070175	0.117302



SUMMARY

FRAUD DETECTION

Model	Sampler	Precision	Recall	F1 Score	Accuracy Score
Decision Tree	RandomUnderSampler	0.126498	0.664336	0.212528	0.695633
Random Forest	RandomUnderSampler	0.137001	0.888112	0.237383	0.647211
Logistic Regression	RandomUnderSampler	0.137310	0.853147	0.236549	0.659533
XGBoost	RandomUnderSampler	0.148855	0.818182	0.251884	0.699524
Decision Tree	SMOTEENN	0.857143	0.020979	0.040956	0.939256
Random Forest	SMOTEENN	0.000000	0.000000	0.000000	0.938176
Logistic Regression	SMOTEENN	0.000000	0.000000	0.000000	0.938176
XGBoost	SMOTEENN	0.000000	0.000000	0.000000	0.938176
Decision Tree	RandomOverSampler	0.190323	0.206294	0.197987	0.896671
Random Forest	RandomOverSampler	0.500000	0.017483	0.033784	0.938176
Logistic Regression	RandomOverSampler	0.131319	0.790210	0.225212	0.663856
XGBoost	RandomOverSampler	0.207317	0.297203	0.244253	0.886295
Decision Tree	SMOTE	0.182109	0.199301	0.190317	0.895158
Random Forest	SMOTE	0.875000	0.024476	0.047619	0.939473
Logistic Regression	SMOTE	0.176471	0.010490	0.019802	0.935798
XGBoost	SMOTE	0.480000	0.041958	0.077170	0.937959

No model performed well

SUMMARY

FRAUD DETECTION

No model performed well

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- Artificial Neural Network

Oversampling vs undersampling

One-hot encoding vs label encoding

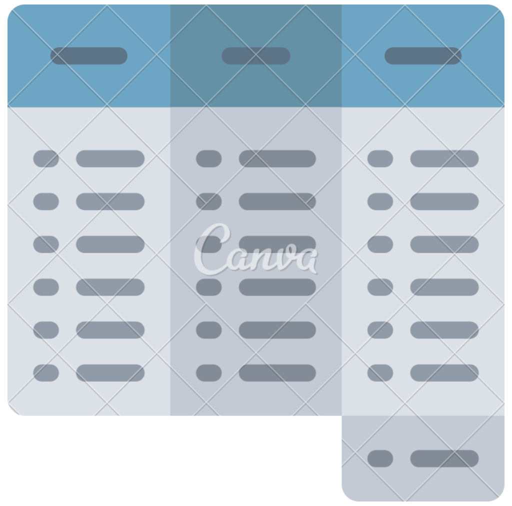
Combined variables, e.g. base-policy + vehicle type

Frequency-encoding, e.g. months/make/day high vs low count

FUTURE DIRECTIONS

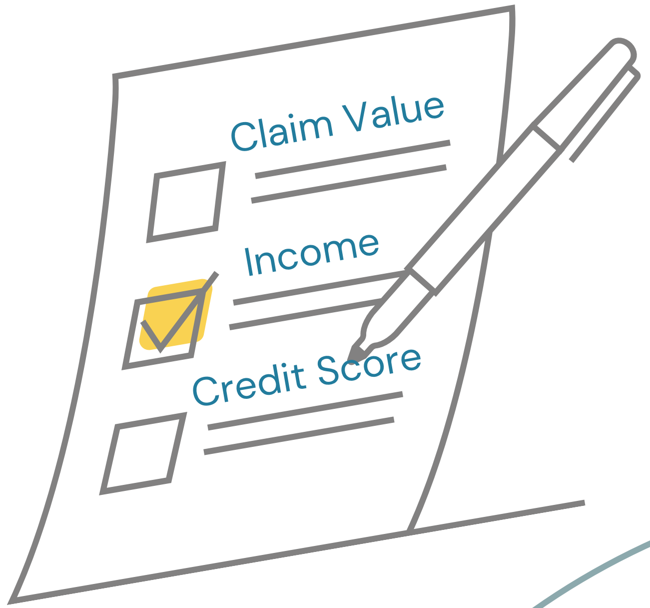
Improving model capability

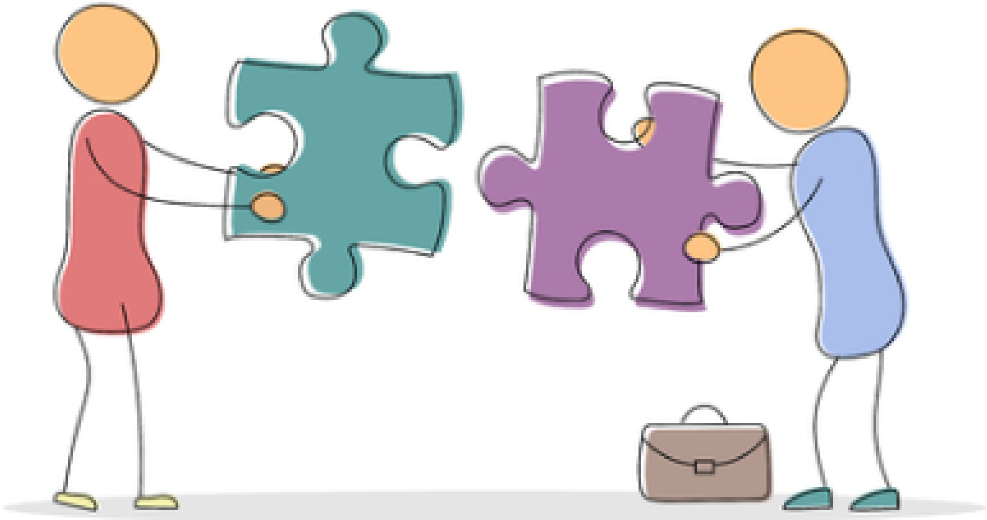
Expanding the dataset



Additional segmentation

Introducing new features





c'est fini

THANK YOU!

Sepi

Yana