DADOS E APRENDIZAGEM AUTOMÁTICA

Mestrado em Engenharia Informártica

Ana Murta (pg50184)

Ana Henriques (pg50196)

Gonçalo Soares (pg50393)

Diogo Pires (pg50334)

METODOLOGIA CRISP-DM

DATASET

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer Name	Segment	City	State	Country
32298	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495	Rick Hansen	Consumer	New York	New York	United States
26341	IN-2013-77878	05-02-2013	07-02-2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongo	New South Wales	Australia
25330	IN-2013-71249	17-10-2013	18-10-2013	First Class	CR-12730	Craig Reiter	Consumer	Brisbane	Queensland	Australia
13524	ES-2013-1579342	28-01-2013	30-01-2013	First Class	KM-16375	Katherine Murray	Home Office	Berlin	Berlin	Germany
47221	SG-2013-4320	05-11-2013	06-11-2013	Same Day	RH-9495	Rick Hansen	Consumer	Dakar	Dakar	Senegal
22732	IN-2013-42360	28-06-2013	01-07-2013	Second Class	JM-15655	Jim Mitchum	Corporate	Sydney	New South Wales	Australia
30570	IN-2011-81826	07-11-2011	09-11-2011	First Class	TS-21340	Toby Swindell	Consumer	Porirua	Wellington	New Zealand

Postal Code	Market	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority
10024	US	East	TEC-AC-10003033	Technology	Accessories	Plantronics CS510 - Over-the	2309,65	7	0	762,1845	933,57	Critical
	APAC	Oceania	FUR-CH-10003950	Furniture	Chairs	Novimex Executive Leather	3709,395	9	0,1	-288,765	923,63	Critical
	APAC	Oceania	TEC-PH-10004664	Technology	Phones	Nokia Smart Phone, with Cal	5175,171	9	0,1	919,971	915,49	Medium
	EU	Central	TEC-PH-10004583	Technology	Phones	Motorola Smart Phone, Cord	2892,51	5	0,1	-96,54	910,16	Medium
	Africa	Africa	TEC-SHA-10000501	Technology	Copiers	Sharp Wireless Fax, High-Sp	2832,96	8	0	311,52	903,04	Critical
	APAC	Oceania	TEC-PH-10000030	Technology	Phones	Samsung Smart Phone, with	2862,675	5	0,1	763,275	897,35	Critical
	APAC	Oceania	FUR-CH-10004050	Furniture	Chairs	Novimex Executive Leather	1822,08	4	0	564,84	894,77	Critical

ANÁLISE E TRATAMENTO DOS DADOS

- Descrição das object de features

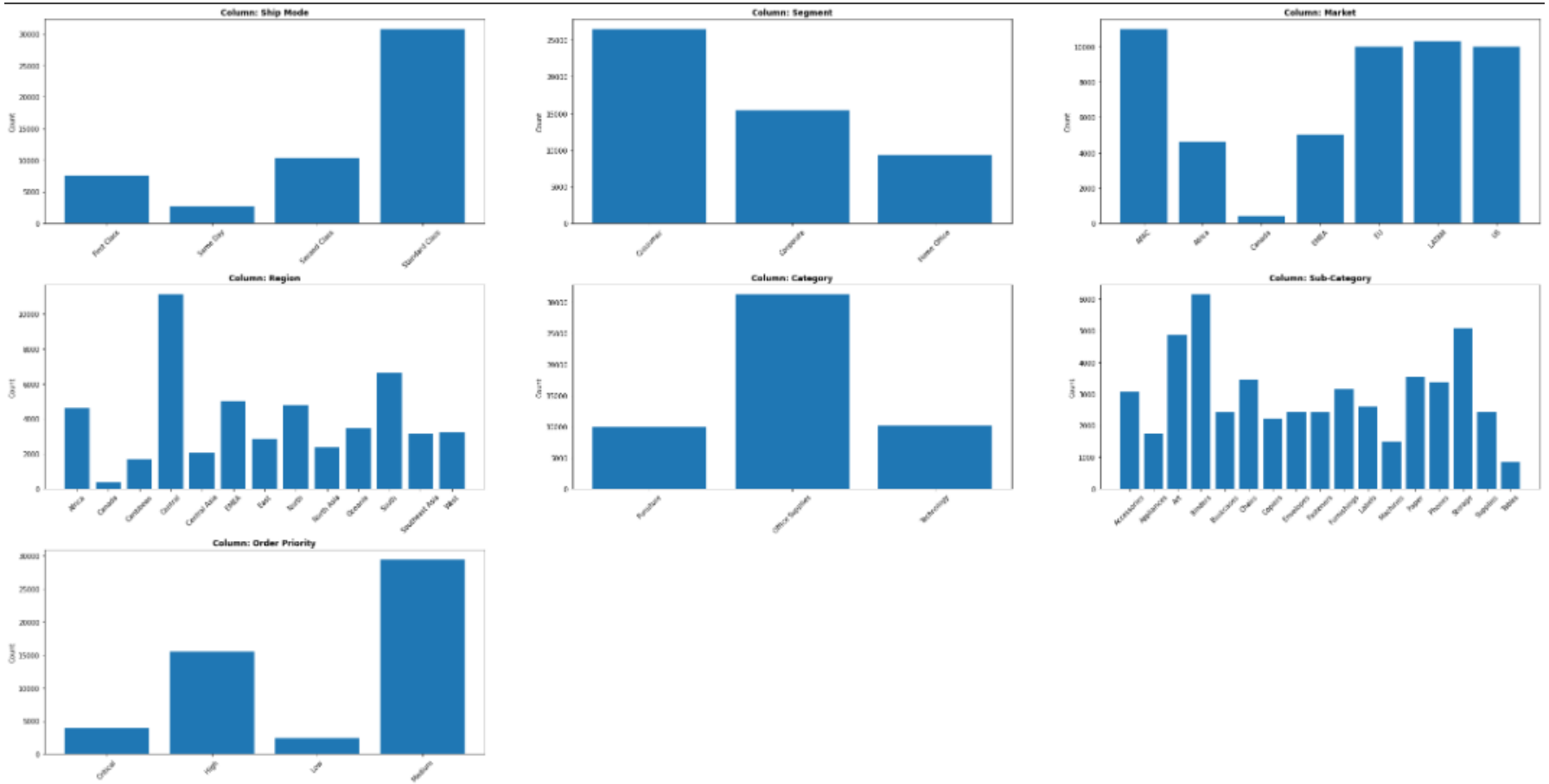
	Order ID	Ship Mode	Customer ID	Customer Name	Segment	City	State	Country	Postal Code	Market	Region	Product ID	Category	Sub-Category	Product Name	Order Priority
count	51290	51290	51290	51290	51290	51290	51290	51290	9994.0	51290	51290	51290	51290	51290	51251	51200
unique	25035	4	1590	795	3	3636	1094	147	631.0	7	13	10292	3	17	3781	4
top	CA-2014-100111	Standard Class	PO-18850	Muhammed Yedwab	Consumer	New York City	California	United States	10035.0	APAC	Central	OFF-AR-10003651	Office Supplies	Binders	Staples	Medium
freq	14	30775	97	108	26518	915	2001	9994	263.0	11002	11117	35	31273	6152	227	29386

- Descrição das features numéricas

	Sales	Quantity	Discount	Profit	Shipping Cost
count	51200.000000	51200.000000	51200.000000	51200.000000	51200.000000
mean	246.865006	3.476426	0.143026	28.639338	26.416522
std	487.908698	2.278923	0.212409	174.491125	57.338685
min	0.444000	1.000000	0.000000	-6599.978000	0.000000
25%	30.880000	2.000000	0.000000	0.000000	2.620000
50%	85.273800	3.000000	0.000000	9.255100	7.810000
75%	251.640000	5.000000	0.200000	36.841500	24.530000
max	22638.480000	14.000000	0.850000	8399.976000	933.570000

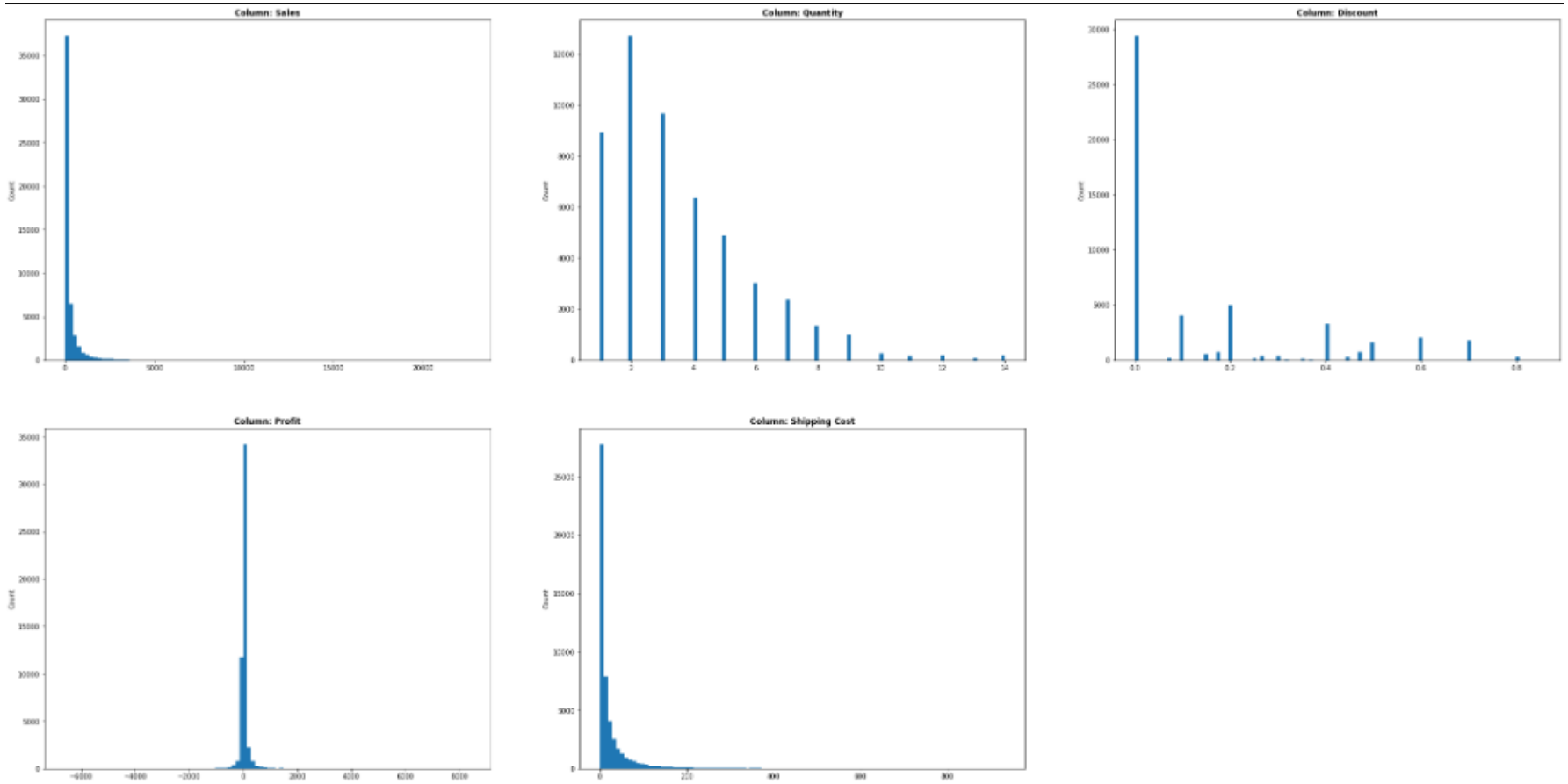
ANÁLISE E TRATAMENTO DOS DADOS

- Distribuição das variáveis categóricas



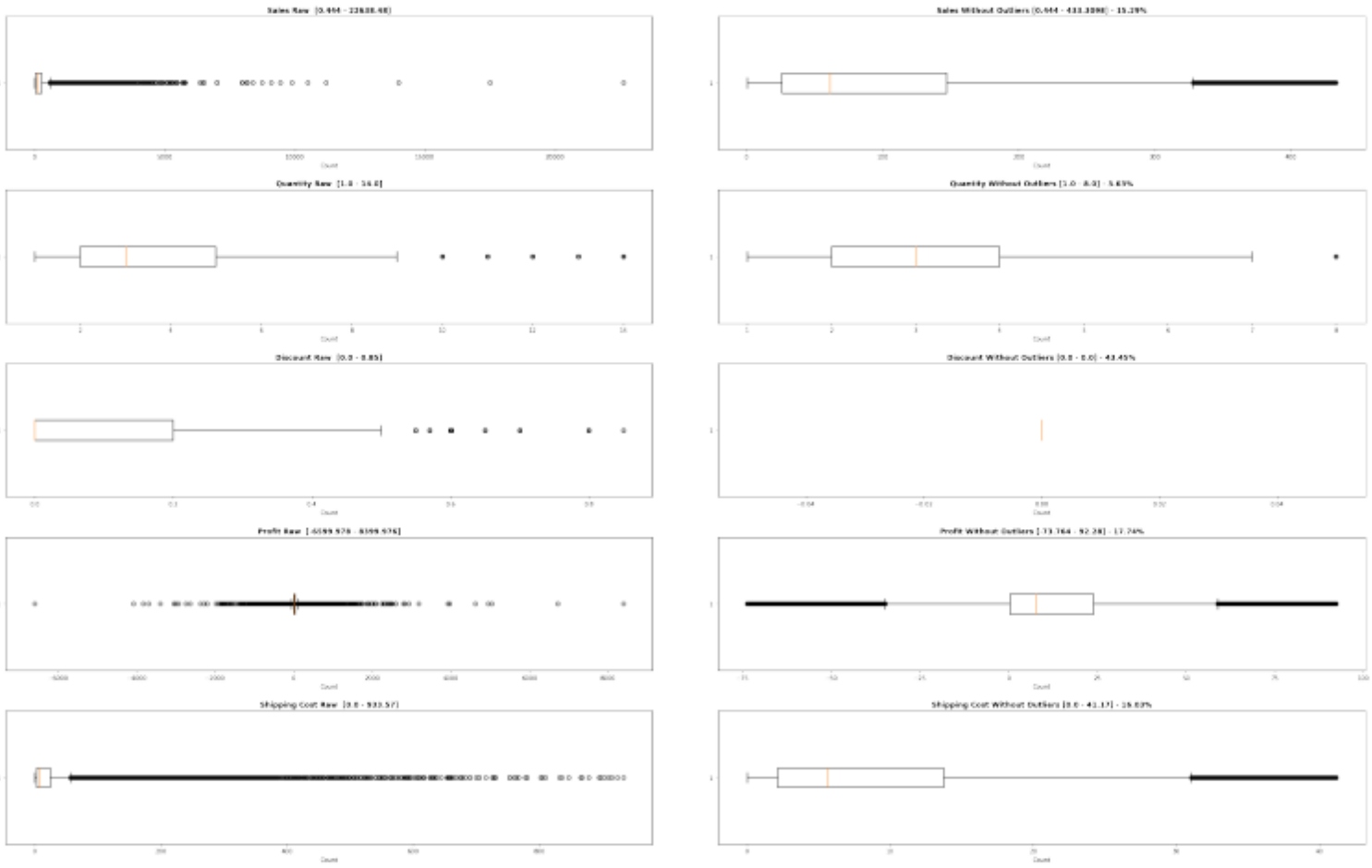
ANÁLISE E TRATAMENTO DOS DADOS

- Distribuição das variáveis numéricas



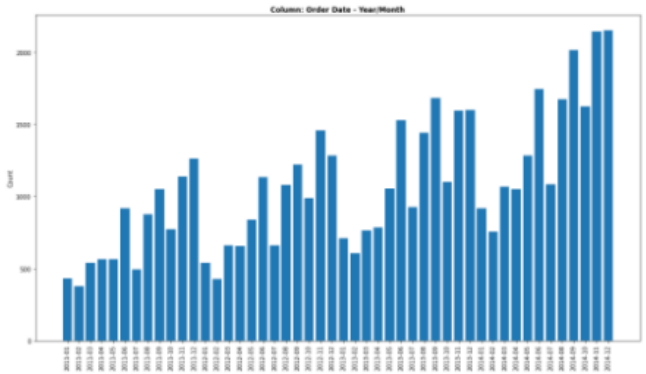
ANÁLISE E TRATAMENTO DOS DADOS

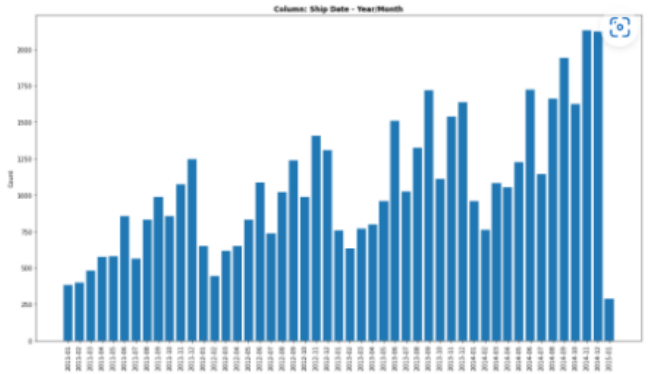
- Análise da dispersão dos dados



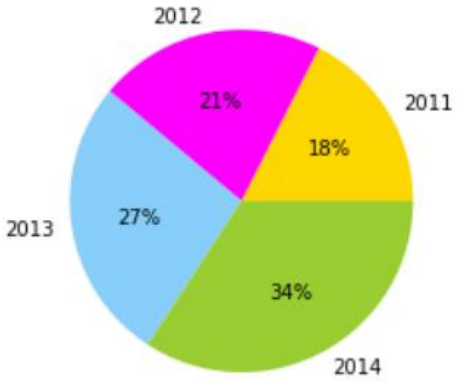
ANÁLISE E TRATAMENTO DOS DADOS

- Dispersão temporal

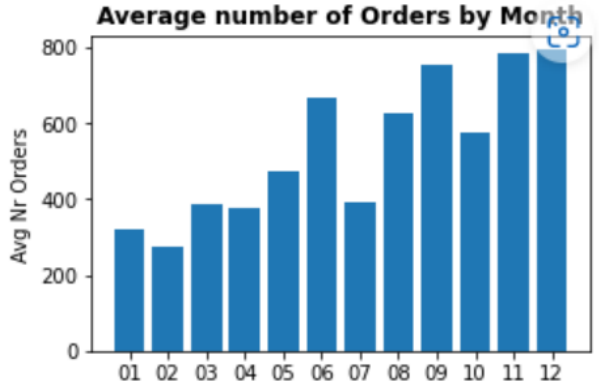




- Quantidade de encomendas por ano

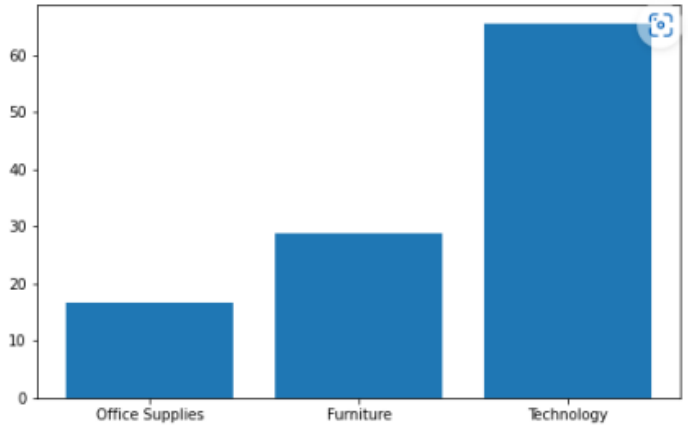


- Média de encomendas por mês ao todo

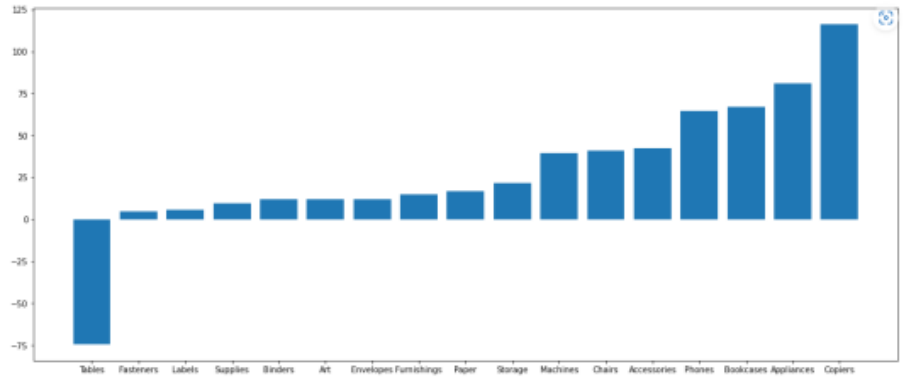


ANÁLISE E TRATAMENTO DOS DADOS

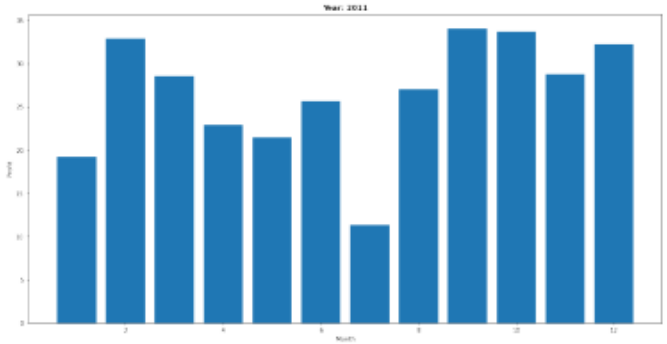
- Lucro por categoria

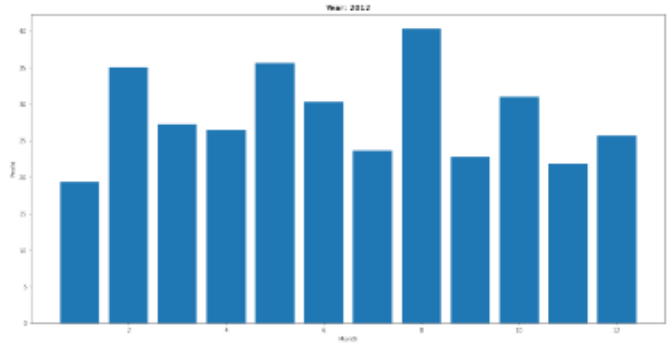


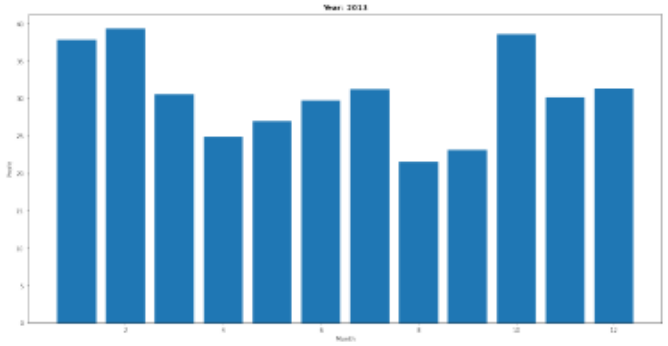
- Lucro por sub-categoria

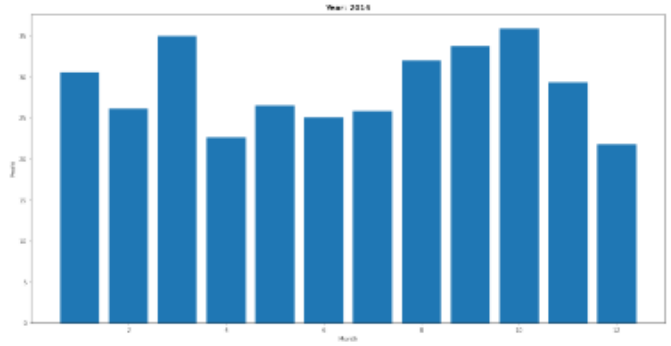


- Lucro por mês de cada ano



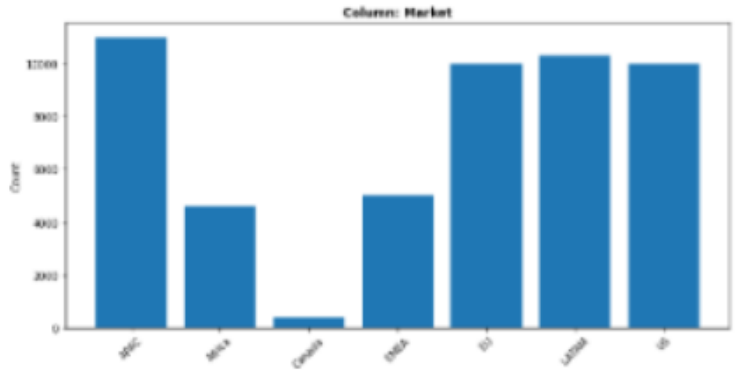


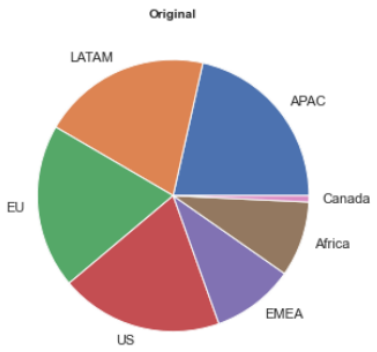


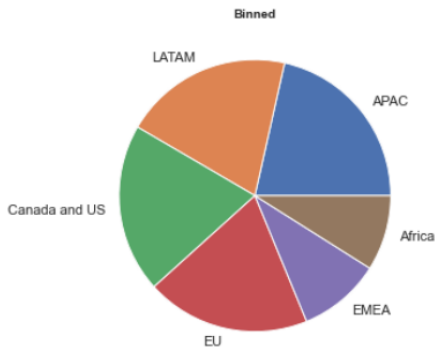


ANÁLISE E TRATAMENTO DOS DADOS

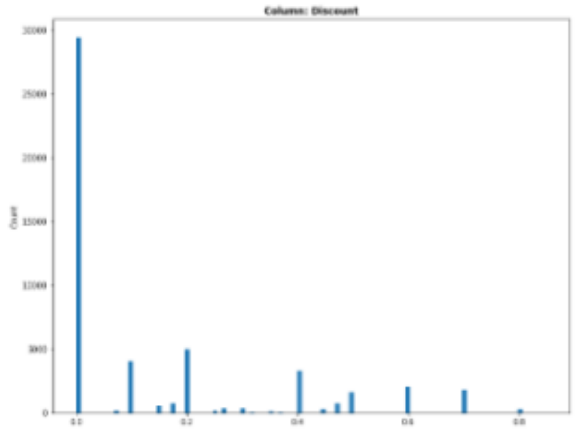
- Market

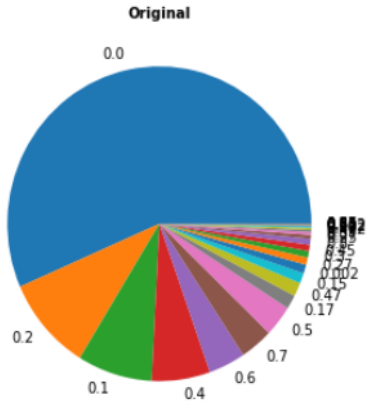


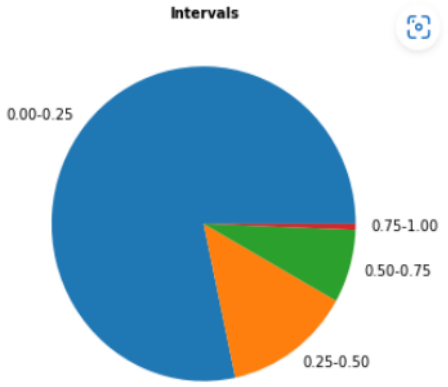




- Discount

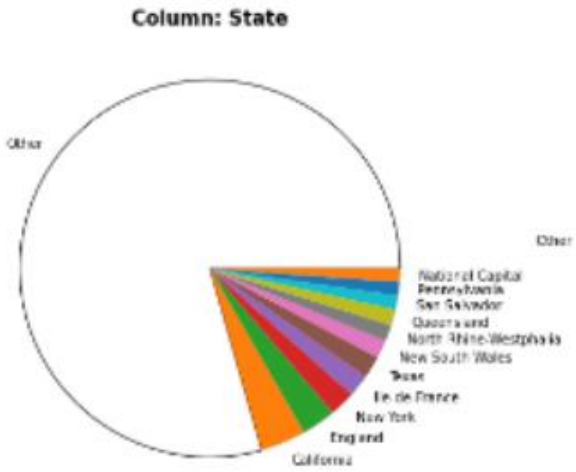


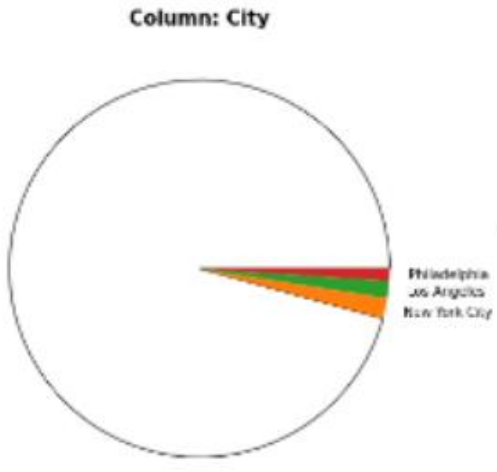


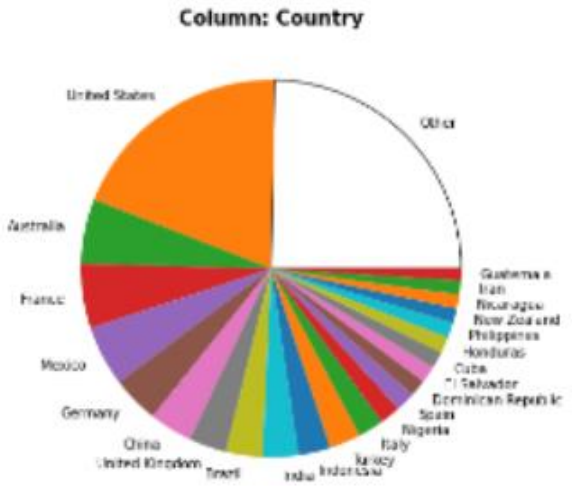


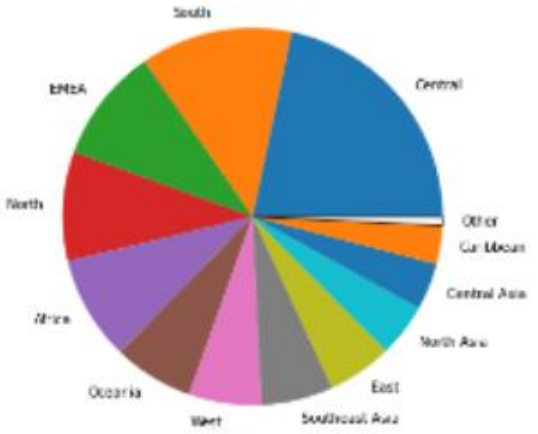
ANÁLISE E TRATAMENTO DOS DADOS

- Estratégia da percentagem



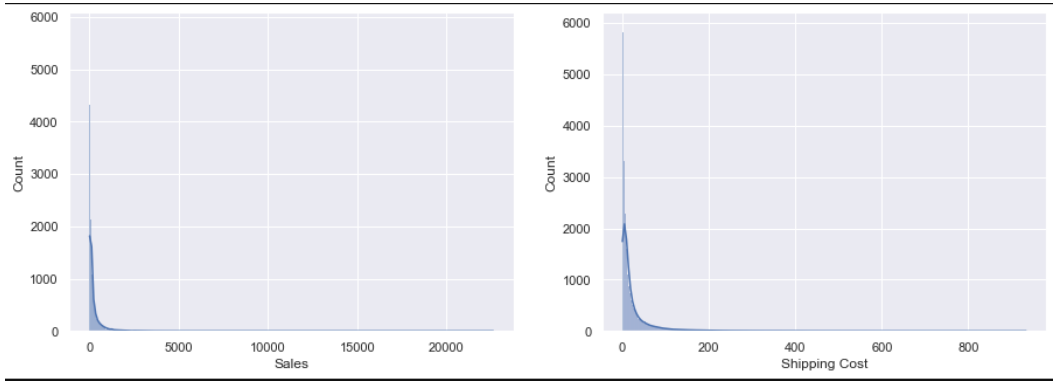




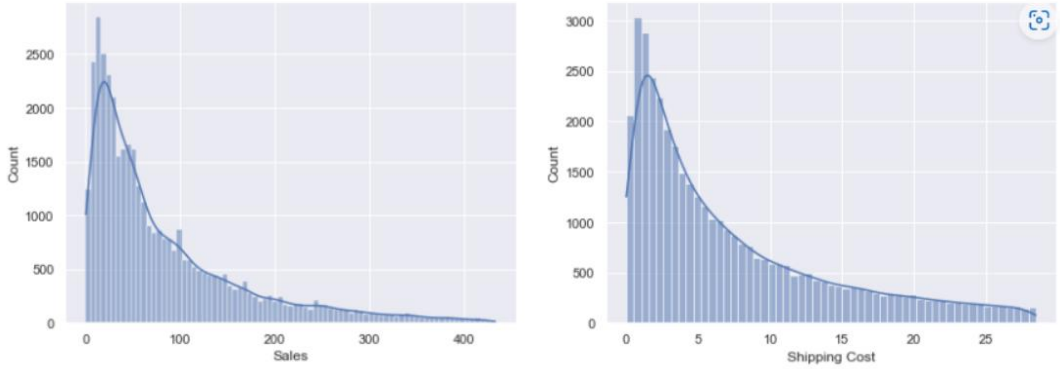


ANÁLISE E TRATAMENTO DOS DADOS

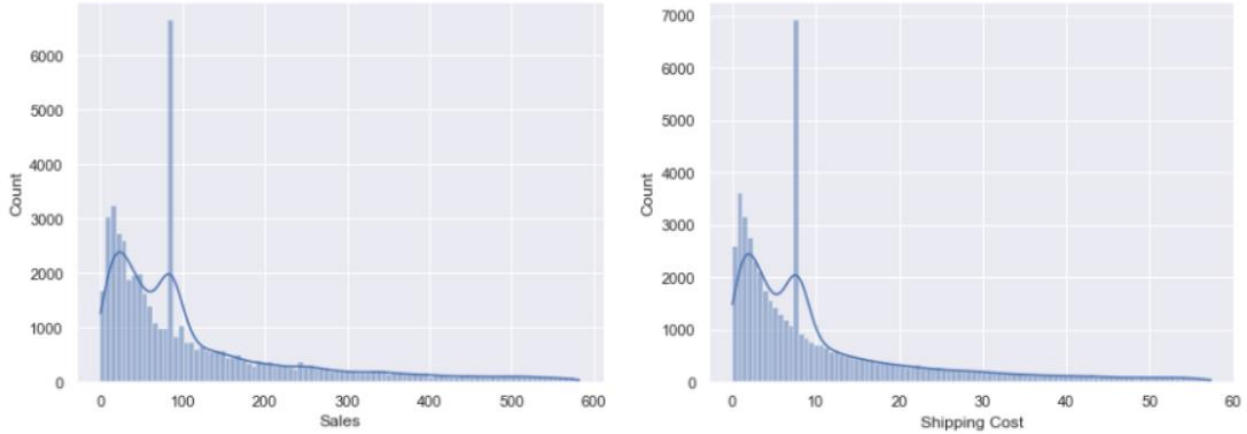
- Distribuição do Sales e Shipping Cost
  - com outliers



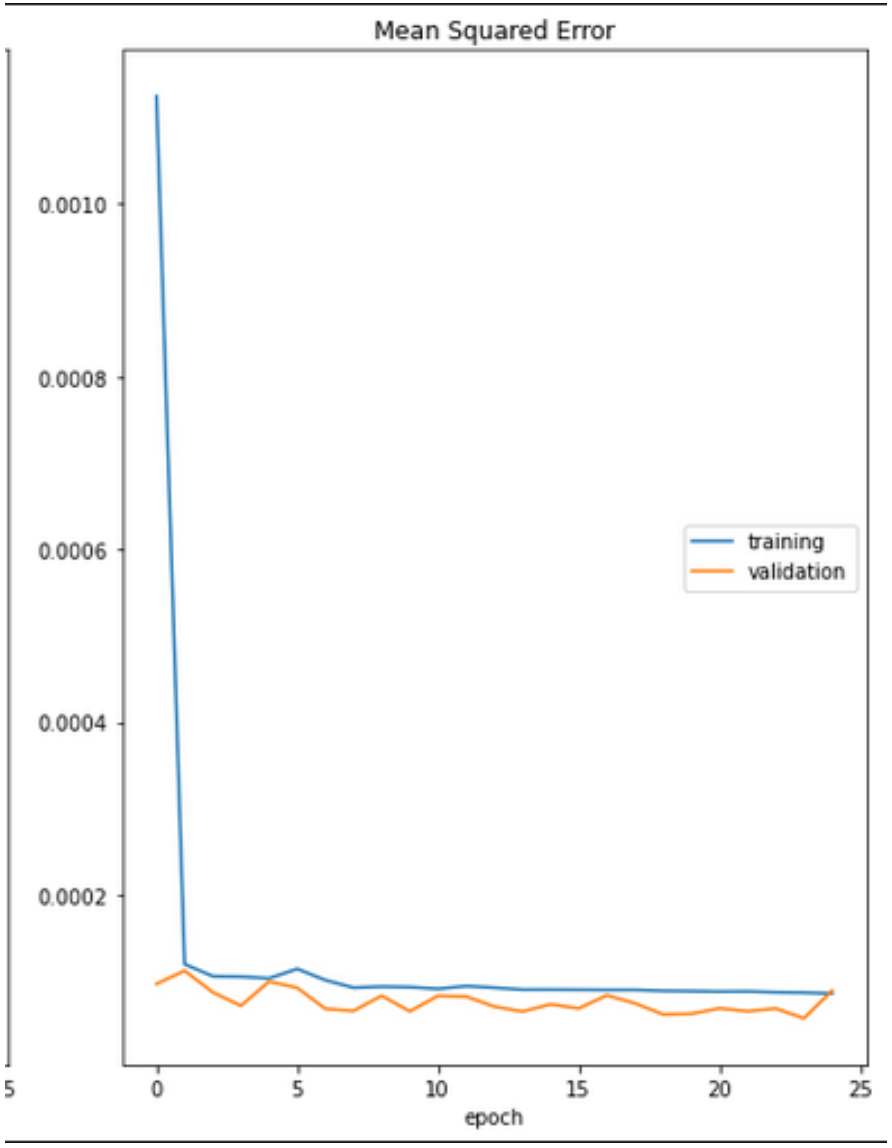
  - Sem outliers

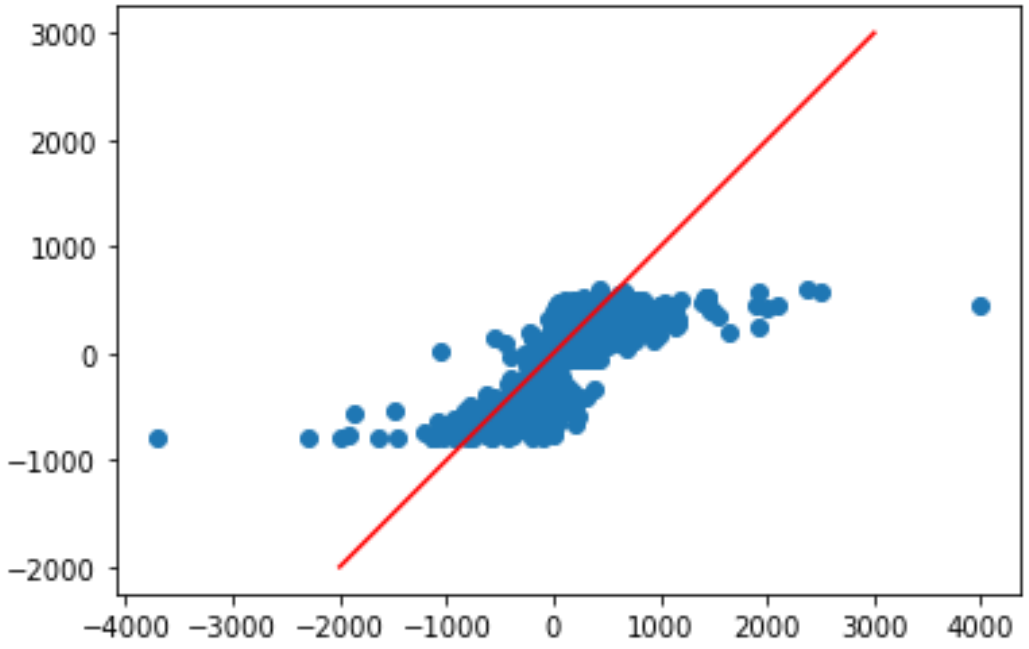


- Substituição dos outliers pela mediana



MODELOS
REDES NEURONAIS

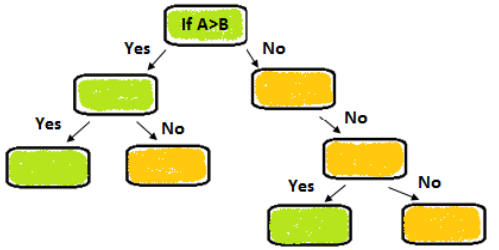


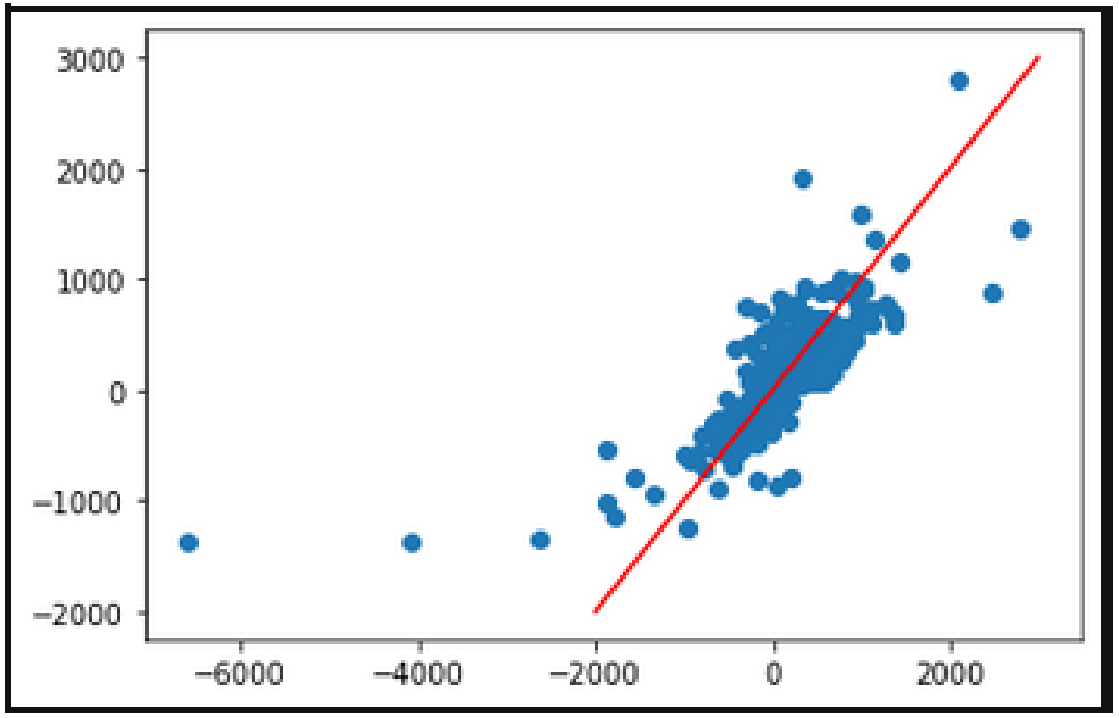


RMSE: 141.74

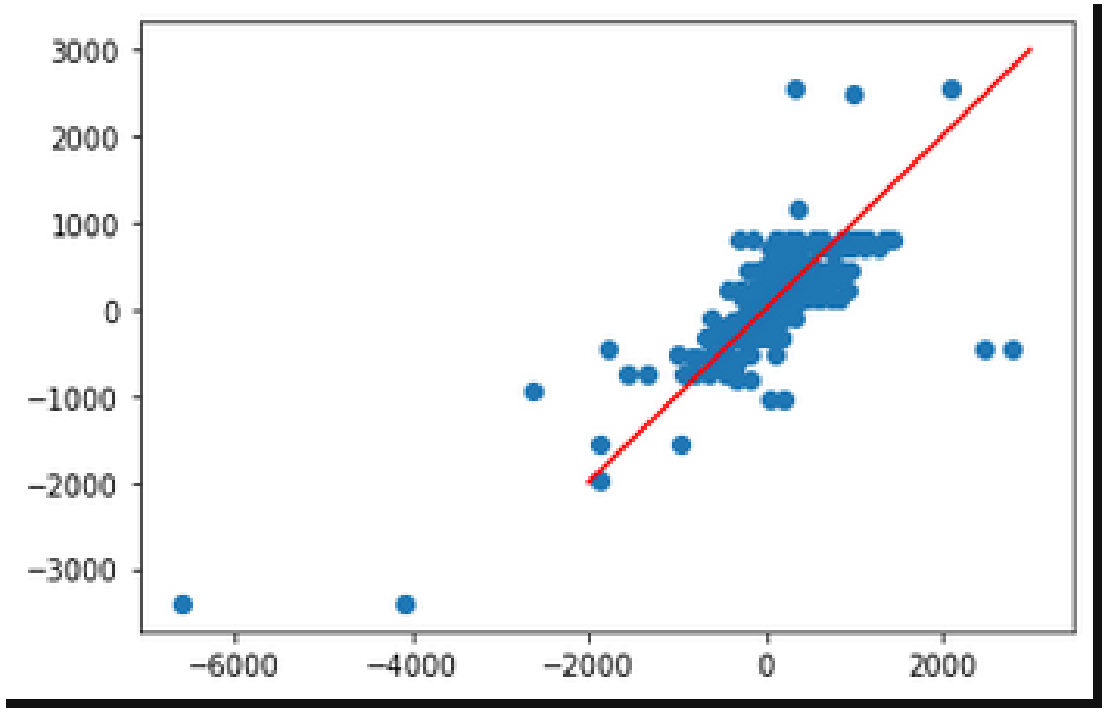
MODELOS

RANDOM FOREST REGRESSOR



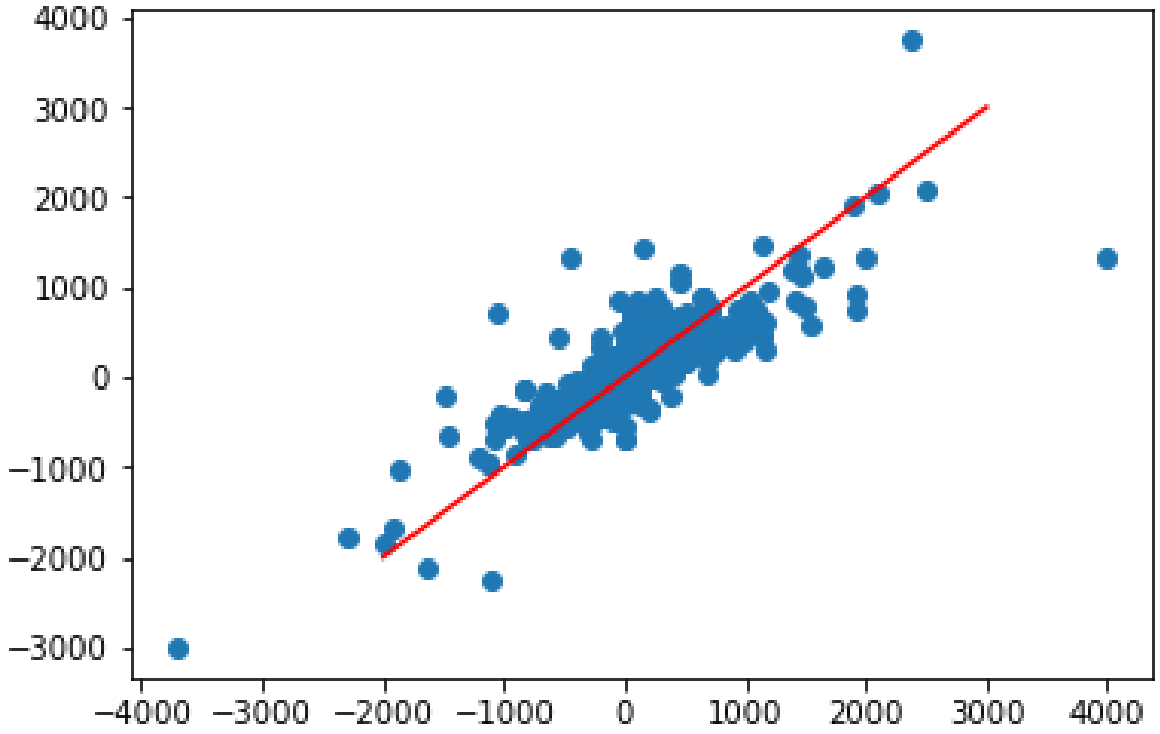


RMSE: 117.08



RMSE: 116.49

MODELOS
XGBOOST



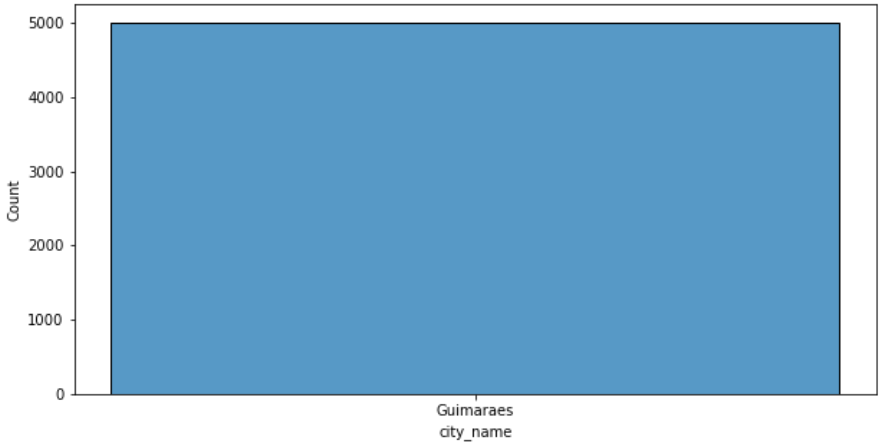

```
RMSE:  91.97666600086227
1 Segment
1 City
1 State
1 Sub-Category
1 Sales
1 Quantity
1 Discount
1 order_month
1 ship_day
2 Country
3 pib_country
4 Shipping Cost
5 Category
6 order_day
7 Region
8 Ship Mode
9 discount_bracket
10 Order Priority
11 order_year
12 ship_year
13 ship_month
14 ship_mode_binned
15 Market_binned
```

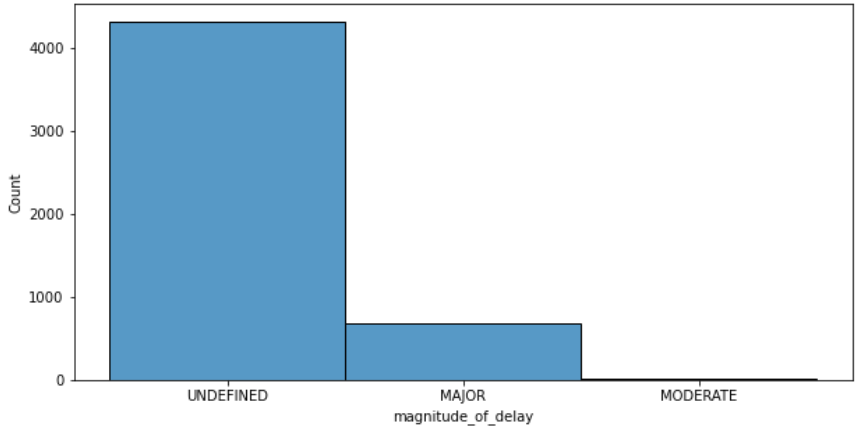
DATASET COMPETIÇÃO

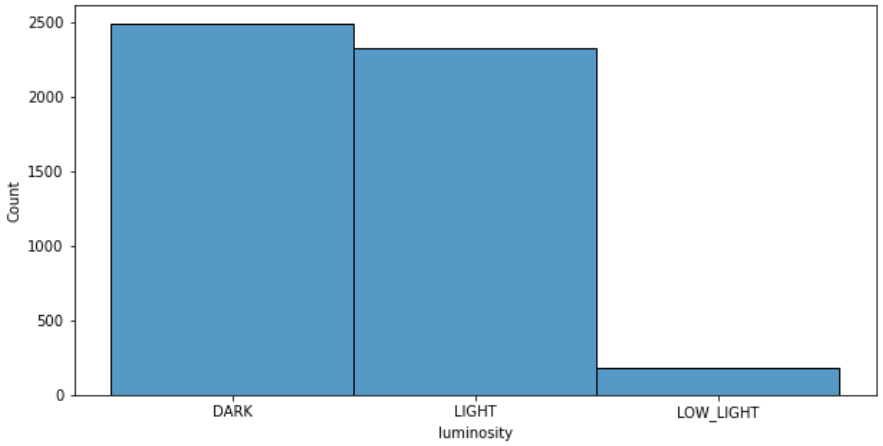
	city_name	magnitude_of_delay	delay_in_seconds	affected_roads	record_date	luminosity	avg_temperature	avg_atm_pressure
0	Guimaraes	UNDEFINED	0	,	2021-03-15 23:00	DARK	12.0	1013.0
1	Guimaraes	UNDEFINED	385	N101,	2021-12-25 18:00	DARK	12.0	1007.0
2	Guimaraes	UNDEFINED	69	,	2021-03-12 15:00	LIGHT	14.0	1025.0
3	Guimaraes	MAJOR	2297	N101,R206,N105,N101,N101,N101,N101,N101,N101,N...	2021-09-29 09:00	LIGHT	15.0	1028.0
4	Guimaraes	UNDEFINED	0	N101,N101,N101,N101,N101,	2021-06-13 11:00	LIGHT	27.0	1020.0

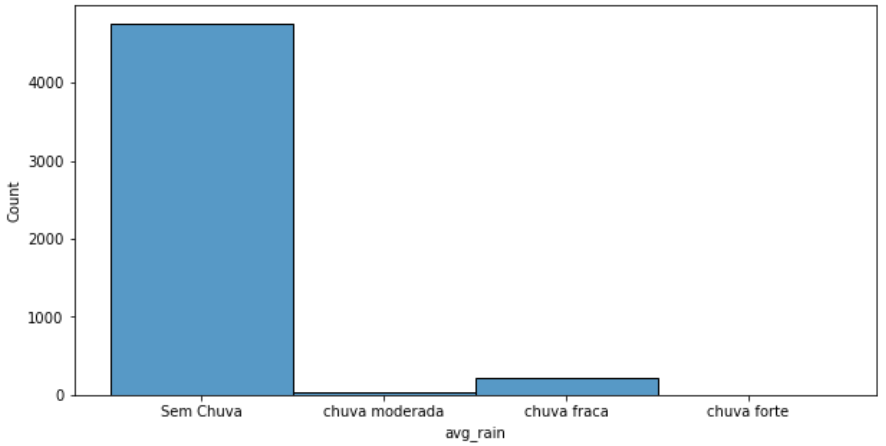
avg_humidity	avg_wind_speed	avg_precipitation	avg_rain	incidents
70.0	1.0	0.0	Sem Chuva	None
91.0	1.0	0.0	Sem Chuva	None
64.0	0.0	0.0	Sem Chuva	Low
75.0	1.0	0.0	Sem Chuva	Very_High
52.0	1.0	0.0	Sem Chuva	High

COMPETIÇÃO - ANÁLISE DE DADOS

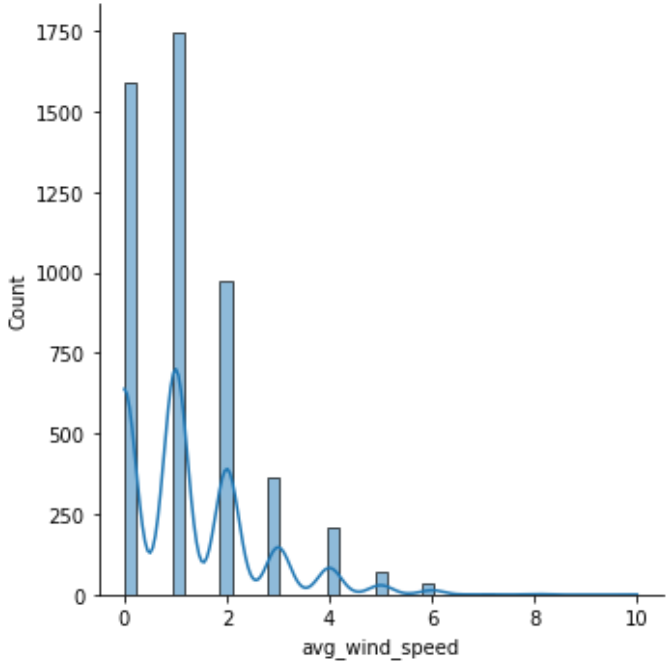


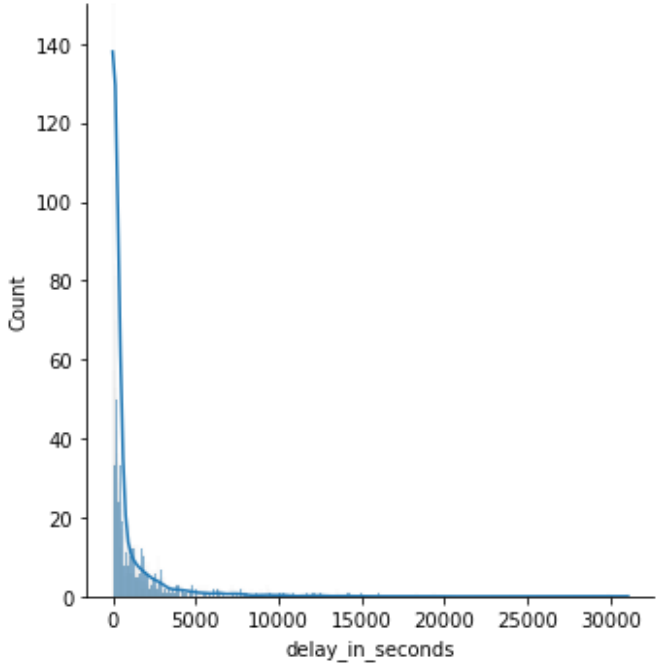


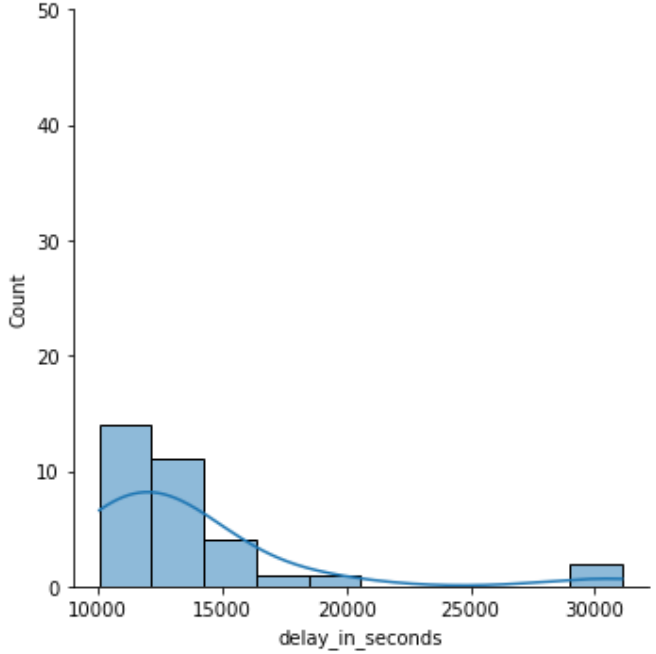




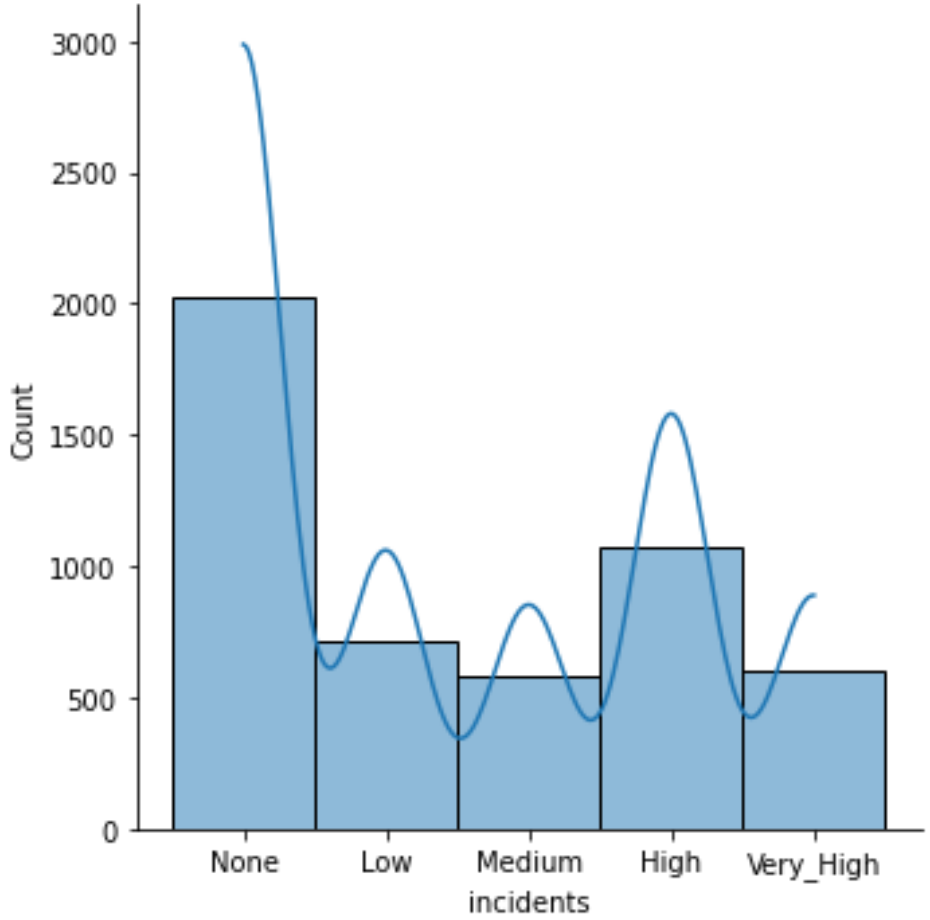
COMPETIÇÃO - ANÁLISE DE DADOS



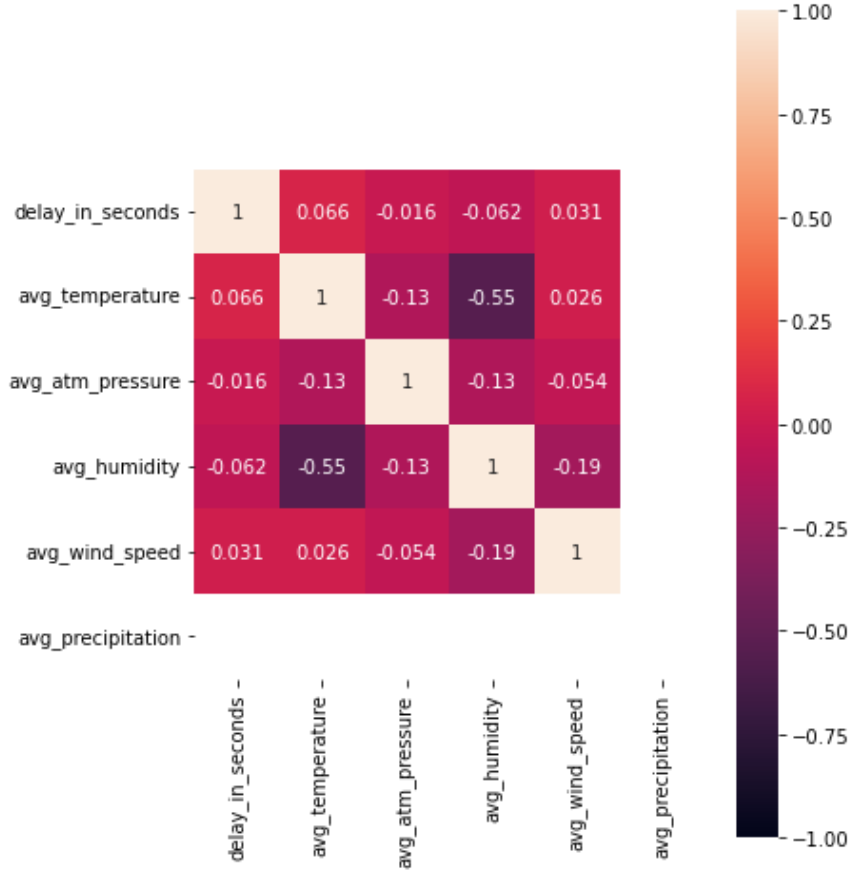




COMPETIÇÃO - ANÁLISE DE DADOS



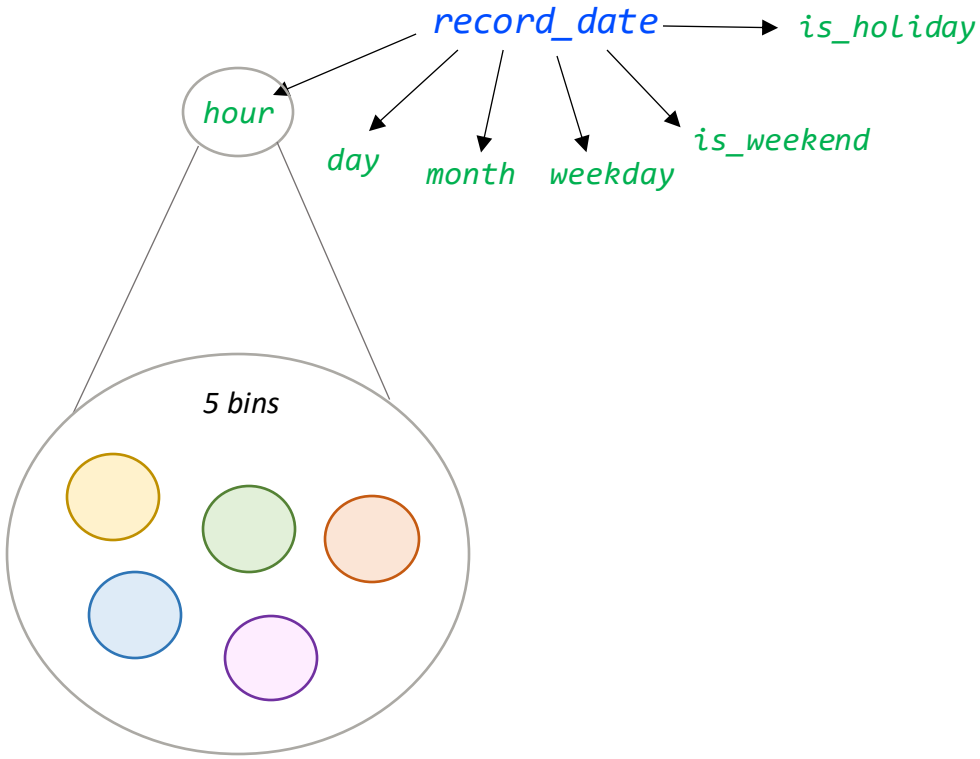
COMPETIÇÃO - ANÁLISE DE DADOS



COMPETIÇÃO - TRATAMENTO DE DADOS

Feature Selection

Feature Engineering



COMPETIÇÃO - TRATAMENTO DE DADOS

One-Hot Encoding

affected_roads	N101	IC5	N105
N101	1	0	0
N105	0	0	1
IC5	0	1	0

delay_in_seconds

delay_in_minutes

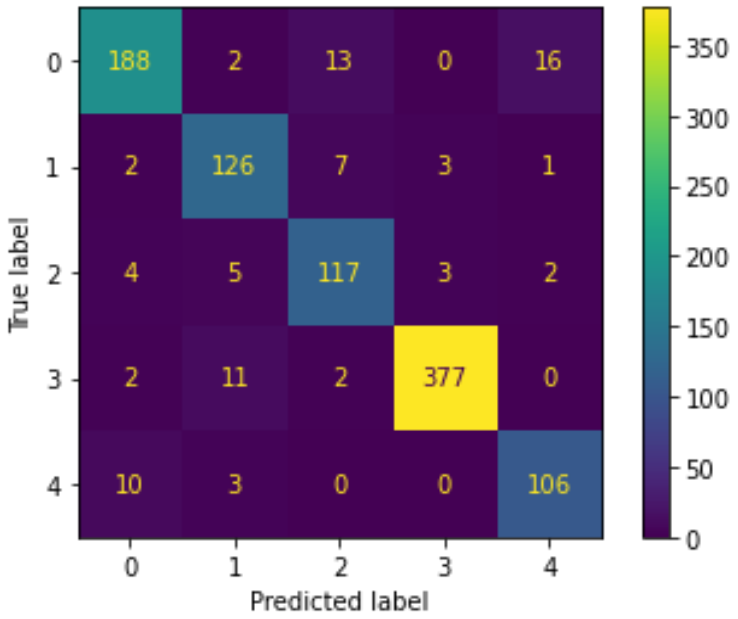
$$X_{normalized} = \frac{(X - X_{minimum})}{(X_{minimum} - X_{minimum})}$$

COMPETIÇÃO - MODELAÇÃO

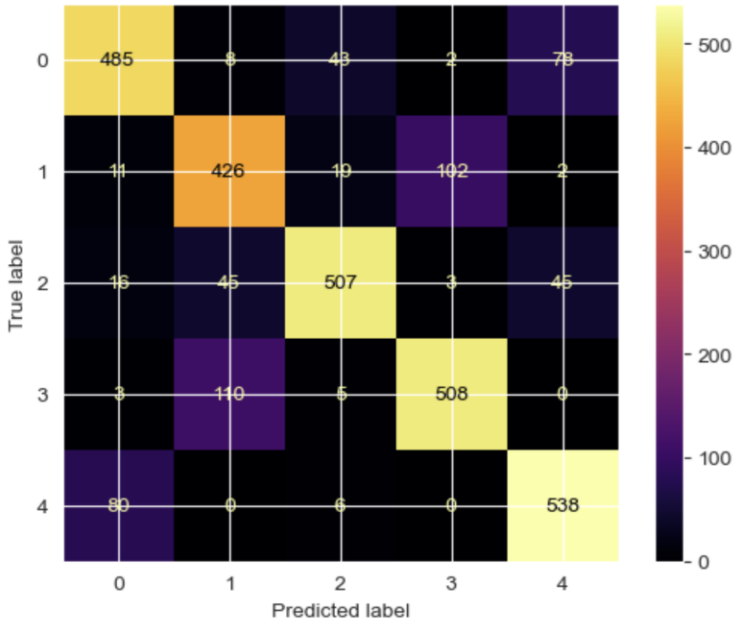
```
    'model': RandomForestClassifier(),
    'params': {
        'n_estimators': [100, 200, 300, 400, 500],
        'max_depth': [3, 5, 7, 9],
        'min_samples_split': [20, 25, 30, 35, 40],
        'min_samples_leaf': [1, 2, 4]
    }
},
'XGBoost': {
    'model': XGBClassifier(gpu_id=0, tree_method='gpu_hist'),
    'params': {
        'n_estimators': [100, 200, 300, 400, 500],
        'max_depth': [5, 6, 7, 8],
        'eta': [0.1, 0.15, 0.2],
    }
},
'CatBoost': {
    'model': CatBoostClassifier(),
    'params': {
        'n_estimators': [100, 200, 300, 400, 500],
        'max_depth': [5, 6, 7, 8],
        'learning_rate': [0.05, 0.1, 0.15, 0.2],
        'early_stopping_rounds': [10],
        'verbose': [False]
    }
},
```

```
grid_search = GridSearchCV(estimator=model, param_grid=params, cv=CV_FOLDS, n_jobs=-1,
                           verbose=1)
grid_search.fit(X, y)

print_model_prefix('CV', model_name, "Best Params", grid_search.best_params_)
print_model_prefix('CV', model_name, "Best Score", grid_search.best_score_)
print_model_prefix('CV', model_name, "Best Estimator", grid_search.best_estimator_)
print_model_prefix('Train', model_name, "Accuracy", grid_search.score(X, y))

models_trained[model_name] = grid_search.best_estimator_
```

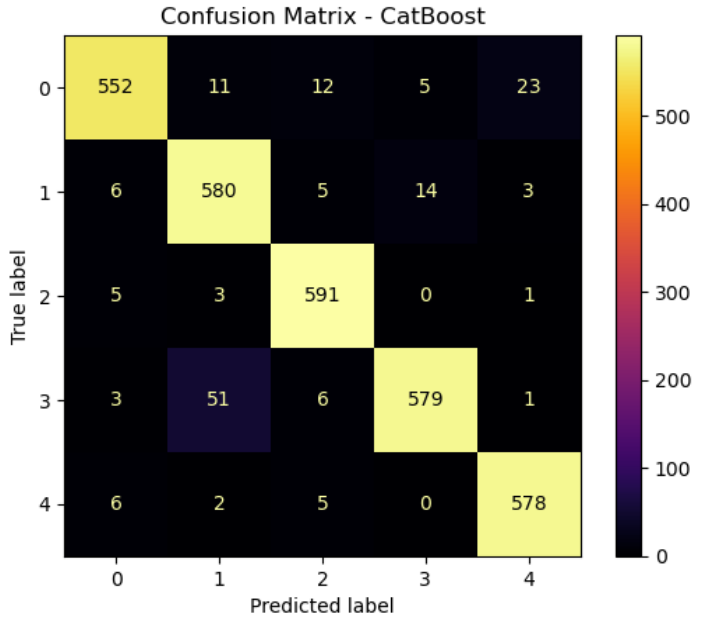
COMPETIÇÃO - AVALIAÇÃO



Decision Tree
Test Accuracy: 0.914



Random Forest
Test Accuracy: 0.960



Cat Boost
Test Accuracy: 0.947

COMPETIÇÃO - *DEPLOYMENT*

Catboost

Private Score	Public Score
0.91005	0.93351

Decision Tree

Private Score	Public Score
0.91242	0.93074