Analise_ML

February 27, 2021

0.0.1 Dados

[2]:

	Pred_class	probabilidade	status	True_class
0	2	0.079892	approved	0.0
1	2	0.379377	approved	74.0
2	2	0.379377	approved	74.0
3	2	0.420930	approved	74.0
4	2	0.607437	approved	NaN
5	2	0.690894	approved	NaN
6	2	0.759493	approved	NaN
7	2	0.834910	approved	NaN
8	2	0.861396	approved	NaN
9	2	1.000000	approved	NaN

0.0.2 Checando os missings

[3]:

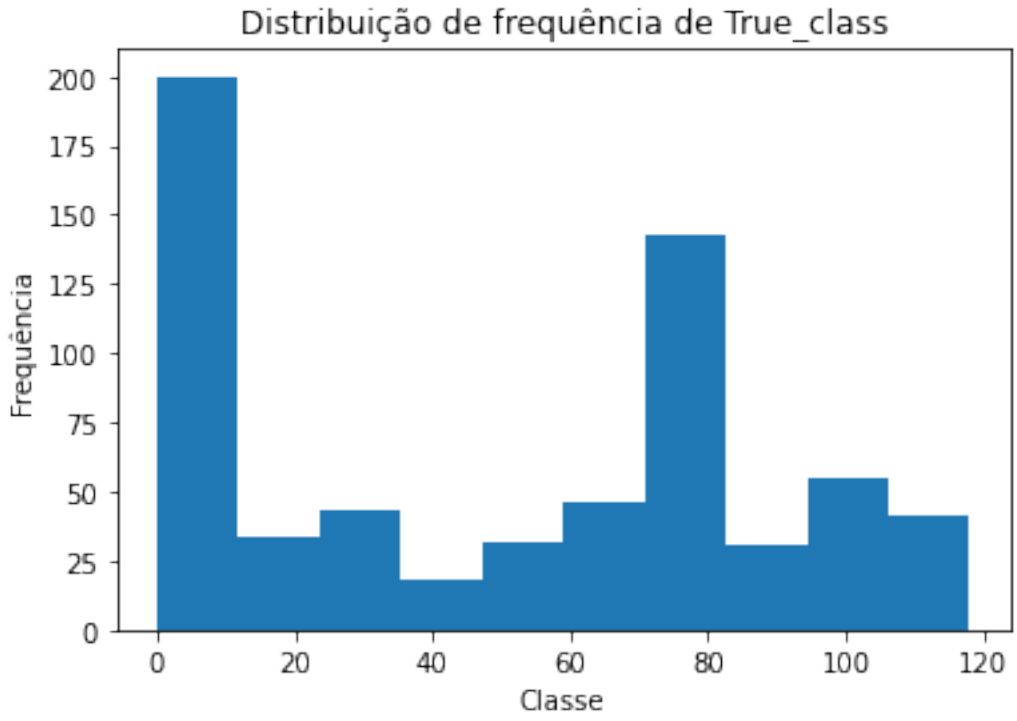
```
Pred_class         0
probabilidade      0
status             0
True_class       462
dtype: int64
```

0.0.3 Substituindo os missings da True_class pelos valores de Pred_class

[5]:

	Pred_class	probabilidade	status	True_class
0	2	0.079892	approved	0
1	2	0.379377	approved	74
2	2	0.379377	approved	74
3	2	0.420930	approved	74
4	2	0.607437	approved	2
5	2	0.690894	approved	2
6	2	0.759493	approved	2
7	2	0.834910	approved	2
8	2	0.861396	approved	2
9	2	1.000000	approved	2

1 1. Análise exploratória dos dados utilizando estatística descritiva e inferencial, considerando uma, duas e/ou mais variáveis;

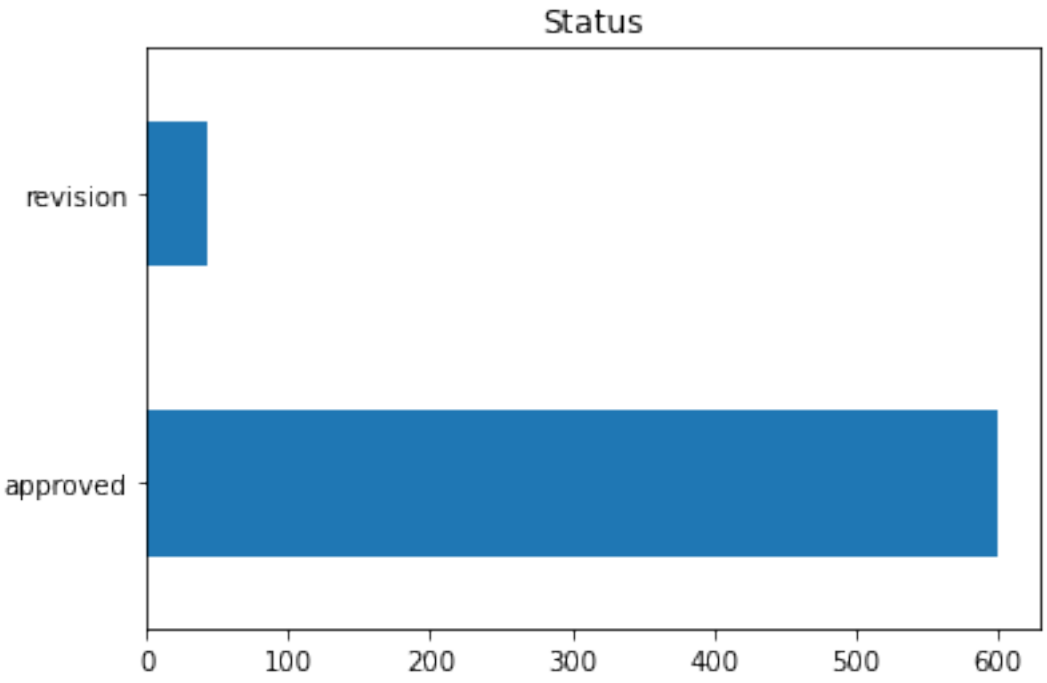


1.0.1 Proporção de approved e revision para status

[7]:

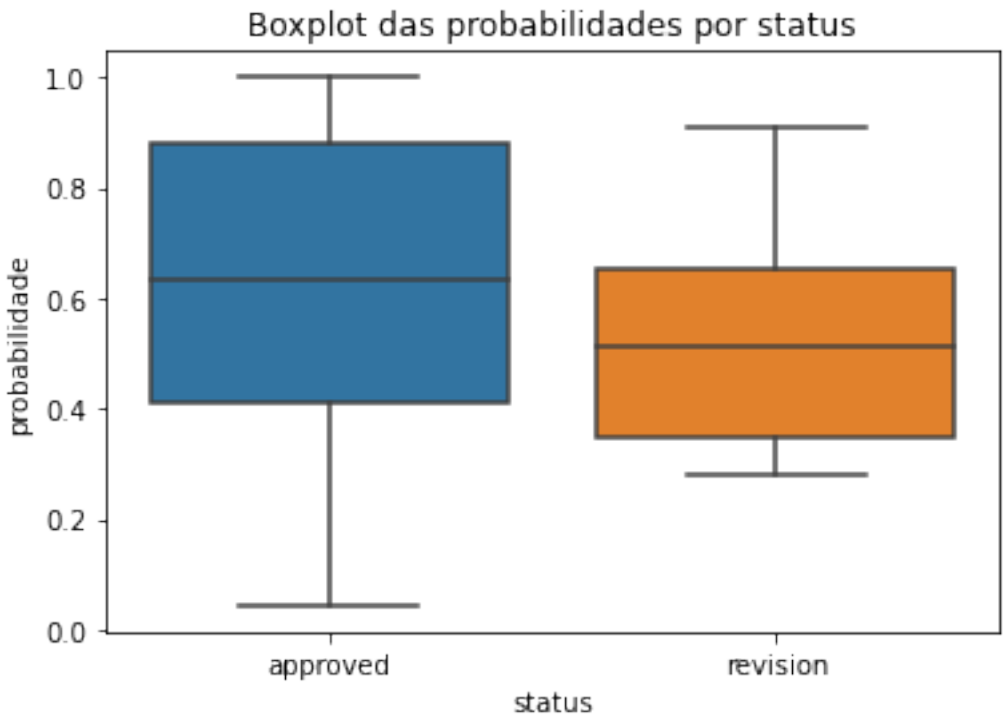
	Freq	Freq(%)
approved	600	93.31%
revision	43	6.69%

1.0.2 Visualização



1.0.3 Probabilidade x status

Bloxplot



Teste t Execução do teste t para testar se há diferença entre a média das probabilidades dos dois grupos (approved e revision)

[11]:
```
Ttest_indResult(statistic=2.5358689388146742, pvalue=0.011453227726784192)
```

Conclusão do teste: Com 95% de confiança, há evidência estatística de que há diferença entre as médias das probabilidades dos grupos 'approved' e 'revision'.

2 2. Calcule o desempenho do modelo de classificação utilizando pelo menos três métricas;

[14]:

	Métricas
Acurácia	0.718507
Precisão	0.653055
Recall	0.809530
F1	0.635151

3 3. Crie um classificador que tenha como output se os dados com status igual a revision estão corretos ou não (Sugestão : Técnica de cross-validation K-fold);

O classificador usado será o Random Forest. Como o interesse é saber se os dados da classe positiva estão classificados corretamente, a métrica apropriada é o recall.

```
Resultados do cross validation:
```

[18]:

	Recall
0	0.20
1	0.20
2	0.00
3	0.25
4	0.00
5	0.00
6	0.50
7	0.00
8	0.25
9	0.00

```
Média por fold: 0.14
```

4 4. Compare três métricas de avaliação aplicadas ao modelo e descreva sobre a diferença;

```
Cross validation para acurácia, precisão e recall
```

[21]:

	Acurácia	Precisão	Recall
0	0.91	0.33	0.20
1	0.89	0.25	0.20
2	0.88	0.00	0.00
3	0.95	1.00	0.25
4	0.81	0.00	0.00
5	0.94	0.00	0.00
6	0.95	0.67	0.50
7	0.91	0.00	0.00
8	0.95	1.00	0.25
9	0.92	0.00	0.00

```
Acurácia média: 0.91
Precisão média: 0.32
Recall médio: 0.14
```

- A acurácia mede a proporção de classificações corretas do modelo. Para cada fold, o modelo acertou, em média, 91% das classificações.

- A precisão mede a proporção de acertos que o modelo teve para a classe positiva; como exemplo, no primeiro fold, o modelo acertou 33% das observações que classificou como “revision”.
- O recall mede a proporção de observações da classe positiva que foram classificadas corretamente. Também no primeiro fold, 20% dos dados de treino com status “revision” foram classificados corretamente pelo modelo.