Relatório Competição 1

Enzo Cunha e Alexandre Costa

Novembro 2022

1 Análise exploratória dos dados

Abrindo o dataset, verifica-se:

- há 32 colunas, sendo 1 delas o target
- 20 colunas não são numéricas
- 5 colunas estão mais de 95% não preenchidas

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	227122 non-null	int64
1	NR_SEQ_REQUISICAO	227122 non-null	int64
2	NR_SEQ_ITEM	227122 non-null	int64
3	DT_REQUISICAO	227122 non-null	int64
4	DS_TIPO_GUIA	227122 non-null	object
5	DT_NASCIMENTO	227112 non-null	float64
6	NR_PRODUTO	227122 non-null	int64
7	DS_TIPO_PREST_SOLICITANTE	227122 non-null	object
8	DS_CBO	227122 non-null	object
9	DS_TIPO_CONSULTA	10511 non-null	object
10	QT_TEMPO_DOENCA	266 non-null	float64
11	DS_UNIDADE_TEMPO_DOENCA	266 non-null	object
12	DS_TIPO_DOENCA	531 non-null	object
13	DS_INDICACAO_ACIDENTE	209539 non-null	object
14	DS_TIPO_SAIDA	0 non-null	float64
15	DS_TIPO_INTERNACAO	59863 non-null	object
16	DS_REGIME_INTERNACAO	59863 non-null	object
17	DS_CARATER_ATENDIMENTO	227122 non-null	object
18	DS_TIPO_ACOMODACAO	59781 non-null	object
19	QT_DIA_SOLICITADO	58995 non-null	float64
20	CD_GUIA_REFERENCIA	37463 non-null	float64
21	DS_TIPO_ATENDIMENTO	168045 non-null	object
22	CD_CID	131250 non-null	object
23	DS_INDICACAO_CLINICA	179944 non-null	object
24	DS_TIPO_ITEM	227122 non-null	object
25	CD_ITEM	227122 non-null	int64
26	DS_ITEM	227122 non-null	object
27	DS_CLASSE	227122 non-null	object
28	DS_SUBGRUPO	227122 non-null	object
29	DS_GRUPO	227122 non-null	object
30	QT_SOLICITADA	227122 non-null	float64
31	DS_STATUS_ITEM	227122 non-null	object

2 Pré-processamentos realizados

Separadas as colunas numéricas das categóricas, os dados numéricos foram normalizados utilizando o sklearn.preprocessing.StandardScaler enquanto as colunas categóricas foram codificadas utilizando o sklearn.preprocessing.OneHotEncoder.

3 Configuração experimental

A línguagem utilizada é Python na sua versão 3.7.12 e as bibliotecas utilizadas são:

- pandas
- sklearn
- numpy

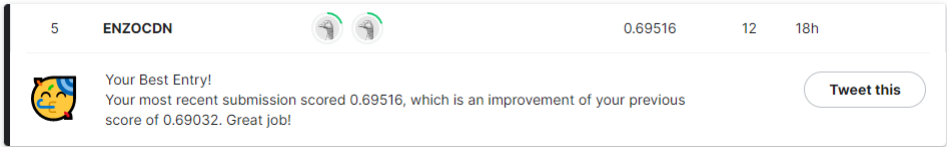
No classificador RandomForestClassifier foram utilizados os parâmetros max_depth=20, max_features=10, min_samples_leaf=3 obtidos por GridSearch.

4 Algoritmos utilizados

Inicialmente, o classificador escolhido foi o Naive Bayes, pos ele teoricamente não necessita de dados numéricos, mas não foi possível implementar a versão CategoricalNB e então foi utilizado então GaussianNB, que supõe parâmetros numéricos com uma distribuição constante.

Posteriormente foi decidido utilizar o classificador RandomForest, pois é um modelo robusto e mais resistente a valores ausentes.

5 Resultados



6 Referências bibliográficas

[1] 6.9.2. Label encoding. Scikit learn. Disponível em: https://scikit-learn.org/stable/modules/preprocessing_ta targets
[2] 1.9. Naive Bayes. Scikit learn. Disponível em: https://scikit-learn.org/stable/modules/naive_bayes.html
[3] curious-attempt-bunny. How to handle categorical data in scikit with pandas. Github. Disponível em: https://github.com/curious-attempt-bunny/ml-info/blob/master/How%20to%20handle%20categorical%20data%20in%20scikit%20with%20pandas.ipynb
[4] Pandas Tutorial. W3 Schools. Disponível em: https://www.w3schools.com/python/pandas/default.asp
[5] 1.13.2. Univariate feature selection. Scikit learn. Disponível em: https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection