UNIVERSIDADE DE ITAÚNA

Curso de Graduação em Ciência da Computação

Paula Luiza Corrêa da Silva Santos

UTILIZAÇÃO DE MACHINE LEARNING PARA PREDIÇÃO DE VAGAS DE TRABALHO FALSAS

Itaúna/MG

2022

Paula Luiza Corrêa da Silva Santos

UTILIZAÇÃO DE MACHINE LEARNING PARA PREDIÇÃO DE VAGAS DE TRABALHO FALSAS

Itaúna/MG

2022

SUMÁRIO

1. Introdução

1.1. Contextualização

De acordo com uma pesquisa feita em 2022 pelo Instituto Brasileiro de Geografia e Estatística (IBGE), cerca de 12 milhões de pessoas estão desempregadas no Brasil. Diante desse cenário, muitas estão suscetíveis a caírem no golpe de falsas vagas com a esperança de se reinserirem no mercado de trabalho. Uma das estratégias utilizadas pelos golpistas é a oferta de vagas aparentemente muito vantajosas para a pessoa trabalhadora.

Por esse motivo, é de suma importância identificar fatores que contribuem na probabilidade da vaga de trabalho ser falsa.

1.2. O problema proposto

Nem sempre é fácil perceber que algumas ofertas de emprego são, na verdade, vagas falsas que servem de isca para atrair pessoas para possíveis golpes. De setembro de 2012 a fevereiro de 2022, foram ofertadas cerca de 608 mil vagas falsas para entrevista de emprego, uma média de 4 mil “golpes” por dia. Esses golpes existem por um simples motivo: eles dão certo.

Há anos, hackers e golpistas com um bom conhecimento de programação para dispositivos móveis contam com um ambiente favorável à prática desse tipo de fraude, e o número de tentativas e de vítimas só aumenta.

1.3. Objetivos

Com o presente trabalho, o modelo desenvolvido será capaz de validar se uma vaga ofertada é falsa ou verdadeira, a fim de auxiliar os candidatos, otimizando seu tempo na busca por um emprego.

2. Coleta de Dados

O dataset utilizado foi obtido através da plataforma Kaggle (https://www.kaggle.com) e é formado por um conjunto de dados que contém cerca de 18 mil descrições de cargos, das quais cerca de 800 são falsas. Os dados consistem em informações textuais e meta-informações sobre os trabalhos, todos na língua inglesa.

A seguir encontra-se uma tabela contendo as informações a respeito das colunas contidas no *dataset*.

Nome da coluna/campo	Descrição	Tipo
job_id	Identificador dos trabalhos	int64
title	Título da entrada do anúncio de emprego.	object
location	Localização geográfica do anúncio de emprego.	object
department	Departamento corporativo.	object
salary_range	Faixa salarial indicativa.	object
company_profile	Breve descrição da empresa.	object
description	Descrição detalhada do anúncio de emprego.	object
requirements	Requisitos para a vaga de emprego.	object
benefits	Recrutou benefícios oferecidos pelo empregador.	object
telecommuting	Verdadeiro para posições de teletrabalho.	int64
has_company_logo	Verdadeiro se o logotipo da empresa estiver presente.	int64
has_questions	Verdadeiro se houver perguntas de triagem.	int64
employment_type	Se é tempo integral, meio período, contrato, etc.	object
required_experience	Executivo, nível de entrada, estagiário, etc.	object
required_education	Doutorado, Mestre, Bacharel, etc.	object
industry	Automotivo, TI, Saúde, Imobiliário, etc.	object

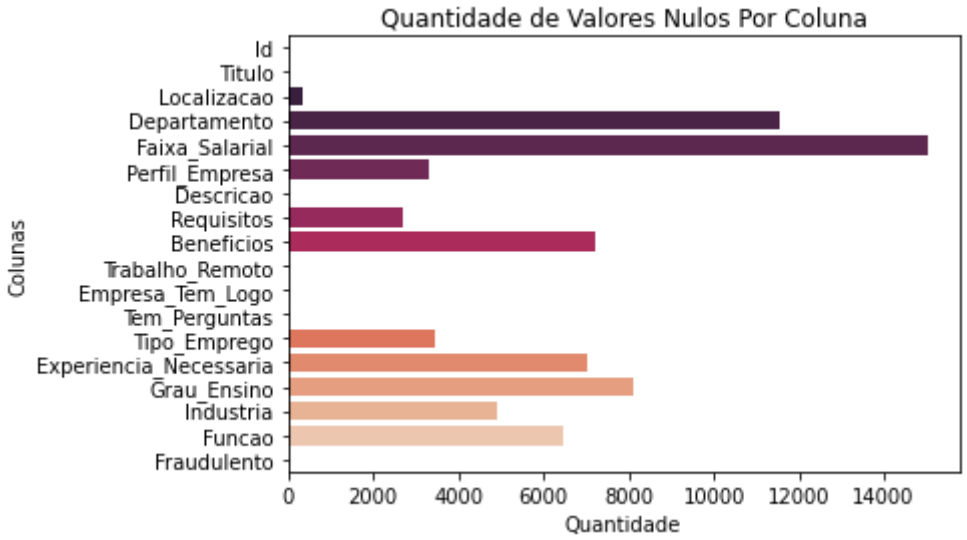
function	Consultoria, Engenharia, Pesquisa, Vendas etc.	object
fraudulent	atributo de classificação (target).	int64

3. Processamento/Tratamento de Dados

O primeiro tratamento realizado foi renomear as colunas, que estavam em inglês para português, para uma melhor compreensão dos dados, da seguinte forma:

Coluna Original	Coluna Nova	Coluna Original	Coluna Nova
job_id	Id	telecommuting	Trabalho_Remoto
title	Titulo	has_company_logo	Empresa_Tem_Logo
location	Localizacao	has_questions	Tem_Perguntas
department	Departamento	employment_type	Tipo_Emprego
salary_range	Faixa_Salarial	required_experiencie	Experiencia_Necessaria
company_profile	Perfil_Empresa	required_education	Grau_Ensino
description	Descricao	industry	Industria
requirements	Requisitos	function	Funcao
benefits	Beneficios	fraudulent	Fraudulento

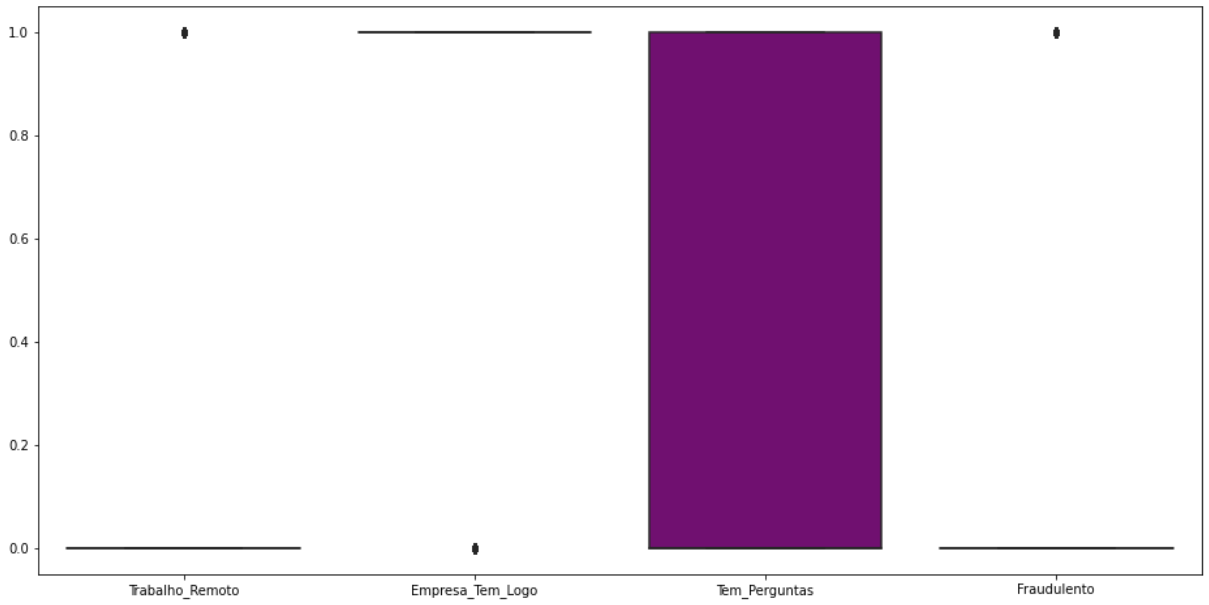
Considerando a base, foi encontrada a seguinte quantidade de valores nulos para cada coluna:



Pode-se observar que as colunas que possuem a maior quantidade de valores nulos são as colunas *Departamento* e *Faixa_Salarial*, onde os dados nulos representam 65% e 84%,

do valor total da coluna, respectivamente. Devido a essa grande quantidade, estas duas colunas foram removidas. Já os valores nulos contidos nas colunas *Beneficios, Requisitos, Descricao, Grau_Ensino* e *Perfil_Empresa* foram substituídos pelas palavras "**Desconhecido"** ou **"Desconhecida"**, pois estas colunas serão concatenadas. Os demais registros com valores nulos foram deletados.

Nas colunas *Trabalho_Remoto, Empresa_Tem_Logo* e *Fraudulento* foram detectados outliers, conforme a imagem abaixo, que foram removidos.



Foi criada uma nova coluna, denominada *Vaga,* composta pela concatenação das colunas *Titulo, Localizacao, Perfil_Empresa, Descricao, Requisitos* e *Beneficios.* Após a concatenação, estas colunas foram deletadas. Após feito isso, foram encontrados e removidos 25 registros duplicados.

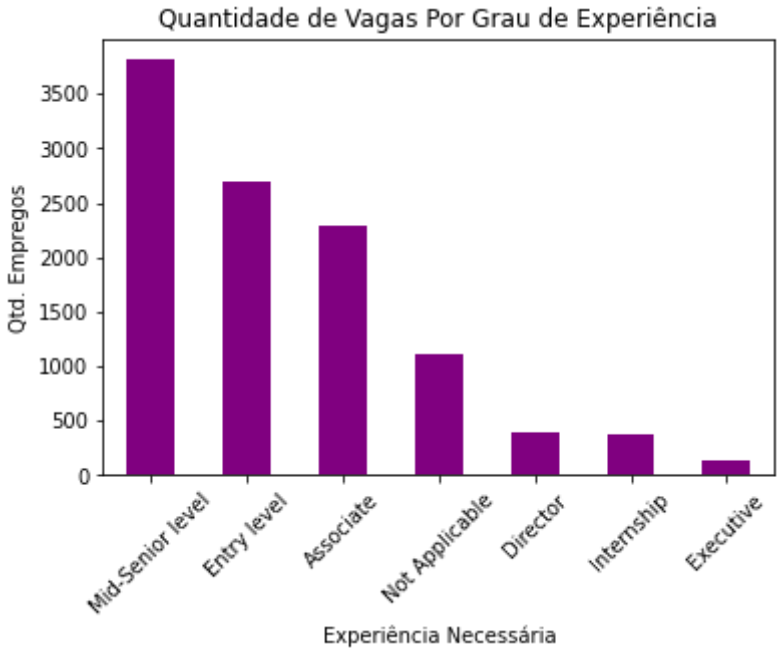
Fazendo um comparativo da base de dados, inicialmente esta possuía 17.880 linhas e 18 colunas. Logo após os tratamentos acima, a nova base passou a conter 1.115 linhas e 5 colunas. Destas 5 colunas restantes, apenas a coluna *Vaga* era do tipo object. Para que adiante fosse aplicado o modelo, foi realizado então o **LabelEncoder** desta coluna, mantendo assim todos os dados com o mesmo tipo.

4. Análise e Exploração dos Dados

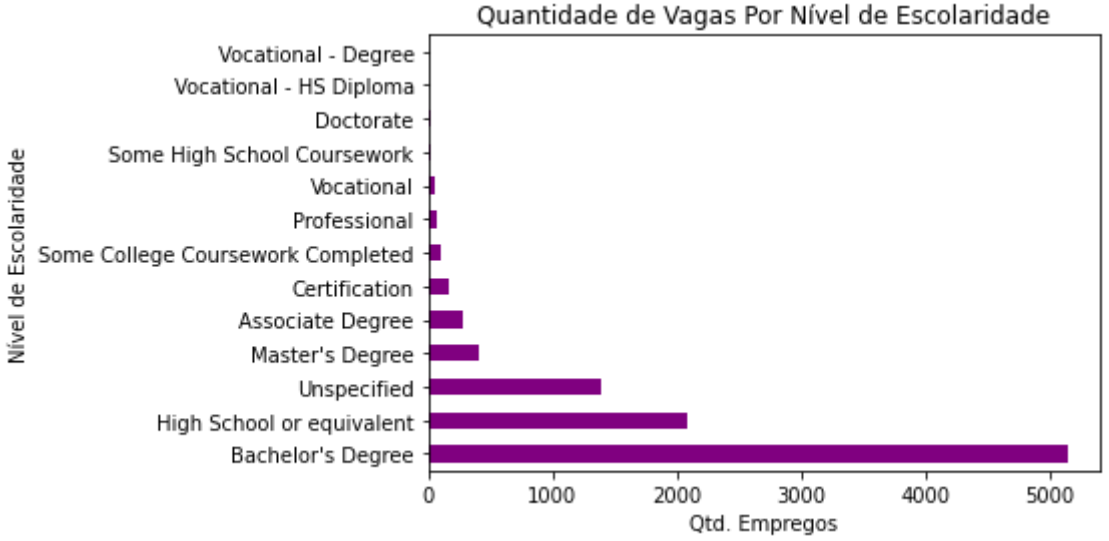
Foi realizada uma parte da análise dos dados antes da criação da coluna Vaga, analisando todas colunas, e outra após a concatenação.

Primeiramente foram feitas análises univariadas, bivariadas e multivariadas. Nestas pudemos analisar algumas das seguintes características:

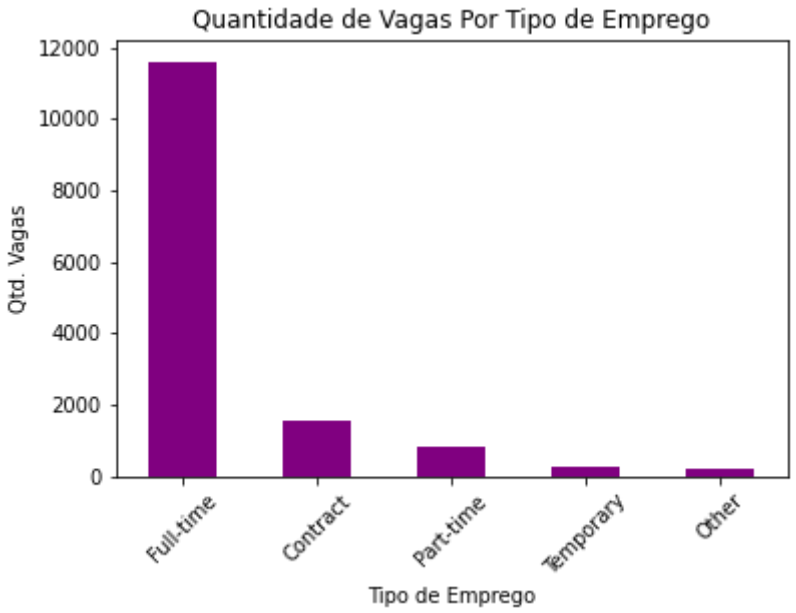
- A maioria das vagas é ofertada para trabalhador de nível Médio/Sênior



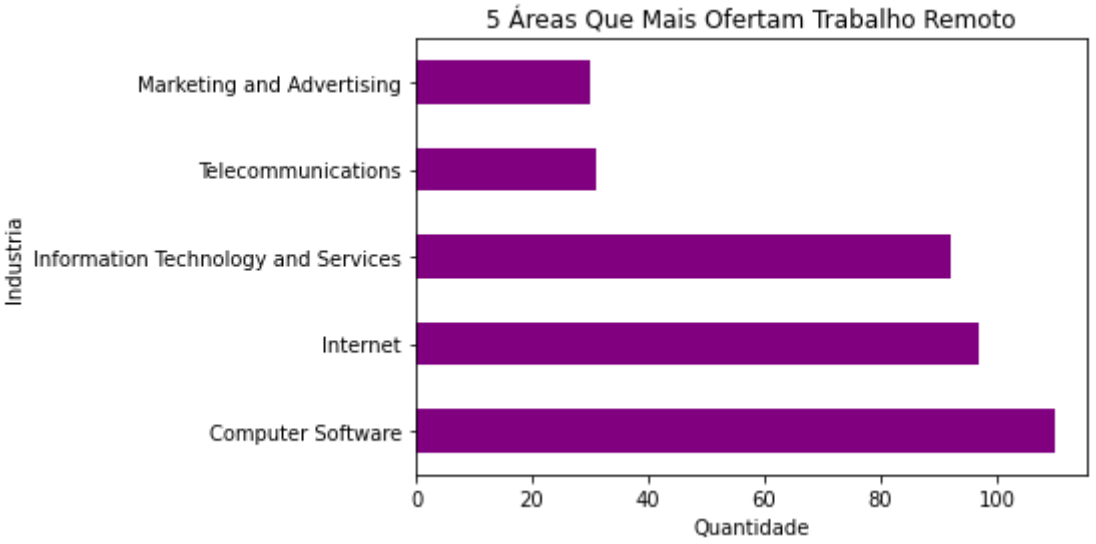
- Já o nível de escolaridade mais exigido é o bacharelado



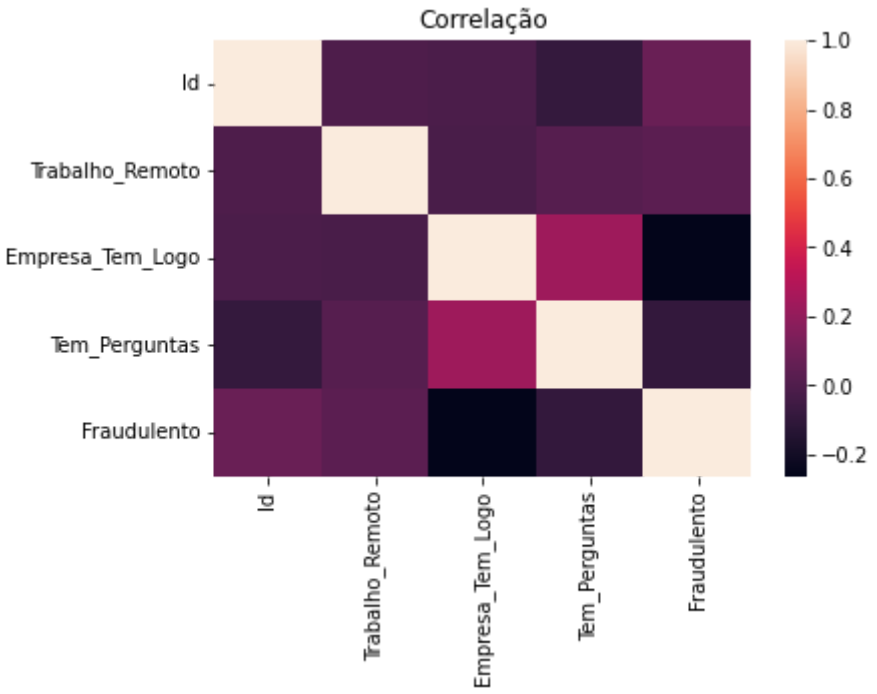
- Na base de estudo, quase 12.000 vagas são oferecidas para trabalho em tempo integral e para as outras modalidades cada uma possui menos de 2.000 vagas disponíveis.



- As vagas de trabalho remoto são concentradas em sua maioria nas indústrias de Software, Internet e T.I.

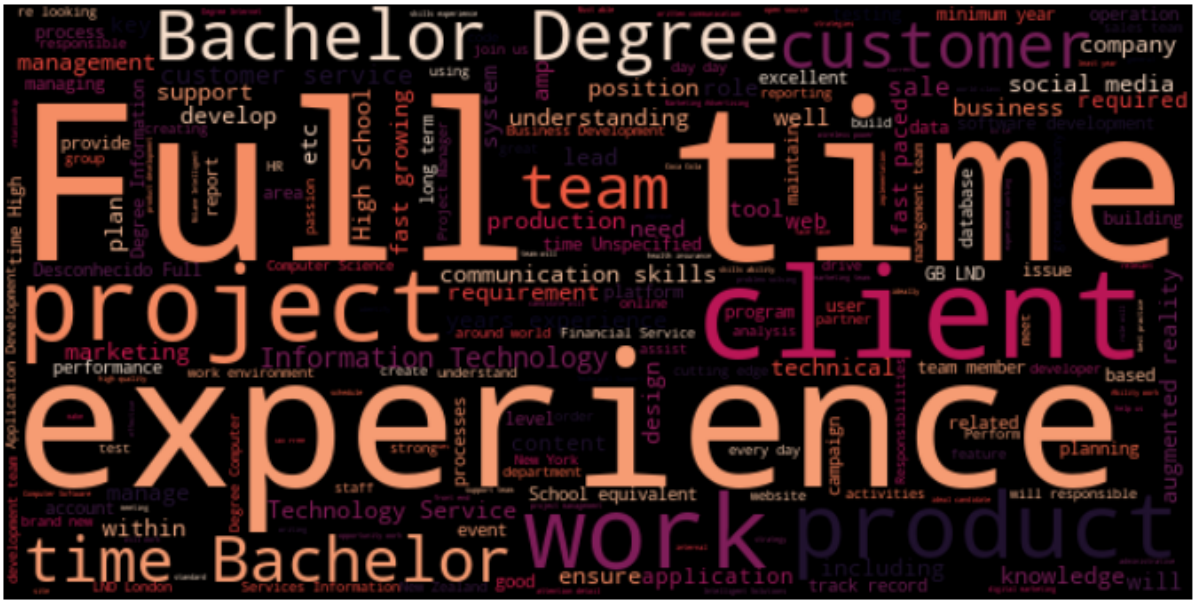


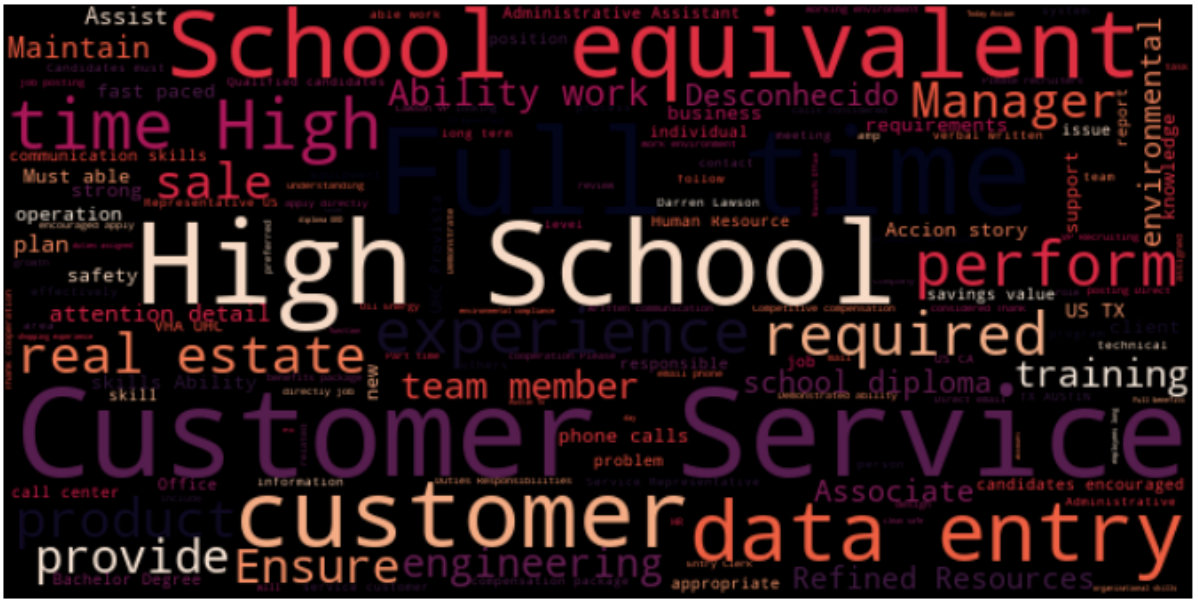
- Verificou-se também que a indústria com o maior número de vagas falsas foi a de "Óleo e Energia", com 109 fraudes.
- Foi feita a correlação das colunas numéricas e verificou-se que elas não possuem quase nenhuma correlação:



- Foi calculada a proporção da target *Fraudulento* e foi visto que a quantidade de vagas falsas corresponde a apenas 4,8% da quantidade de registros totais, sendo necessário balancear esses dados.

Já após a concatenação das colunas e da geração da *Vaga*, foi feita uma análise nos textos dessa coluna e verificadas as palavras que aparecem com maior frequência. Estas palavras são apresentadas nas imagens a seguir, encontradas nas vagas reais e falsas, respectivamente:





Após feitas essas análises, foi realizado o balanceamento dos dados, para dar início a construção dos modelos. Para isso, foi utilizada a técnica denominada **Random Under Samplin,** que consiste basicamente em remover os dados para ter um conjunto de dados mais balanceado e assim evitar o overfitting dos modelos. A primeira coisa a se fazer foi determinar o quão desequilibrada era a classe. Depois de determinadas quantas instâncias foram consideradas fraudulentas, foi necessário trazer as não fraudulentas para a mesma proporção (50/50). O próximo passo então foi embaralhar os dados para ver se os modelos podem manter uma certa precisão toda vez que executar esse script.

Ao final, a base de dados que continha 1062 vagas verdadeiras e 53 vagas falsas, passou a conter 110 registros verídicos e 110 registros falsos, totalizando 220 registros.

5. Criação de Modelos de Machine Learning

Nesta etapa, foram utilizados três tipos de classificadores, observando sua capacidade na resolução do problema proposto, a fim de descobrir qual deles seria mais eficaz na detecção de vagas fraudulentas. Os classificadores são:

1. Árvore de Decisão (Decision Tree): com critério gini
2. KNN (K Nearest Neighbours): com k = 3
3. Floresta Aleatória (Random Forest)

Para a separação dos dados em treinamento e teste, foi utilizada a proporção de 70% para treino e 30% para teste, resultando em 154 e 66 registros respectivamente. Abaixo temos o código em Python utilizado para esta divisão:

```
x = pd.DataFrame(df_Novo[["Trabalho_Remoto",  "Empresa_Tem_Logo", "Tem_Perguntas", "Vaga"]])
y = pd.DataFrame(df_Novo["Fraudulento"])

#Divide o dataframe em treinamento e teste
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 11)

#Verifica os tamanhos
print("Dados de treinamento: ", y_train.shape, x_train.shape)
print("Dados de teste: ", y_test.shape, x_test.shape)
```

```
Dados de treinamento:  (154, 1) (154, 4)
Dados de teste:  (66, 1) (66, 4)
```

Em todos os três classificadores foram utilizadas as quatro métricas de avaliação (Acurácia, Precisão, Recall e F1 Score).

Para implementação dos modelos foi utilizada a biblioteca **Sklearn**, e foi utilizado um padrão no desenvolvimento destes, contendo o treinamento e o teste dos mesmos, em seguida a apresentação do relatório da classificação e a matriz de confusão dos modelos (tabela que indica os erros e acertos do modelo). Abaixo segue o modelo de Árvore de

Decisão:

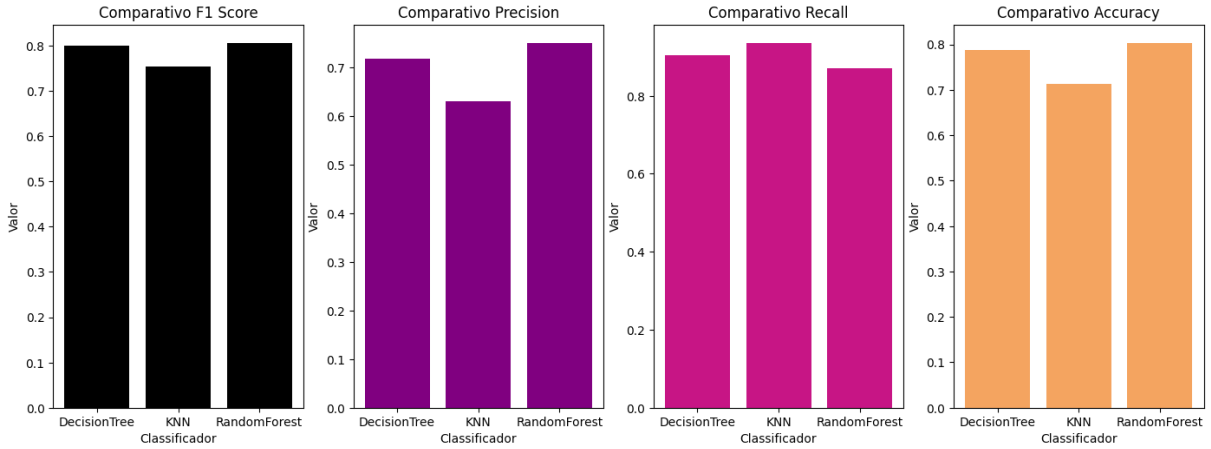
```
#treinamento e teste
x_trainDT = x_train
x_testDT = x_test
y_trainDT = y_train
y_testDT = y_test

#Treina o modelo
dT =  tree.DecisionTreeClassifier()
dT.fit(x_trainDT, y_trainDT)

#Testa o modelo
dTPred = dT.predict(x_testDT)

#Apresenta o classification_report
print(classification_report(y_testDT, dTPred))

#Plota a matriz de confusão
sns.heatmap(confusion_matrix(y_testDT, dTPred),
            annot = True, fmt=".0f", annot_kws={"size": 18}, cmap = "rocket")
```

6. Interpretação dos Resultados

Com base nos modelos de classificação utilizados (*Decision Tree*, KNN e *Random Forest*), observou-se que os modelos *Decision Tree* e *Random Forest* obtiveram resultados semelhantes entre si. Porém, de maneira geral, o *Random Forest* apresentou os melhores resultados, com acurácia, precisão e f1 score superiores aos demais modelos.

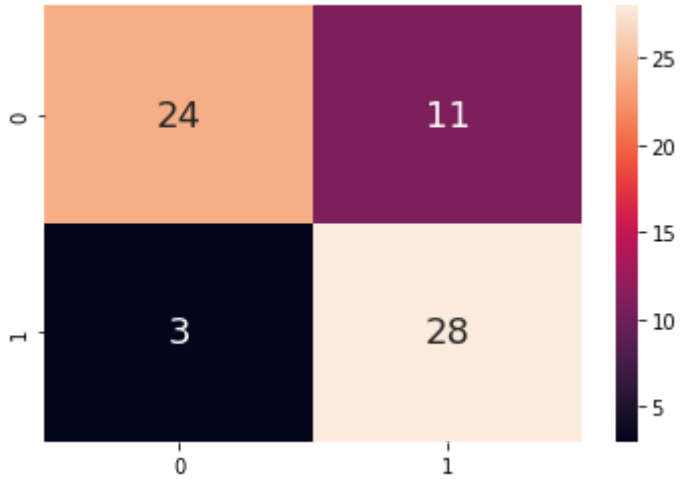
7. Apresentação dos Resultados

De maneira geral, como dito anteriormente, o modelo *Random Forest* apresentou o melhor desempenho na predição de vagas de trabalho falsas. Abaixo estão os comparativos entre as métricas utilizadas em cada modelo, onde foi possível verificar a eficiência de cada um e feita a verificação do melhor modelo:

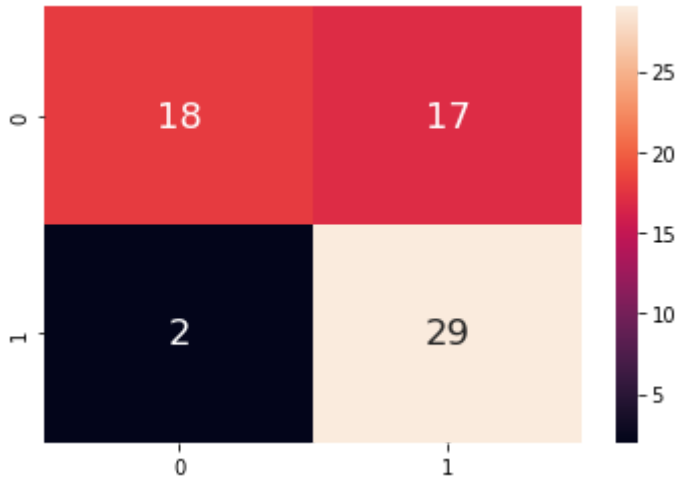


De maneira individual, consegue-se perceber uma boa aplicação em cada um dos modelos. Para melhor verificação desta afirmação, abaixo serão apresentadas as matrizes de confusão de cada um deles:

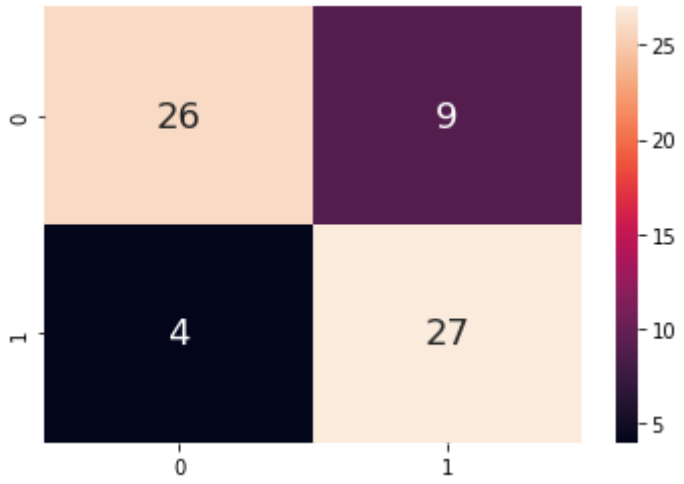
1. *Decision Tree*



2. KNN



3. *Random Forest*



8. Links

Link para o apresentação no Canva:

https://www.canva.com/design/DAFFF5txRn8/k0W5qiE4xogdxhOU3y3W7g/view?utm_content=DAFFF5txRn8&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

Link para o projeto:

https://colab.research.google.com/drive/1pouzhiMwLZScQCL7dwMC1_GfCp1MfusQ?usp=sharing

Link para a base de dados:

https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction

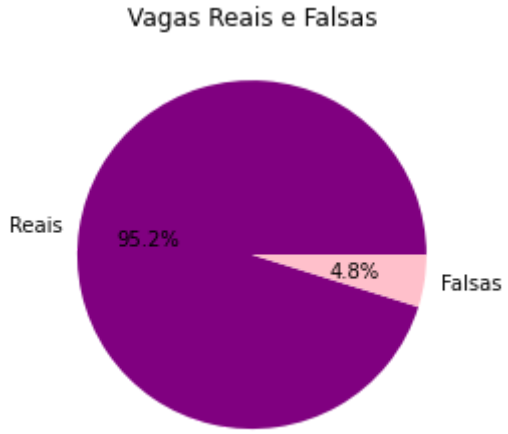
REFERÊNCIAS

- RODRIGUES, Vitor. Métricas de Avaliação: acurácia, precisão, recall… quais as diferenças?. **Medium**. Disponível em : <https://vitorborbarodrigues.medium.com/métricas-de-avaliação-acurácia-precisão-recall-quais-as-diferenças-c8f05e0a513c>**.** Acesso em 25 de junho de 2022.

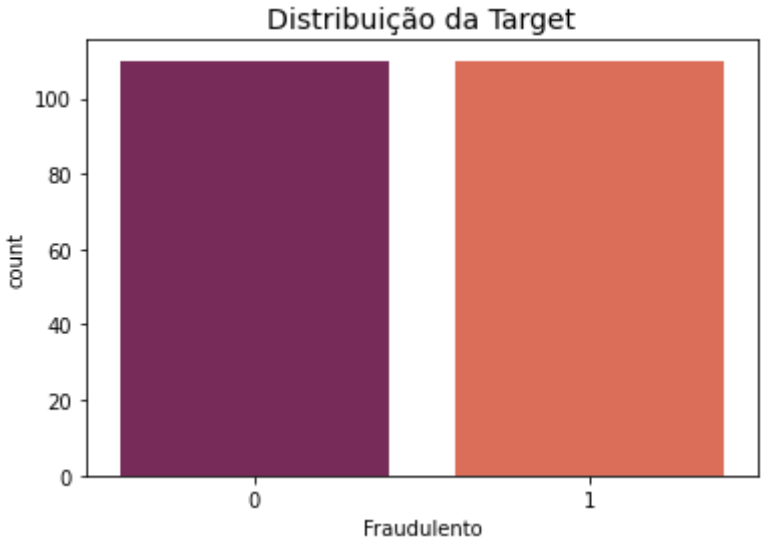
APÊNDICE

Gráficos

- Distribuição da target antes do balanceamento



- Distribuição da target após o balanceamento



Tabelas

- Quantidade de valores nulos na base de dados

	Qtd_Nulos	%Valores_Nulos
Localizacao	346	2.0
Departamento	11547	65.0
Faixa_Salarial	15012	84.0
Perfil_Empresa	3308	19.0
Descricao	1	0.0
Requisitos	2695	15.0
Beneficios	7210	40.0
Tipo_Emprego	3471	19.0
Experiencia_Necessaria	7050	39.0
Grau_Ensino	8105	45.0
Industria	4903	27.0
Funcao	6455	36.0