Curso: Ciência de Dados
Semestre: 2° semestre de 2023
Componente Curricular/Tema: Projeto Aplicado II
Nome Completo dos Aluno: Thainá Vieira dos Santos TIA: 22500081 Vinicius Caumo Segatto TIA: 22506861 Leonardo dos Reis Olher TIA: 22510249 Nicolas Pinotti TIA: 22514112 Vinícius Vieira da Cunha Oliveira TIA: 22505865
Nome dos Professores: Anderson Adaime de Borba

1) Aplicando Conhecimento

Definição do método analítico.

O grupo deve fazer uma descrição detalhada do que pretende usar de metodologia no projeto.

Portanto, nesta etapa deve ser entregue:

- Definição da linguagem de programação usada no projeto.
- Análise exploratória da base de dados escolhida.
- Tratamento da base de dados (Preparação e treinamento).
- Definição e descrição das bases teóricas dos métodos.
- Definição e descrição de como será calculada a acurácia.

SUMÁRIO

1. APRESENTAÇÃO DO PROJETO.....

1.1. DEFINIÇÃO DA ORGANIZAÇÃO ESCOLHIDA.....

1.2. ÁREA DE ATUAÇÃO.....

1.3. DADOS UTILIZADOS.....

2. OBJETIVOS E METAS.....

3. CRONOGRAMA DE ATIVIDADES.....

4. DEFINIÇÃO DO MÉTODO ANALÍTICO.....

4.1. DEFINIÇÃO DA LINGUAGEM DE PROGRAMAÇÃO.....

4.2. ANÁLISE EXPLORATÓRIA DE DADOS.....

4.3. TRATAMENTO DA BASE DE DADOS.....

4.4. DEFINIÇÃO E DESCRIÇÃO DAS BASES TEÓRICAS DOS MÉTODOS.....

5. REFERÊNCIAS BIBLIOGRÁFICAS.....

1) APRESENTAÇÃO DO PROJETO

1.1. DEFINIÇÃO DA ORGANIZAÇÃO ESCOLHIDA

Nome da Empresa: DataTrend Insights
Missão: Nossa missão é fornecer análises de dados de alta qualidade para empresas que desejam tomar decisões estratégicas com base nas tendências salariais em Ciência de Dados 2023.
Visão: Tornar-se a principal fonte de insights e recomendações de remuneração em Ciência de Dados, ajudando empresas a atrair, reter e motivar os melhores talentos.

DataTrend Insights está enfrentando um problema interno relacionado à retenção de Analistas de Dados Júnior. Nossa empresa está perdendo talentos promissores para concorrentes devido a diferenças salariais e benefícios. Precisamos identificar a faixa salarial competitiva para Analistas de Dados Júnior em 2023 e criar uma estratégia de retenção eficaz.

1.2. AREA DE ATUAÇÃO

A empresa "DataTrend Insights" atua principalmente nos ramos de consultoria em remuneração oferecendo serviços de consultoria para outras empresas, ajudando-as a compreender e ajustar suas políticas de remuneração em relação às tendências salariais em Ciência de Dados, envolvendo análises detalhadas do mercado, identificação de faixas salariais competitivas e recomendações específicas. E no ramo soluções de software, onde desenvolve e vende soluções de software especializadas para a gestão de remuneração em Ciência de Dados. Isso incluiria ferramentas de análise de dados, modelagem de salários e previsão de tendências salariais.

1.3. DADOS UTILIZADOS

A DataTrend Insights baseia suas análises de tendências salariais em Ciência de Dados em 2023 em um conjunto de dados cuidadosamente selecionado e abrangente. Este conjunto de dados serve como a espinha dorsal de nossos esforços analíticos, fornecendo informações essenciais para compreender a dinâmica salarial na indústria de Ciência de Dados. Abaixo estão os principais componentes desse conjunto de dados:

- Ano de Trabalho: Esta coluna representa o ano específico da coleta de dados salariais.
- Nível de Experiência: Os funcionários são categorizados de acordo com seu nível de experiência, incluindo iniciantes, experientes, de nível médio e sêniores.
- Tipo de Emprego: Cada profissional é rotulado com seu tipo de emprego, que pode ser tempo integral, contratado, freelancer ou meio período.
- Cargo: Registramos os cargos dos funcionários, abrangendo uma variedade de títulos, como "Cientista Aplicado" e "Analista de Qualidade de Dados".
- Salário: Esta coluna contém os valores salariais, expressos em suas respectivas moedas locais.
- Moeda do Salário: Indica o código da moeda que representa o salário em questão.
- Salário em USD: Todos os salários foram convertidos para dólares americanos (USD) para permitir uma comparação uniforme.
- Localização da Empresa: Esta coluna especifica a localização das empresas, identificadas por códigos de país, como "US" para Estados Unidos e "NG" para Nigéria.
- Tamanho da Empresa: As empresas são classificadas em categorias de tamanho, que incluem grande, média e pequena.

Esses dados, cuidadosamente coletados e preparados, servirão como a base para nossas análises estatísticas avançadas, modelagem preditiva e visualizações de dados. Com esses insights, nossa equipe pode ajudar empresas a tomar decisões estratégicas informadas sobre remuneração, retenção de talentos e estratégias de aquisição de talentos em Ciência de Dados.

A integridade, qualidade e relevância desses dados são fundamentais para garantir que nossas análises sejam precisas e confiáveis, permitindo-nos fornecer serviços de consultoria e soluções personalizadas de alta qualidade para nossos clientes.

2) OBJETIVOS E METAS

Objetivo	Meta
Identificar a faixa salarial competitiva para Analistas de Dados Júnior em 2023	Determinar uma faixa salarial competitiva com base em dados coletados e análise estatística até o final do primeiro trimestre de 2023.

Criar um conjunto de recomendações para a empresa	Fornecer um conjunto de recomendações específicas para a empresa com base nas descobertas, incluindo ajustes salariais e estratégias de retenção, até o final do segundo trimestre de 2023.
Visualizar e interpretar tendências salariais em Ciência de Dados de 2021 a 2023	Desenvolver visualizações de dados interativas que mostrem as tendências salariais ao longo de três anos e interpretar os resultados até o final do terceiro trimestre de 2023.
Avaliar a eficácia das recomendações	Realizar uma análise de acompanhamento no quarto trimestre de 2023 para avaliar como as recomendações implementadas afetaram a retenção de Analistas de Dados Júnior e a competitividade salarial da empresa.
Publicar um relatório de tendências salariais	Lançar um relatório detalhado e informativo sobre as tendências salariais em Ciência de Dados em 2023 até o final do ano de 2023.
Oferecer treinamento sobre gestão de remuneração em Ciência de Dados	Desenvolver e lançar um programa de treinamento em gestão de remuneração voltado para profissionais de RH e gerentes de contratação no primeiro trimestre de 2024.
Expandir os serviços de consultoria em remuneração	Ampliar o portfólio de serviços de consultoria em remuneração para outras indústrias e setores até o final do segundo trimestre de 2024.

Esses objetivos e metas são formulados para orientar as atividades da empresa ao longo do projeto de análise de tendências salariais, garantindo que os resultados sejam alcançados de maneira eficaz e dentro do prazo estabelecido.

3) CRONOGRAMA DE ATIVIDADES

Fase 1: Preparação do Projeto	• Definir os objetivos e escopo do projeto. • Montar a equipe de projeto e designar responsabilidades. • Estabelecer as metas de qualidade e critérios de sucesso. • Preparar uma lista de verificação de coleta de dados.
Fase 2: Coleta de Dados 11/09	• Identificar as fontes de dados relevantes. • Realizar a coleta de dados das fontes identificadas. • Limpar e pré-processar os dados brutos para remover valores ausentes e erros. • Validar a qualidade dos dados coletados.
Fase 3: Manipulação de Dados com Python 15/10	• Importar os dados coletados para um ambiente de análise, como Jupyter Notebook. • Realizar a manipulação inicial de dados usando bibliotecas Python, como Pandas. • Executar transformações de dados, como filtragem, agregação e criação de novas variáveis. • Verificar a integridade e consistência dos dados manipulados.
Fase 4: Manipulação de Dados com R 15/10	• Importar os dados manipulados anteriormente para um ambiente de análise R, como RStudio. • Realizar manipulações de dados adicionais usando pacotes R, como dplyr e tidyr.

	• Aplicar análises estatísticas e modelagem de dados usando bibliotecas R relevantes. • Criar visualizações de dados usando ggplot2 ou outras bibliotecas de visualização.
Fase 5: Análise de Dados e Relatórios 01/11	• Realizar análises estatísticas avançadas usando Python e R, conforme necessário. • Interpretar os resultados das análises. • Preparar relatórios de análise de dados que incluam visualizações, gráficos e conclusões. • Revisar e validar as conclusões com a equipe e partes interessadas.
Fase 7: Documentação e Apresentação 01/11	• Documentar todo o processo de análise de dados, desde a coleta até a implementação de recomendações. • Preparar uma apresentação executiva para compartilhar com partes interessadas e a equipe de gerenciamento.
Fase 8: Encerramento do Projeto 29/11	• Avaliar o sucesso do projeto em relação às metas estabelecidas. • Realizar uma revisão pós-projeto para identificar lições aprendidas. • Arquivar todos os dados, códigos e documentação relevantes.

4) DEFINIÇÃO DO MÉTODO ANALÍTICO

4.1 DEFINIÇÃO DA LINGUAGEM DE PROGRAMAÇÃO

Após cuidadosa consideração, a equipe de desenvolvedores da DataTrend Insights optou por conduzir a análise exploratória do dataset em Python, em vez da linguagem R.

Essa decisão foi tomada com base em vários benefícios que o Python oferece para a nossa equipe e para o projeto como um todo.

Um dos principais benefícios de escolher Python para a AED é a versatilidade da linguagem. Python é amplamente conhecido por sua capacidade de integração com várias bibliotecas de análise de dados, machine learning e visualização. Isso nos permite criar um fluxo de trabalho contínuo, onde podemos realizar a análise exploratória, a preparação de dados e a modelagem preditiva, tudo em um único ambiente.

Além disso, Python é uma linguagem de programação de propósito geral, o que significa que muitos de nossos desenvolvedores já têm experiência com ela. Isso facilita a colaboração e o compartilhamento de código entre a equipe. Também permite que nossos desenvolvedores usem suas habilidades de programação Python para automatizar tarefas e criar pipelines de análise de dados personalizados.

Outro benefício importante é a vasta comunidade Python e a disponibilidade de recursos educacionais. Isso significa que podemos encontrar suporte, soluções para desafios técnicos e documentação facilmente. Além disso, as bibliotecas de Python, como Pandas, são bem documentadas e oferecem ampla funcionalidade para análise de dados e modelagem preditiva.

Eu destaco um cuidado crucial na abordagem deste projeto: a análise da base de dados não pode se limitar apenas ao salário, pois o salário não é uma variável textual, mas sim uma variável numérica. Na realidade, o salário pode ser considerado uma variável contínua.

O sucesso da nossa análise depende da consideração de variáveis textuais, como o nível de experiência (junior, senior), a moeda (BRL, USD) e outras variáveis textuais presentes na base de dados. Essas variáveis desempenham um papel fundamental na compreensão das dinâmicas salariais em Ciência de Dados.

Ao considerar essas variáveis textuais, seremos capazes de explorar relacionamentos e tendências significativas que podem influenciar os salários, como a diferença salarial entre Analistas de Dados Júnior e Sênior, a variação de salários em diferentes moedas, e outros fatores que podem ser cruciais para a nossa estratégia de retenção de talentos.

4.2 ANÁLISE EXPLORATÓRIA DE DADOS

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

```
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

df = pd.read_csv('/kaggle/input/data-science-salary-2021-to-2023/Data
Science Salary 2021 to 2023.csv')
```

- NumPy (import numpy as np):

O NumPy é uma biblioteca fundamental para cálculos numéricos em Python. Ele fornece estruturas de dados eficientes, como arrays multidimensionais (numpy arrays), que são usados para armazenar e manipular dados numéricos. O NumPy é especialmente útil para realizar operações matemáticas em larga escala em arrays, como cálculos estatísticos e operações de álgebra linear.

- Pandas (import pandas as pd):

O Pandas é uma biblioteca essencial para a manipulação e análise de dados tabulares. Ele oferece as estruturas de dados DataFrame e Series, que facilitam a importação, limpeza, transformação e análise de dados. O Pandas é ideal para carregar e explorar conjuntos de dados, executar operações de filtro, agrupamento e agregação, lidar com dados faltantes e preparar dados para análise estatística.

- Matplotlib (import matplotlib.pyplot as plt):

O Matplotlib é uma biblioteca de visualização de dados que oferece uma ampla variedade de opções para criar gráficos e visualizações personalizadas. É usado para criar gráficos de linhas, gráficos de dispersão, histogramas, gráficos de barras e muitos outros tipos de visualizações. É particularmente útil na fase de AED para ajudar a visualizar os dados e identificar tendências, padrões e anomalias.

- Seaborn (import seaborn as sns):

O Seaborn é uma biblioteca de visualização de dados baseada no Matplotlib, que simplifica a criação de gráficos estatísticos atraentes. Ele fornece funções de alto nível para criar gráficos estatísticos complexos, como boxplots, gráficos de violino, mapas de calor (heatmap) e muito mais. O Seaborn é uma ferramenta valiosa para explorar as relações entre variáveis e apresentar resultados de forma mais informativa.

- Warnings (import warnings):

4.3. TRATAMENTO DA BASE DE DADOS

A biblioteca Warnings é usada para controlar mensagens de aviso durante a execução do código. Ao usar warnings.filterwarnings('ignore'), você pode suprimir mensagens de aviso que não deseja ver durante a execução do código. Isso é útil para manter o ambiente de desenvolvimento mais limpo e focado na análise.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3761 entries, 0 to 3760
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   work_year         3761 non-null   int64 
 1   experience_level  3761 non-null   object
 2   employment_type   3761 non-null   object
 3   job_title         3761 non-null   object
 4   salary            3761 non-null   int64 
 5   salary_currency   3761 non-null   object
 6   salary_in_usd     3761 non-null   int64 
 7   company_location  3761 non-null   object
 8   company_size      3761 non-null   object
dtypes: int64(3), object(6)
memory usage: 264.6+ KB
```

DataFrame Info:

<class 'pandas.core.frame.DataFrame'> indica que o objeto criado é um DataFrame do Pandas, que é uma estrutura tabular para armazenar e manipular dados.

RangeIndex: 3761 entries, 0 to 3760 informa que o DataFrame possui um total de 3761 linhas (entradas) e um índice que vai de 0 a 3760. Isso sugere que você tem 3761 registros no conjunto de dados.

Data Columns:

As informações sob o cabeçalho "Data columns (total 9 columns)" fornecem uma visão geral das colunas presentes no DataFrame.

Cada coluna é listada com um número (0 a 8) e um nome (por exemplo, "work_year"). Isso indica que o DataFrame tem nove colunas no total.

Column Information:

As informações na tabela logo abaixo das colunas descrevem cada coluna individualmente, incluindo seu nome, a quantidade de valores não nulos (Non-Null Count) e o tipo de dado (Dtype).
Por exemplo:

"work_year" é uma coluna numérica (int64) com 3761 valores não nulos.

"experience_level" é uma coluna de texto (object) com 3761 valores não nulos.

"salary" é uma coluna numérica (int64) com 3761 valores não nulos.

É possível checar se alguma coluna possui valor nulo com a seguinte execução:

```
df.isnull().sum()
```

```
work_year           0
experience_level    0
employment_type     0
job_title           0
salary              0
salary_currency     0
salary_in_usd       0
company_location    0
company_size        0
dtype: int64
```

Memory Usage:

memory usage: 264.6+ KB indica o uso de memória do DataFrame, que é de aproximadamente 264.6 KB.

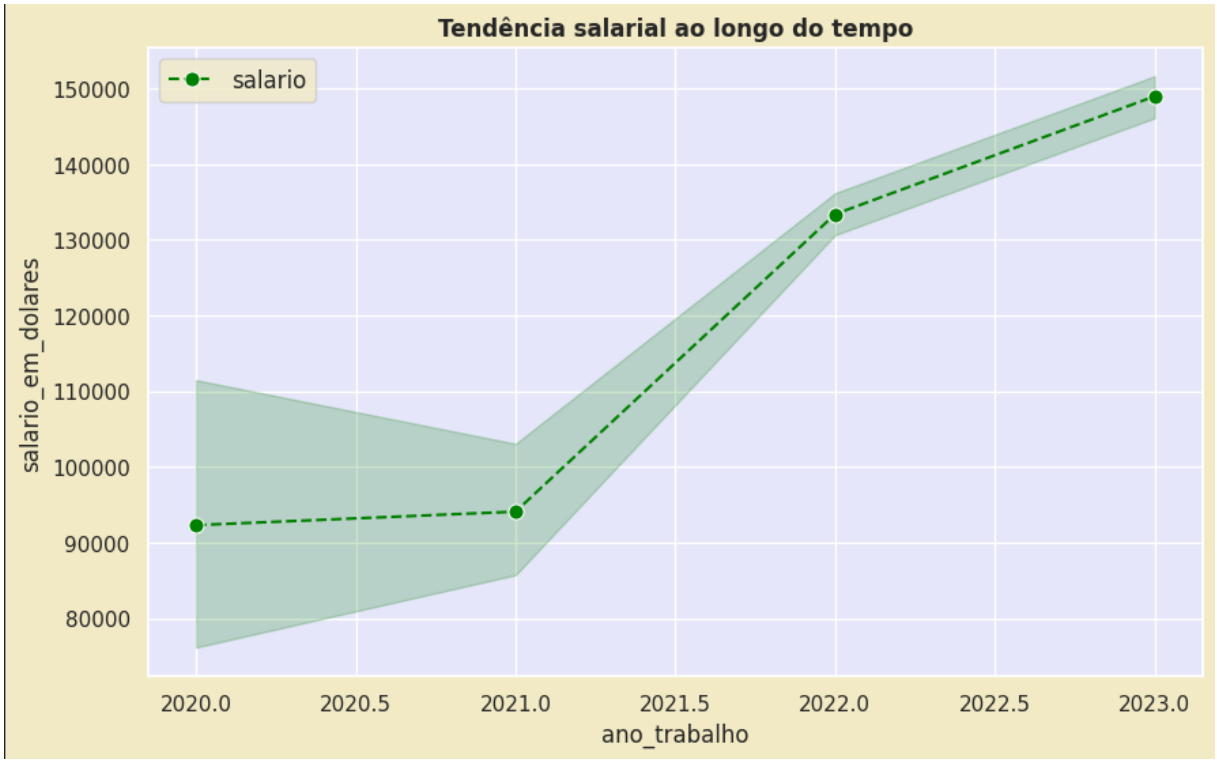
Essas informações iniciais fornecem uma visão geral importante do conjunto de dados que irá ser trabalhado pela equipe do DataTrends. Por exemplo, é perceptivel não há valores nulos em nenhuma das colunas, o que é um bom sinal para a qualidade dos dados. Além disso, agora você tem uma lista das colunas disponíveis e informações sobre os tipos de dados em cada uma delas.

Para facilitar o entendimento da equipe e a criação dos próximos passos, foi traduzido as colunas do dataset para uma melhor compreensão:

```
df['experience_level'] = df['experience_level'].replace('EN', 'Junior')
df['experience_level'] = df['experience_level'].replace('EX', 'Experiente')
df['experience_level'] = df['experience_level'].replace('MI', 'Nivel_medio')
df['experience_level'] = df['experience_level'].replace('SE', 'Senior')
```

```
df['employment_type'] = df['employment_type'].replace('FT', 'Full-Time')
df['employment_type'] = df['employment_type'].replace('CT', 'Contratante')
df['employment_type'] = df['employment_type'].replace('FL', 'Freelancer')
df['employment_type'] = df['employment_type'].replace('PT', 'Meio_periodo')

df['company_size'] = df['company_size'].replace('L', "Grande")
df['company_size'] = df['company_size'].replace('M', "Media")
df['company_size'] = df['company_size'].replace('S', "Pequena")

df = df.rename(columns={'work_year':
'ano_trabalho','experience_level':'nivel_experiencia',
'employment_type':'tipo_vaga'})
df = df.rename(columns ={'job_title':'titulo_emprego','salary':'salario',
'salary_currency': "moeda_pagamento"})
df = df.rename(columns={'salary_in_usd':'salario_em_dolares',
'company_location':'lugar_empresa', 'company_size': 'tamanho_empresa'})
```

```
plt.figure(figsize = (10,6))
salary_trend = df[['salario_em_dolares', 'ano_trabalho']].sort_values(by =
'ano_trabalho')
p = sns.lineplot(data =salary_trend ,x = 'ano_trabalho', y =
'salario_em_dolares', marker = 'o',linestyle='--', color='Green',
markersize=8 )
plt.title('Tendência salarial ao longo do tempo', fontsize=12,
fontweight='bold')


# Customize the background color
p.set_facecolor("#E6E6FA")
plt.legend(['salario'], loc='best', fontsize=12)


p.grid(True)


plt.show()
```



O comando executado cria um gráfico de linha que exibe a tendência salarial ao longo do tempo. Isso é útil para destacar como os salários na área de Ciência de Dados mudaram ao longo dos anos, fornecendo informações valiosas para sua análise. Os elementos do gráfico, como a cor, os marcadores, o estilo de linha e a legenda, são escolhidos para tornar a representação visualmente informativa e atraente.

Integração na Análise

Esse gráfico pode ser uma parte integrante da análise exploratória de dados. Ele permite que identificar tendências gerais nos salários, como aumentos ou quedas, e pode ser usado para destacar anos específicos de interesse. Isso pode ajudar a responder a perguntas relacionadas à evolução salarial em Ciência de Dados.

Comunicação de Resultados

Além de apoiar a análise, esse gráfico é útil para comunicar os resultados a outros membros da equipe ou partes interessadas. A representação visual é frequentemente mais eficaz na comunicação de tendências do que tabelas de dados.

4.5 DEFINIÇÃO E DESCRIÇÃO DAS BASES TEÓRICAS DOS MÉTODOS

Nossa escolha de bibliotecas e ferramentas é fundamentada nas seguintes considerações teóricas:

4.6 Python como Linguagem de Programação

Python é amplamente reconhecido como uma linguagem eficaz para análise de dados devido à sua vasta comunidade, bibliotecas robustas e capacidade de integração.

Pandas para Manipulação de Dados: O Pandas é uma biblioteca Python projetada para análise de dados tabulares. Sua estrutura de dados, o DataFrame, simplifica a manipulação de dados, tornando-o uma escolha lógica para nosso projeto.

Matplotlib e Seaborn para Visualização: O Matplotlib é uma biblioteca de visualização de dados versátil, enquanto o Seaborn fornece ferramentas adicionais para criar visualizações estatísticas atraentes. Ambos são escolhas sólidas para representar graficamente nossos resultados.

Adaptação às Necessidades do Projeto

A escolha de bibliotecas e ferramentas deve ser adaptada às necessidades específicas do projeto. Isso inclui considerar os tipos de dados que estamos lidando, os objetivos da análise e a eficiência na execução de tarefas específicas.

Documentação e Comunidade de Suporte

A disponibilidade de documentação detalhada e uma comunidade de suporte ativa são fatores importantes a serem considerados. Garantir que as bibliotecas e ferramentas selecionadas sejam bem documentadas e tenham uma base de usuários ativa facilita a resolução de problemas e a aprendizagem.

Revisão da Literatura

A revisão da literatura também desempenha um papel crucial na escolha de bibliotecas e ferramentas, pois permite avaliar a relevância e a eficácia das soluções em projetos semelhantes. Em resumo, a escolha de bibliotecas e ferramentas no nosso projeto é uma decisão informada, baseada em considerações teóricas, práticas e na adequação às necessidades específicas da análise de tendências salariais em Ciência de Dados. Isso garante a eficácia da análise e a qualidade dos resultados obtidos.

4.7. Machine Learning para Previsão de Salários Futuros

A aplicação de técnicas de Machine Learning (Aprendizado de Máquina) é uma parte fundamental do nosso projeto, que visa prever os salários futuros de profissionais em Ciência

de Dados. Nesta seção, discutiremos as bases teóricas e conceituais que sustentam a escolha de métodos de ML e as ferramentas relacionadas ao nosso projeto.

Importância do Machine Learning

Machine Learning desempenha um papel central em nosso projeto de análise de tendências salariais. Suas contribuições incluem:

Modelagem Preditiva: Utilizamos algoritmos de ML para construir modelos que preveem salários futuros com base em dados históricos e outras variáveis relevantes.

Exploração de Relações Complexas: ML permite a identificação de relações complexas entre as variáveis que podem não ser facilmente detectadas por métodos tradicionais de análise estatística.

Automatização: Os modelos de ML automatizam o processo de previsão, economizando tempo e aumentando a eficiência.

Bases Teóricas da Escolha de Algoritmos de Machine Learning

Nossa escolha de algoritmos de ML é fundamentada nas seguintes considerações teóricas:

Random Forest Regressor: O uso do algoritmo Random Forest Regressor é motivado pela sua capacidade de lidar com problemas de regressão, nos quais tentamos prever um valor contínuo, como salários. O algoritmo é baseado em árvores de decisão e é conhecido por seu desempenho sólido.

Divisão de Dados: Utilizamos o train_test_split para dividir nossos dados em conjuntos de treinamento e teste, permitindo a avaliação do desempenho do modelo.

Imputação de Dados Ausentes: O SimpleImputer é aplicado para tratar valores ausentes nos dados, garantindo que o modelo seja alimentado com dados completos.

Pipeline e Transformação de Colunas: O uso de pipelines e transformadores de colunas (ColumnTransformer) ajuda na organização do fluxo de trabalho do ML, incluindo a codificação one-hot para variáveis categóricas e escalonamento padrão para variáveis numéricas.

Adaptação aos Objetivos do Projeto

A escolha do algoritmo de Random Forest Regressor foi adaptada aos objetivos específicos do projeto, que envolvem a previsão de salários com base em várias variáveis, incluindo experiência, tipo de emprego e localização.

Hiperparâmetros e Otimização

A otimização de hiperparâmetros é uma etapa crítica no treinamento do modelo. O uso de técnicas como RandomizedSearchCV é fundamentado na busca de combinações ideais de hiperparâmetros para maximizar o desempenho do modelo.

Revisão da Literatura

A revisão da literatura forneceu insights sobre a aplicação bem-sucedida de técnicas de ML em projetos semelhantes e validou a escolha do Random Forest Regressor como um algoritmo apropriado.

REFERÊNCIAS BIBLIOGRÁFICAS

I) KUMARDATALAB, harish. Data Science Salary 2021 to 2023. Kaggle, 2023. Disponível em: https://www.kaggle.com/datasets/harishkumardatalab/data-science-salary-2021-to-2023. Acesso em: 10 set. 2023.

II) Mello, Oliveira Leornado. CIÊNCIA DE DADOS APLICADA A GESTÃO DE PROJETOS DE QUALITY ASSURANCE. Universidade Federal do Rio Grande do Sul, 2021.

III) BEHESHTI, Nima. Random Forest Regression. Medium, 2022. Disponível em: https://towardsdatascience.com/random-forest-regression-5f605132d19d. Acesso em: 11 out. 2023

IV) What is a Decision Tree?. IBM. Disponível em: https://www.ibm.com/topics/decision-trees. Acesso em: 11 out. 2023

V) Github do projeto: https://github.com/ViniSegatto/Analise-salario