UNIVERSIDADE FEDERAL DO CEARÁ

ERICK RAMOS COUTINHO
SHELDA DE SOUZA RAMOS
MAVERICK ALEKYNE DE SOUSA RIBEIRO

PROJETO DE LABORATÓRIO DE CIÊNCIA DE DADOS

ITAPAJÉ-CE
2023

RESUMO

A área de dados tem ganhado relevância no Brasil devido ao avanço tecnológico e à disponibilidade crescente de dados. Empresas, governos e organizações reconhecem a importância dos dados para melhorar decisões e impulsionar a inovação. Setores como finanças, saúde e governo lideram projetos de análise de dados, apoiados por tecnologias como inteligência artificial. No entanto, desafios como segurança de dados e falta de profissionais qualificados persistem. Ao longo deste relatório, será feita a descrição de tudo que foi realizado neste projeto da disciplina de laboratório de ciência de dados, as plataformas utilizadas, a motivação para a escolha deste tema e as metodologias que contribuíram para a análise descritiva dos dados. Este relatório busca analisar a situação e tendências da área de dados no Brasil, fornecendo informações para orientar decisões futuras e insights sobre esse ramo em relação ao mercado de trabalho.

OBJETIVO

O objetivo deste trabalho é realizar uma análise descritiva básica das questões relevantes relacionadas à área de dados no Brasil. Desejamos compreender e comunicar a situação atual, tendências e desafios enfrentados na área de dados em relação ao mercado de trabalho. Por meio dessa análise, nosso propósito é fornecer informações significativas que possam orientar a tomada de decisões e influenciar o desenvolvimento futuro da área de dados no Brasil.

JUSTIFICATIVA

A seleção deste dataset é guiada por uma motivação intrínseca à nossa paixão e compromisso com a área de Ciência de Dados. Compreender as tendências e desafios relacionados ao mercado de trabalho em nossa área é essencial para nosso crescimento profissional e para o progresso contínuo do campo da Ciência de Dados como um todo.

Vivemos em uma era em que os dados desempenham um papel crucial em todas as esferas da sociedade, da economia à saúde, da educação à tecnologia. Como cientistas de dados, nossa missão é explorar as infinitas possibilidades que os dados oferecem e aplicar nosso conhecimento para solucionar problemas complexos. No entanto, essa jornada requer um entendimento profundo do cenário atual, das demandas do mercado de trabalho e dos desafios que enfrentamos.

Ao escolher este dataset, estamos buscando não apenas adquirir informações valiosas sobre as tendências emergentes na área de Ciência de Dados, mas também entender como podemos nos preparar e contribuir de maneira mais significativa. Queremos estar na vanguarda das inovações e tendências, e a análise deste dataset nos permitirá fazer escolhas informadas e estratégicas em nossa trajetória profissional.

1. INTRODUÇÃO

A área de dados tem desempenhado um papel cada vez mais significativo no cenário brasileiro, impulsionada pelo avanço tecnológico, a crescente disponibilidade de dados e o reconhecimento de sua importância estratégica. No Brasil, como em muitos outros lugares do mundo, a revolução dos dados está moldando a maneira como empresas, governos e organizações tomam decisões e conduzem operações.

Nos últimos anos, o Brasil testemunhou um aumento significativo na conscientização sobre a importância dos dados, com um crescente número de empresas e instituições governamentais investindo em iniciativas de análise de dados e ciência de dados. Isso se deve, em parte, ao reconhecimento de que a capacidade de coletar, processar e interpretar dados pode fornecer insights valiosos, melhorar a tomada de decisões e impulsionar a inovação em vários setores da economia.

Setores como finanças, saúde, agronegócio, e-commerce e governo têm liderado o caminho na implementação de projetos de dados robustos. Além disso, a adoção de tecnologias emergentes, como inteligência artificial e aprendizado de máquina, tem se expandido, permitindo análises mais avançadas e previsões precisas.

No entanto, apesar dos avanços, a área de dados no Brasil ainda enfrenta desafios significativos, incluindo questões de privacidade e segurança de dados, a falta de profissionais qualificados em ciência de dados e análise de dados, e a necessidade de regulamentações e políticas mais claras para governança de dados.

Este relatório tem como objetivo realizar uma análise básica de algumas questões-chave relacionadas à área de dados no Brasil, como por exemplo: a distribuição dos níveis de ensino. Abordaremos alguns aspectos específicos que ajudarão a compreender a situação atual e as tendências no uso de dados na educação brasileira. Por meio dessa análise, esperamos fornecer informações valiosas para orientar a tomada de decisões e o desenvolvimento futuro da área de dados no Brasil.

2. MÉTODOS UTILIZADOS

Durante a análise dos dados, foram utilizados vários métodos para classificar as variáveis e entender melhor o conjunto de dados. Alguns dos métodos e técnicas incluíram:

Média: Calculamos a média para analisar os salários por cargo atual, a fim de identificar os cargos com salários mais altos, e também para fazer a média geral das idades.

Desvio Padrão: Foi usado desvio padrão para calcular o desvio das idades dos entrevistados, a fim de saber o quanto as idades se dispersam.

Tabelas de Frequência: Criamos tabelas de frequência para analisar a distribuição de valores em variáveis categóricas, como nível de ensino e gestores.

Gráficos: Durante a análise foi utilizado gráficos para visualizar a distribuição de algumas variáveis categóricas, como nível de ensino e cargo atual.

Filtragem e Agrupamento: Realizamos filtragem e agrupamento de dados para segmentar o conjunto de dados com base em critérios específicos, como gestores e nível de ensino.
Análise de Dados Faltantes: Identificamos dados ausentes e discutimos como tratá-los, dependendo do contexto
Transformação de Dados: Realizamos transformações nos dados, como a conversão de faixas salariais em valores numéricos.
Análise Descritiva: Realizamos análises descritivas para resumir e entender as características do conjunto de dados.

Neste projeto, adotamos uma abordagem rigorosa para o desenvolvimento e gerenciamento de código, utilizando as seguintes ferramentas e plataformas-chave:

1. Ambiente de Desenvolvimento (IDE): PyCharm

Para a criação, edição e execução de nosso código, optamos pelo ambiente de desenvolvimento PyCharm. A escolha do PyCharm foi motivada por sua interface amigável, recursos avançados de depuração e integração perfeita com Python, facilitando o desenvolvimento eficiente e aprimorado do código.

2. Controle de Versionamento: GitHub

Gerenciar as várias iterações de nosso projeto e colaborar com a equipe de forma eficaz foi possível graças ao GitHub. Utilizamos o GitHub como nosso repositório central de controle de versão, permitindo o rastreamento preciso das mudanças no código, a colaboração simultânea de membros da equipe e a implementação de um fluxo de trabalho de desenvolvimento contínuo (CI/CD) para manter nosso projeto sempre atualizado e funcional.

Essas escolhas estratégicas de ferramentas e plataformas desempenharam um papel fundamental na execução bem-sucedida deste projeto, proporcionando eficiência no desenvolvimento, rastreamento de mudanças e colaboração entre os membros da equipe.

3. CONCLUSÕES PÓS PRÉ-PROCESSAMENTO

3.1. PRÉ-PROCESSAMENTO

Após a conclusão do processo de pré-processamento dos dados, várias observações e conclusões importantes podem ser citadas. Primeiramente, notamos uma melhoria significativa na qualidade geral dos dados. Isso foi alcançado através da limpeza de dados ausentes, algumas mudanças na estrutura do dataset que estavam causando problemas e alteração dos nomes das colunas para melhor visualização. Essas alterações permitiram uma melhor análise dos dados, pois poderiam causar problemas futuros.
Além disso, o pré-processamento dos dados teve um impacto positivo na visualização e interpretação das informações. As visualizações dos dados se tornaram mais claras e

informativas, facilitando a identificação de tendências e padrões. Essa clareza nas visualizações é crucial para a tomada de decisões informadas e a comunicação eficaz dos resultados. Em termos de descobertas específicas, identificamos algumas tendências interessantes após a limpeza dos dados. Por exemplo, notamos que a média salarial para a função de "Arquiteto de Dados" é de aproximadamente 20.000 reais. Essa informação pode ser valiosa para a equipe de recursos humanos e a tomada de decisões relacionadas a salários e benefícios, além de um chamativo para pessoas que pensam em trabalhar nessa área buscarem o cargo de Arquiteto de Dados.
Em resumo, o pré-processamento de dados desempenhou um papel crucial na preparação dos dados para análises posteriores. Ele melhorou a qualidade, clareza e consistência dos dados, permitindo-nos extrair insights valiosos e tomar decisões mais informadas.

3.2. ANÁLISE DESCRITIVA DOS DADOS

Uma análise descritiva é uma técnica estatística que visa resumir e descrever os principais aspectos de um conjunto de dados, de forma a proporcionar uma compreensão mais clara das informações contidas nele. Após realizar algumas análises sobre o dataset em questão, podemos tirar conclusões sobre ele, como por exemplo:

- Área Demográfica:

A maioria dos respondentes é do gênero masculino, na faixa etária de 30-34 anos, com maior representação na região Sudeste do Brasil. A maioria possui nível de educação pós-graduação e tem mais de 10 anos de experiência na área de dados.
Existem entrevistados empregador de até 18 anos, em período de graduação, mostrando que muitas vezes é possível conseguir um emprego ainda nesse período, como por exemplo, um estágio. Enquanto a idade máxima é 54 anos.

- Emprego e Setor:

A maioria dos profissionais trabalha sob regime CLT em consultoria ou indústria. A faixa salarial mais comum varia de R$ 6.001/mês a R$ 12.000/mês, e a maioria tem menos de 2 anos de experiência na empresa atual.
Existe uma preferência significativa por modelos de trabalho totalmente remotos ou híbridos. A maioria dos profissionais consideraria procurar outra oportunidade se a empresa optasse pelo modelo presencial.
A área que mais possui pessoas empregadas dentre os entrevistados é o cargo de Cientista de Dados, com 357. É possível prever esse fato, pois é o cargo mais abrangente, tendo áreas de atuação em diversos mercados.

- Desafios como Gestor:

Os gestores enfrentam desafios como contratar e reter talentos, convencer a empresa a investir em dados e gerar valor para as áreas de negócios.

- Tecnologias e Ferramentas:

Python e SQL são as linguagens de programação mais amplamente utilizadas, enquanto AWS, Google Cloud e Azure são as principais plataformas em nuvem. PowerBI, Tableau e

Looker são as ferramentas de BI mais comuns. O uso de ferramentas de AutoML e Analytics é relativamente alto.

- Atividades Diárias:

As atividades diárias variam amplamente, desde o desenvolvimento técnico até a gestão de projetos e de pessoas.

- Conhecimento do Data Hackers:

A maioria dos respondentes conhece o Data Hackers, principalmente através do blog e do podcast.

4. REFERÊNCIAS BIBLIOGRÁFICAS

Data Hackers. Kaggle: Your Machine Learning and Data Science Community, 2021. Datasets. Disponível em: <https://www.kaggle.com/datasets/datahackers/state-of-data-2021>. Acesso em: 05 de out. de 2023

LVES-MAZZOTI, A. J.; GEWANDSZNAJER, F. O método nas ciências naturais e sociais: pesquisa quantitativa e qualitativa. 2.ed. São Paulo: Pioneira, 1998

BURREL, G.; MORGAN, G. Sociological paradigms and organizational analysis. Londres: Heinemann Books, 1979

Link de acesso aos códigos: https://github.com/ErickCoutinho/LAB_CD