Análise de Dados – Exploratória e ML

Sumário

Análise de Dados para *(nome do aplicativo)*

Nome da startup

Nome dos integrantes da equipe do 2° Tech AD – em ordem alfabética

Introdução

Descrição do aplicativo e do público-alvo – 1 a 2 parágrafos, contendo:

→ o problema que tenta resolver

→ o público-alvo

Objetivo do aplicativo

Um parágrafo que defina de forma específica como o aplicativo resolve o problema que foi descrito.

Recomenda-se citar aqui que os integrantes do 1° ano já fizeram a especificação do aplicativo e que os detalhes sobre essa parte podem ser encontrados no material desenvolvido por eles para documentação do trabalho.

Objetivo da Análise Exploratória de Dados

Montar, analisar e limpar um conjunto de dados relacionado ao aplicativo para ser utilizado em análise de dados exploratória e também para fins de aprendizado de máquina (ML).

Os dados que serão incluídos nesse conjunto podem ser provenientes de bases de dados governamentais (Brasil, Estados Unidos, União Europeia etc.), científicas ou de sondagem própria das equipes envolvidas (Forms, por exemplo).

Objetivo da Análise de Dados (para Aprendizado de Máquina)

Liste quais os modelos que serão utilizados e o escreva o que vão prever (se uma pessoa é, ou não, um cliente em potencial para o seu APP, qual o seu nível do interesse, quais produtos podem interessar a ela, quais meios de comunicação podem atingir essa pessoa, etc.

Levantamento dos dados na análise exploratória

Busca dos dados

Aqui, o grupo descreve os critérios e os métodos para buscar os dados. Por exemplo:

- Levantamento de dados por questionário criado pelo próprio grupo
- Dados em bases oficiais/do governo/da ONU
- Bases de veículos de comunicação – jornais, revistas especializadas (não serve a Caras)

Coloque aqui os questionários que a equipe de Dados fez e indique quantas pessoas responderam. Se os questionários foram feitos pelos outros integrantes da equipe do Projeto (Dev ou 1° ano), não é necessário incluir aqui.

Justificativa de uso

Explique por que é importante esses dados que foram escolhidos, qual a importância deles para o trabalho. Essas justificativas podem ser colocadas em uma seção separada, ou então indicadas junto da descrição da busca de dados.

Descrição da base de dados de trabalho

Aqui, faça como nas atividades do primeiro semestre. Inclua coisas como:

- Tipos dos dados: inteiro, data/hora, string etc.
- Valores-limite: máximo e mínimo para dados numéricos, lista de valores para dados categóricos
- ~~Plotagem dos gráficos de análise preliminar dos dados brutos usando Python.~~
- Parâmetros estatísticos que descrevem os dados: média, moda, desvio padrão, variância etc.

Sempre indique como fez a análise, seja com seu código Python próprio, ou dizendo qual a biblioteca de análise exploratória de dados (por exemplo, ydata-profiling).

No relatório, coloque os resultados da análise dos dados brutos (parâmetros estatísticos e gráficos). O código (notebook) deve ser fornecido em um anexo ou como um link para o GitHub.

Para plotagem dos dados brutos encontrados, monte um dashboard, de acordo com os requisitos da disciplina Business Intelligence (aproveite o que for desenvolvido para essa disciplina).

Limpeza dos dados

Descrição do procedimento de limpeza e preparação dos dados para poder prosseguir para a análise preditiva, ou seja:

- quais colunas/campos foram limpos e por quê (remoção de outliers, dados claramente incorretos etc)
- gráficos que ajudem na justificativa
- como transformou os dados (mudança de tipo, arredondamento etc)

No relatório, coloque os resultados da limpeza (parâmetros estatísticos e gráficos). O código (notebook) deve ser fornecido em um anexo ou como um link para o GitHub.

Para plotagem dos dados limpos, monte um dashboard, de acordo com os requisitos da disciplina Business Intelligence (aproveite o que for desenvolvido para essa disciplina).

Condicionamento para alimentar o modelo de ML

O condicionamento de dados será realizado antes de rodar o modelo de ML. Também será feito após rodar o modelo de ML e for constatado que é necessário algum ajuste para reprocessamento.

Este item deverá conter uma subseção para cada etapa de condicionamento dos dados.

Condicionamento inicial

- Normalização, padronização
- Redução de dimensionalidade

- Escolha de uma variável dentro de um conjunto de variáveis com alta correlação etc.
- Avaliação dos dados preparados: Dashboard de Business Intelligence.

Os itens “Condicionamento após rodar o modelo pela primeira/segunda vez” podem ser retirados, porque ficam melhor na parte de Análise de dados, conforme indicado mais adiante, na seção “Mudanças na Base de Dados:”.

~~Condicionamento após rodar o modelo pela primeira vez~~

~~Indicar aqui o que foi alterado~~

~~Condicionamento após rodar o modelo pela segunda vez~~

~~Indicar aqui o que foi alterado~~

AQUI COMEÇA A PARTE DE ANÁLISE DE DADOS (MÓDOLO)

Definição dos objetivos e das classes

Explicar em detalhes os objetivos que desejam prever com a aplicação dos modelos. Escrever quais são as perguntas ou problemas específicos que precisam resolver com a aplicação dos modelos.

Listar e descrever a(s) resposta(s) (y) e as classes que cada uma dessas respostas vai prever.

Definição dos modelos mais adequados para analisar os dados

Escrever uma visão geral inicial dos modelos que levantaram para serem aplicados nas análises e quais critérios utilizados para selecionar aqueles que serão aplicados nos seus dados. (Aqui é uma explicação teórica antes de aplicar os modelos aos dados)

Apresentar uma visão geral dos modelos de classificação que analisaram para aplicar aos seus dados. Explicar por que esses modelos são potenciais candidatos para resolver o seu problema de classificação e os critérios utilizados para essa seleção, considerando aspectos como:

- Natureza dos Dados: dados categóricos ou contínuos que precisam ser classificados.
- Complexidade dos Modelos: modelos simples ou modelos mais complexos.
- Objetivos de Classificação: tarefa é binária ou multi-classe.

Descrição dos modelos selecionados

Descrever cada modelo selecionado em uma subseção a seguir (criar quantas subseções forem necessárias para os modelos selecionados). Escrever os critérios que utilizou para seleção do modelo, considerando aspectos como:

- Precisão e Robustez: capacidade do modelo de classificar corretamente novos dados e lidar com variações nos dados.
- Interpretação e Explicabilidade: facilidade em interpretar os resultados e entender como as previsões são feitas.
- Desempenho Computacional: tempo e recursos necessários para treinar e aplicar o modelo.
- Capacidade de Generalização: habilidade do modelo de manter um bom desempenho em dados não vistos.

Modelo 1: [Nome do Modelo]

Descrever o modelo 1.

Escrever os critérios que utilizou para seleção do modelo 1.

Modelo 2: [Nome do Modelo]

Descrever o modelo 2.

Escrever os critérios que utilizou para seleção do modelo 2.

Modelo n: [Nome do Modelo]

Descrever o modelo n.

Escrever os critérios que utilizou para seleção do modelo n.

Aplicação dos modelos selecionados

Explicar como cada modelo foi aplicado aos dados, como foram feitos a divisão, o treinamento e a avaliação dos modelos.

- Divisão dos Dados: detalhar a divisão dos dados em conjuntos de treinamento e teste (ex.: 80% treinamento e 20% teste, validação cruzada, etc).
- Treinamento: explicar o processo de treinamento para cada modelo, incluindo ajuste de hiperparâmetros.
- Métricas de Avaliação: apresentar as métricas utilizadas para avaliar o desempenho dos modelos, como acurácia, precisão, recall e F1-score, e descrever como essas métricas ajudam a avaliar a eficácia dos modelos.

Análise dos Resultados

Discutir os resultados obtidos para cada modelo com base nas métricas de avaliação.

Modelo 1: [Nome do Modelo]

Métricas de Avaliação: apresentar as métricas de avaliação obtidas para o modelo.

Discussão dos Resultados: analisar como o modelo performou em cada uma das métricas e o que esses resultados indicam sobre seu desempenho, considerando aspectos como:

- Equilíbrio entre Precisão e Recall: se o modelo está tendendo a obter alta precisão mas baixo recall, ou vice-versa.
- Classes Desbalanceadas: se o modelo tem dificuldade em classificar classes menos frequentes.
- Erro e Overfitting: se o modelo apresenta sinais de overfitting, ou seja, se ele tem um desempenho muito bom no conjunto de treinamento mas ruim no conjunto de teste.

Modelo 2: [Nome do Modelo]

Métricas de Avaliação.

Discussão dos Resultados.

Modelo n [Nome do Modelo]

Métricas de Avaliação.

Discussão dos Resultados.

Comparação Entre Modelos:

Fazer a comparação dos resultados das métricas de avaliação e identificar as forças e fraquezas de cada modelo.

Comparação Direta: comparar os resultados das métricas de avaliação entre os modelos. Discuta qual modelo obteve melhor desempenho em cada métrica e quais aspectos de cada modelo contribuem para essas diferenças.

Forças e Fraquezas: identificar as forças e fraquezas de cada modelo com base nas métricas de avaliação e nos objetivos da análise.

Ajustes Necessários

Realizar mudanças na base de dados ou nos modelos para melhorar os resultados, se necessário, e explicar as modificações realizadas e o impacto delas nos resultados.

Identificação de Problemas:

Problemas de Desempenho: identificar problemas específicos encontrados com os modelos, como baixa acurácia, baixa precisão, ou baixa recall.

Dados Desbalanceados: se as classes estão desbalanceadas, considerar como isso pode estar afetando o desempenho do modelo.

Mudanças na Base de Dados:

Ajustar o pré-processamento e/ou o balanceamento das classes para melhorar o desempenho.

Pré-processamento:

- Limpeza de Dados: revisar e ajustar a limpeza de dados para remover outliers ou erros que podem estar afetando o desempenho.
- Transformação de Dados: considerar técnicas de normalização ou padronização adicionais.

- Engenharia de Características: adicionar ou modificar características para melhorar a representação dos dados, tais como, criar novas variáveis a partir das existentes ou usar técnicas de seleção de características.

Balanceamento de Classes: usar técnicas de balanceamento como oversampling (ex.: SMOTE) ou undersampling para tratar desbalanceamento de classes e melhorar o desempenho do modelo.

Ajustes nos Modelos:

Verificar a necessidade de ajustes nos hiperparâmetros e/ou utilização de validação cruzada, caso ainda não tenham feito:

Hiperparâmetros: realize ajuste de hiperparâmetros para otimizar o desempenho do modelo.

Validação Cruzada: usar validação cruzada para avaliar o desempenho dos modelos de forma mais robusta e evitar overfitting.

Impacto das Modificações

Discutir novamente os resultados obtidos para cada modelo com base nas métricas de avaliação.

Modelo 1: [Nome do Modelo]

Discussão dos Resultados: explicar como as mudanças na base de dados e os ajustes nos modelos ajudaram a melhorar as métricas de avaliação.

Modelo 2: [Nome do Modelo]

Discussão dos Resultados: explicar como as mudanças na base de dados e os ajustes nos modelos ajudaram a melhorar as métricas de avaliação.

Modelo n [Nome do Modelo]

Discussão dos Resultados: explicar como as mudanças na base de dados e os ajustes nos modelos ajudaram a melhorar as métricas de avaliação.

Comparação Entre Modelos:

Fazer a comparação dos resultados das métricas de avaliação, identificar as forças e fraquezas de cada modelo.

Escolher o modelo mais adequado para seus dados e explicar porque é o mais adequado.

Conclusão

Resumir as principais descobertas da análise e a eficácia dos modelos aplicados.

Discutir o potencial impacto das previsões no seu projeto e as próximas etapas recomendadas para aprimorar a análise.