**Template para entrega do projeto da disciplina**

**Ciência de Dados e Inteligência Artificial**

**Fase 2**

|  |  |
| --- | --- |
| **Nome do estudante** | **Leonardo Ladeira Rodrigues** |

Desenvolva um processo de ciência de dados no Orange Data Mining, cobrindo os elementos abaixo. **Para cada um dos itens solicitados é necessário inserir imagens que evidenciem o trabalho realizado**.

|  |
| --- |
| Exploração dos dados |
| *Que tipo de experimentos você fez na exploração dos dados (verificação de outliers, cálculos de médias, dados inválidos etc.).* |
| Análise Preditiva:  A análise preditiva é uma técnica que utiliza dados históricos e algoritmos de aprendizado de máquina para prever eventos futuros ou resultados. |

|  |
| --- |
| Escolha de, ao menos, três algoritmos de aprendizado para a modelagem |
| *Apresente os algoritmos utilizados e justifique a escolha.* |
| 1. Support Vector Machine (SVM):    Justificativa: O SVM é tipo um super filtro que separa os dados em categorias diferentes, como se fosse uma linha ou plano imaginário que divide tudo certinho. Escolhi ele porque funciona bem quando temos muitas variáveis e os dados não são tão simples de separar.  2. k-Nearest Neighbors (KNN):    Justificativa: O KNN é como perguntar a opinião dos vizinhos. Ele olha para os dados mais próximos para decidir a categoria de um novo dado. É fácil de entender e usar, porque basta definir quantos vizinhos (k) você quer considerar para tomar a decisão.  3. Árvore de Decisão:  Justificativa: A árvore de decisão é como um jogo de perguntas e respostas. Cada "pergunta" separa os dados até chegar a uma decisão final. É ótimo porque dá para ver exatamente como as decisões são tomadas, o que ajuda a entender o que está acontecendo com os dados. |

|  |
| --- |
| Preparação dos dados de acordo com as características dos algoritmos de aprendizado escolhidos |
| *Descreva o processo realizado para essa etapa.* |
| 1. Importação dos Dados: carreguei os dados no Orange usando o widget File. Isso é fundamental para começar qualquer análise, pois precisamos dos dados para trabalhar.  2. Seleção de Colunas: Em seguida, utilizei o widget Select Columns. Onde, pode escolher quais colunas (variáveis) quer incluir ou excluir da análise. Isso ajuda a focar nas variáveis mais relevantes e a evitar ruídos nos dados.  3. Visualização dos Dados: Com o widget Data Table, visualizando os dados em uma tabela. Isso facilita a inspeção manual dos dados, permitindo verificar se tudo está conforme o esperado e identificar possíveis problemas ou inconsistências.  4. Aplicação de Cores: Usei o widget Color para atribuir cores às variáveis ou valores específicos. Isso é útil para destacar padrões ou categorias nos dados, tornando a visualização mais intuitiva.  5. Visualização com Scatter Plot: O Scatter Plot é um gráfico de dispersão que permite visualizar a relação entre duas variáveis. Isso ajuda a identificar correlações e possíveis padrões nos dados. Ao colorir os pontos, você pode ver como diferentes categorias se distribuem no gráfico.  6. Conexão com Data Table: Depois de usar o Feature Statistics, eu conectei o Data Table ao próximo widget para visualizar os dados processados. Isso ajudou a garantir que os dados estavam formatados corretamente antes de prosseguir.  7. Uso do Data Sample: Eu utilizei o widget Data Sample duas vezes:  Primeiro Data Sample: Para criar uma amostra dos dados de treinamento e teste. Este passo é crucial para dividir os dados originais em partes que podem ser usadas para treinar os modelos e verificar sua precisão.  Segundo Data Sample: Para criar uma nova amostra, talvez para validação cruzada ou outro tipo de análise específica.  8. Análise no Test & Score: Depois, eu usei o widget Test & Score para avaliar a performance dos modelos:  Teste de Dados e Dados: Configurei para realizar a validação dos modelos utilizando diferentes conjuntos de dados, como treinamento e teste.  Confusion Matrix: Conectei a análise a duas matrizes de confusão em paralelo com dois modelos KNN para visualizar e comparar os resultados.  9. Árvore de Decisão: Depois de criar a segunda amostra de dados com o Data Sample, utilizei o widget Árvore de Decisão.  Propósito: O objetivo foi treinar um modelo de árvore de decisão e visualizar como o modelo toma decisões com base nas características dos dados.  Benefício: As árvores de decisão são fáceis de interpretar, permitindo entender quais variáveis são mais importantes e como elas influenciam a classificação.  10. Box Plot: Em seguida, utilizei o widget Box Plot para visualizar a distribuição das variáveis.  Propósito: O box plot ajuda a identificar a mediana, quartis e possíveis outliers nas variáveis, proporcionando uma visão clara da distribuição dos dados.  Benefício: Isso é útil para entender a variabilidade dos dados e verificar a presença de valores atípicos que podem influenciar os modelos.  11. SVM - Scatter Plot e Silhouette Plot: Finalmente, utilizei o algoritmo SVM e visualizei os resultados com Scatter Plot e Silhouette Plot.  SVM com Scatter Plot:  Propósito: O scatter plot ajudou a visualizar como o SVM separa os dados em diferentes classes, mostrando a fronteira de decisão entre as classes.  Benefício: Isso permite ver claramente como o SVM está classificando os dados e identificar quaisquer dados que possam estar incorretamente classificados.  SVM com Silhouette Plot:  Propósito: O silhouette plot foi utilizado para avaliar a qualidade do clustering feito pelo SVM, medindo quão bem os dados estão agrupados.  Benefício: Isso ajuda a entender se os clusters formados pelo SVM são compactos e bem separados, o que é um indicativo de bom desempenho do modelo. |

|  |
| --- |
| Execução dos experimentos de aprendizado e coleta das métricas |
| *Descreva o processo e os resultados obtidos.* |
| Para a execução dos experimentos de aprendizado, segui um processo estruturado e meticuloso para garantir que os modelos fossem bem treinados e avaliados. Aqui está uma descrição detalhada do processo:  1. Preparação dos Dados:  Dividi os dados em conjuntos de treinamento e teste utilizando o widget Data Sample. Isso assegurou que tivéssemos uma parte dos dados para treinar os modelos e outra para testar a sua eficácia.  2. Treinamento dos Modelos:  Utilizei os widgets específicos para cada algoritmo:  SVM (Support Vector Machine): Treinei o modelo SVM com os dados de treinamento.  k-Nearest Neighbors (k-NN): Configurei dois modelos KNN para treinamento, variando os parâmetros para comparar os resultados.  Árvore de Decisão: Treinei o modelo de árvore de decisão para compreender melhor as decisões baseadas nas características dos dados.  Para cada modelo, ajustei os hiperparâmetros para otimizar o desempenho, como a escolha do kernel no SVM ou o número de vizinhos no k-NN.  3. Avaliação dos Modelos:  Utilizei o widget Test & Score para validar os modelos. Esse widget permitiu realizar a validação cruzada e avaliar a performance dos modelos com dados que não foram utilizados no treinamento.  As métricas de avaliação incluíram:  Acurácia: A porcentagem de previsões corretas feitas pelo modelo.  Precisão: A proporção de verdadeiros positivos em relação ao total de positivos previstos pelo modelo.  Recall: A capacidade do modelo de identificar corretamente todas as ocorrências de uma classe.  F1-score: A média harmônica da precisão e recall, oferecendo um equilíbrio entre as duas métricas.  4. Análise dos Resultados:  Confusion Matrix: Utilizei o widget Confusion Matrix para visualizar a performance dos modelos KNN. Isso ajudou a entender onde os modelos estavam acertando e errando nas previsões.  Scatter Plot e Silhouette Plot: Para o SVM, utilizei esses gráficos para visualizar a separação das classes e a qualidade do agrupamento.  Box Plot: Utilizado para analisar a distribuição dos dados após a aplicação dos modelos, ajudando a identificar possíveis outliers e variabilidade nos resultados.  Resultados obtidos:  SVM  Acurácia Alta:  Observação: O modelo SVM apresentou uma acurácia alta, o que demonstra a eficiência do modelo em separar as classes corretamente.  Implicação: Isso significa que o modelo é confiável para prever as categorias baseadas nos dados de entrada fornecidos.  Scatter Plot:  Análise: Utilizei o scatter plot para visualizar a separação das classes com base no tipo de operação que a aeronave estava realizando durante o incidente ou acidente ao longo dos anos.  Resultado: A separação das classes foi clara e bem definida, indicando que o modelo conseguiu discernir os diferentes tipos de operações de forma eficaz.  Silhouette Plot:  Análise: O silhouette plot foi utilizado para avaliar a qualidade do agrupamento formado pelo SVM, mostrando o que causou o acidente ou incidente por região.  Resultado: Os clusters foram bem definidos e compactos, o que indica que os dados foram agrupados de maneira consistente e significativa.  KNN Os resultados mostram uma variação no desempenho dos modelos k-NN com diferentes valores de CA. O Modelo 2, com um CA maior, apresentou melhorias em várias métricas-chave, como acurácia e recall, indicando uma melhor capacidade de generalização e desempenho global mais estável. Contudo, o valor de MCC baixo em ambos os modelos sugere que ainda há espaço para ajustes e melhorias.  Árvore de decisão  A árvore de decisão proporcionou insights valiosos sobre os fatores que contribuem para acidentes e incidentes aéreos. A alta interpretabilidade do modelo ajudou a entender melhor os dados e a identificar padrões importantes que podem ser utilizados para melhorar a segurança aérea. |

|  |
| --- |
| Relato dos experimentos e lições aprendidas |
| *Apresente uma reflexão acerca dos resultados obtidos com este projeto.* |
| Ao longo deste projeto, realizamos uma série de experimentos de aprendizado de máquina utilizando diferentes algoritmos (SVM, k-NN, Árvore de Decisão) para analisar dados de acidentes e incidentes aéreos.  Este projeto foi extremamente enriquecedor, proporcionando insights valiosos sobre a análise de dados de acidentes e incidentes aéreos. Algumas das lições mais importantes incluem a necessidade de uma preparação cuidadosa dos dados, a escolha apropriada dos algoritmos e a importância de uma avaliação detalhada usando múltiplas métricas.  Essas lições não só ajudaram a melhorar a qualidade dos modelos neste projeto específico, mas também forneceram uma base sólida para futuros projetos de ciência de dados. A compreensão dos padrões e características dos dados nos permitirá desenvolver melhores estratégias de prevenção e segurança no futuro. |

|  |
| --- |
| LINK PARA O ARQUIVO DO PROJETO DO ORANGE E DOS DADOS UTILIZADOS |
| *Insira os links para os arquivos.* |
| https://drive.google.com/drive/folders/11-SvcFON\_CVGOJwmcy0GLn-e6ZxF1pgy |