**Template para entrega do projeto da disciplina**

**Ciência de Dados e Inteligência Artificial**

**Fase 2**

|  |  |
| --- | --- |
| **Nome do estudante** | **Nicholas Custódio Boreto Francatti** |

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.).* |
| Diagrama, Mapa  Descrição gerada automaticamente  Em ambos os prints demonstro que tipos de testes de exploração eu fiz no meu dataset e quais widgets usei para tal. No caso do primeiro print fiz uso dos widgets de visualização para encontrar relações fortes as quais eu pudesse usar para os algoritmos a fim de treinar os dados para eles.  Como no caso do **Scatter Plot** em que eu pude ver que grande parte dos filmes do grupo com **IMDB Rating** de **8.7 a 9.3** (**laranja**) tem as maiores quantidades de votos e ficam mais entre os **anos 90** para cima:    No **Boxplot** foi possível ver a proporção de filmes com base no **certificado** para cada **gênero**:    No caso do **Linear** **Projection** pude ver a relação forte entre **Número de Votos, Bilheteria e Nota no IMDB**, tendo um olhar especial para o grupo de filmes com notas entre **8.7-9.3** (**laranja**):    Ademais, usei o widget **Correlations** para me auxiliar a ver combinações relevantes entre as colunas:    Por fim, usei o **Data Table, Pie Chart e Feature Statistics**, sendo o **último** o mais **útil** para visualizações de **moda, média, mínimo, máximo** e entre outras informações importantes:    Quanto aos **outliers**, fiz essa **seleção de colunas** mais importantes, tirando as colunas que não seriam relevantes antes de usar o widget **Data Sampler** para treinar os modelos escolhidos, usando o widget **Select Columns**, para selecionar as colunas relevantes, **excluindo** no caso as colunas **Director, Overview, Poster\_link, Series\_title e Star1 até Star4**.    Em relação aos dados **inválidos** ou **nulos**, tratei da forma como vi grande parte das pessoas tratando nesse tipo de dataset, nas três colunas em que havia dados faltantes **Certificate**, **Metascore** e **Gross**, preenchi com a média dos valores, mas caso o professor/tutor que corrigir preferir, posso reenviar o trabalho utilizando outra forma, no caso excluindo as linhas em que ao menos 1 desses dados estiver nulo/vazio. Para tal usei o widget **Impute**. |

|  |
| --- |
| Escolha de, ao menos, três algoritmos de aprendizado para a modelagem |
| *Apresente os algoritmos utilizados e justifique a escolha.* |
| Os algoritmos utilizados foram, **Random Forest, KNN, Word Cloud e Sentiment Analysis**. No caso dos dois últimos, são **ADDONS** do **Orange** em que adicionei a ferramenta para fazer **Data Mining de texto**, pois achei interessante para o dataset escolhido. Sendo assim, começando por eles, o **Word Cloud** foi usado justamente para análise da coluna **Overview** (Resumo/Sinopse Curta) podendo-se visualizar as palavras mais utilizadas entre todos os Resumos. Já o **Sentiment Analsys** nada mais é do que uma **Análise de Sentimento** mais generalizada, limitada obviamente pela ferramenta do Orange, mas que também foi usada na coluna **Overview** para que fossem analisados os **resumos**, predizendo os sentimentos que cada um deles mais passaria.  No caso do algoritmo **Random Forest**, ele foi usado para predizer as notas do IMDB (**IMDB\_Rating**) com base em algumas **features** escolhidas do dataset.  Por fim, o algoritmo **KNN** foi escolhido para tentar adivinhar **o gênero principal** do filme com base também em algumas **features** escolhidas do dataset. |

|  |
| --- |
| Preparação dos dados de acordo com as características dos algoritmos de aprendizado escolhidos |
| *Descreva o processo realizado para essa etapa.* |
| Para preparar os dados de acordo com as características dos algoritmos tive que realizar algumas mudanças. A primeira foi **alterar os gêneros**, pois como pode ser visto na primeira fase, o dataset leva em consideração de 1 até 4 gêneros por filme, o que atrapalha bastante na hora de tentar predizê-lo, sendo assim, preferi limitar essa coluna **Genre** só para o seu primeiro valor, ou seja, só o primeiro gênero que consta de cada filme seria exposto, facilitando o trabalho do algoritmo.    Em seguida percebi que a coluna **IMDB\_Rating** por ser numérica e em float poderia atrapalhar a predição quando usado o **kNN**, por esse motivo preferi usar o widget **Discretize** para fazer **endpoints** para as notas, separando-as em grupos com os intervalos **7.6, 8.1, 8.7, 9.3** e depois nomeando os **breakpoints** por **E, D, C, B e A** respectivamente, da mais baixa para a mais alta, facilitando assim o trabalho do algoritmo ao tentar prever em qual dos grupos o filme pertencia, passando a tratá-los **categoricamente**.      Para a **mineração de dados com texto** eu tive que fazer um pré-processamento com o widget de **pré**-**processamento** específico de data mining. Nesse caso utilizei **Transformation** para colocar todas as palavras em letra minúscula e retirar acentos, **Tokenization** para separar as palavras por pontuação com a configuração **Word** **Punctuation** que foi a que se saiu melhor e por fim usei **Filtering** com uma lista de **stopwords** adicionada na pasta ao final do trabalho que deve ser usada já que se saiu melhor do que as **stopwords** padrões do Orange em inglês: |

|  |
| --- |
| Execução dos experimentos de aprendizado e coleta das métricas |
| *Descreva o processo e os resultados obtidos.* |
| No algoritmo **Random Forest**, utilizei a configuração padrão, com **10 árvores**:    E obtive os seguintes resultados, me surpreendendo com tamanha precisão para o uso do algoritmo, predizendo muito bem a **categoria** que o filme se encaixava com base no **Rating do IMDB**:    A **matriz de confusão** do **Random Forest** ficou assim:    Já no algoritmo **kNN**, tentei diversas alterações, mudando tanto as configurações do algoritmo quanto as **features** que seriam escolhidas para seu uso, mas a melhor configuração que cheguei foi a da **métrica** **Chebyshev** e com **2 vizinhos**. Caso eu trocasse o **peso de uniforme para distância**, tinha um problema com os resultados que se tornavam perfeitos demais dando **overfitting**:    Notei que como explicado em aula no meu caso quanto mais vizinhos eu adicionava pior ficava a precisão do algoritmo, com razão, por isso a redução para **2** para tentar chegar em um valor de precisão mais alto, mas ainda assim, não me contentei com os **0.656 de precisão**. Porém, como não foi determinado um valor mínimo de precisão para os algoritmos, mantive o **kNN**, segue o resultado:    A **matriz** **de** **confusão** do **kNN** ficou assim:    Por fim, no **Word Cloud** o resultado foi bem satisfatório, mostrando palavras realmente relevantes nos **Resumos/Sinopses** dos filmes:    Já no **Sentiment Analysis** foi usado o método **SentiArt** em inglês que foi o que mais se encaixou e mais trouxe possibilidades e resultados mais satisfatórios dentro do possível:    Obtive os seguintes **resultados** com essa **Análise de Sentimentos**: |

|  |
| --- |
| Relato dos experimentos e lições aprendidas |
| *Apresente uma reflexão acerca dos resultados obtidos com este projeto.* |
| A partir dos resultados obtidos pude perceber a quantidade de algoritmos que poderiam ser utilizados para predição do **IMDB\_Rating**, mas a minha escolha foi a Random Forest, mesmo não tendo a visualização que o algoritmo **Tree** tem, porém, obtendo melhores resultados em acurácia.  Quanto ao **kNN**, como já dito, não foi a acurácia que eu desejava, mas como não houve nenhuma exigência no trabalho quanto a isso, mantive o **kNN** justamente como forma de aprendizado e como forma de amostragem de que é algo experimental e que me trouxe muito mais esforço por tentar ajustar o algoritmo tanto em suas configurações e pré-processamentos, seleção de colunas do que o **Random Forest**, o que me trouxe muito mais aprendizado.  Em relação aos algoritmos de **Text Mining (Addons)**, caso o tutor/professor corretor prefira que eu utilize de outros algoritmos para o trabalho posso entregá-lo novamente fazendo outras escolhas, mas acredito que seja bastante interessante esse estudo e análise dos textos curtos de sinopse do dataset do IMDb.  Primeiramente por conseguir produzir uma **Word Cloud** bem concisa e com palavras chave expressivas, conseguindo com sucesso demonstrar quais são as mais utilizadas, bem como dar um simples relance de como funciona uma **Análise de Sentimentos**, por mais superficial possa ser por conta dos textos serem curtos e do algoritmo ser fixo, por mais que tenha várias opções, sem poder ser mexido ou alterado, o que acabou causando inclusive alguns erros cômicos, em que um filme de comédia trazia sentimentos de **Medo**, por haver na sinopse curta justamente a palavra **Medo**. |

|  |
| --- |
| LINK PARA O ARQUIVO DO PROJETO DO ORANGE E DOS DADOS UTILIZADOS |
| *Insira os links para os arquivos.* |
| <https://drive.google.com/drive/folders/18wk8Q-iNPOOTW1EEutXieZ1KesSWwonj?usp=sharing>  Fiz upload de uma pasta no Google Drive com os respectivos arquivos:   * Lista de stopwords em inglês para serem usadas no widget/algoritmo de Word Cloud * Projeto do dataset no Orange * Arquivo XLSX do dataset |