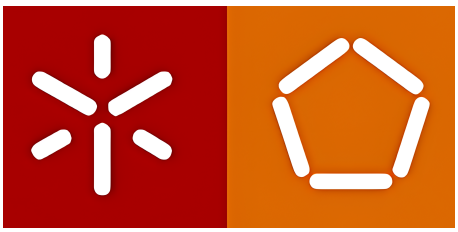

Universidade do Minho
Licenciatura em Engenharia Informática

Trabalho Prático

Aprendizagem e Decisão Inteligente 2023/24

Conceção de Modelos de Aprendizagem e Decisão

Grupo 14:

Bernardo Lima [A93258]
David Teixeira [A100554]
João Pedro Pastore [A100543]
Luís Ferreira [A91672]

2 de março de 2024

Conteúdo

Lista de Figuras

1 Introdução

Este documento foi elaborado no âmbito da Unidade Curricular de Aprendizagem e Decisão Inteligentes, na qual temos a responsabilidade de desenvolver modelos de aprendizagem. O projeto em questão aborda duas atividades distintas. A primeira envolve a pesquisa, análise e seleção de um conjunto de dados escolhido pelo nosso grupo, seguido por uma exploração, análise e preparação dos mesmos. Já a segunda atividade requer a exploração, análise e preparação de um conjunto de dados fornecido pelos professores da disciplina, seguido pela concepção de modelos de *machine learning* de classificação e regressão, além de uma análise crítica dos resultados obtidos.

2 Tarefa Dataset Atribuído

Nesta tarefa, a equipa docente forneceu um conjunto de dados que possuem informações de vários pacientes médicos (cerca de 583), incluindo idade, género e diversos marcadores bioquímicos. Para abordar este problema, iremos adotar a metodologia **CRISP-DM**.

Vamos seguir as suas etapas para entender melhor o problema, explorar os dados disponíveis, preparar os dados para análise, desenvolver modelos preditivos, avaliar esses modelos e, por fim, implementar a solução para a tarefa em questão.

2.1 Compreensão do Negócio

Na primeira fase da metodologia **CRISP-DM**, é essencial compreender os objetivos e requisitos do projeto do ponto de vista do negócio. Isso envolve identificar o problema, determinar os objetivos do projeto e estabelecer um critério de aceitação para definir o projeto como entregue.

2.1.1 Problema

O problema central é a crescente taxa de mortalidade por cirrose hepática, impulsionada pelo aumento do consumo de álcool, infecções crónicas por hepatite e doenças hepáticas relacionadas à obesidade. Embora a detecção precoce da patologia hepática seja crucial para melhorar os resultados do paciente, existem disparidades observadas, especialmente em relação ao diagnóstico precoce da patologia hepática em pacientes do sexo feminino. O conjunto de dados fornecido contém registos de pacientes no Nordeste de *Andhra Pradesh*, Índia, com a tarefa de prever se um paciente sofre de doença hepática com base em marcadores bioquímicos.

2.1.2 Objetivo

Os objetivos estabelecidos pela equipa de trabalho são os seguintes:

1. Desenvolver um modelo que determine se um paciente sofre de doença hepática com base em marcadores bioquímicos.
2. Investigar disparidades de género no diagnóstico precoce de patologia hepática e avaliar a eficácia dos marcadores bioquímicos para pacientes do sexo masculino e feminino.
3. Analisar eventuais inclinações ou distorções em algoritmos de saúde, particularmente durante análises segmentadas por sexo na previsão de doenças hepáticas.

2.1.3 Critério de Aceitação

O critério de aceitação para este projeto pode ser definido como:

- O modelo desenvolvido deve apresentar uma taxa de precisão acima de um limite estabelecido (por exemplo, 80%) na previsão de doença hepática com base nos marcadores bioquímicos fornecidos.
- A análise das disparidades de género deve ser realizada com sucesso, identificando diferenças significativas na eficácia dos marcadores bioquímicos para pacientes do sexo masculino e feminino.
- A investigação sobre possíveis viéses ou distorções em algoritmos de saúde, especialmente em análises segmentadas por sexo, deve ser concluída e documentada.

2.2 Compreensão dos Dados

A fase de compreensão dos dados é crucial para entender a estrutura e a natureza das informações disponíveis. Analisando o conjunto de dados fornecido, podemos destacar alguns pontos importantes:

- Formato dos Dados: O conjunto de dados é apresentado num formato .csv, onde cada linha representa um paciente e cada coluna representa uma característica específica. Este *dataset* possui 583 linhas e 17 colunas (17 atributos).
- Variáveis Disponíveis:
  - **id_code**: Identificador da linha
  - **Age**: Idade
  - **birth_year**: Ano de nascimento
  - **birth_month**: Mês de nascimento
  - **birth_date**: Data de nascimento
  - **Gender**: Género
  - **TB**: Bilirrubina total
  - **DB**: Bilirrubina direta
  - **Alkphos**: Fosfatase alcalina
  - **Sgpt**: Alanina aminotransferase
  - **Sgot**: Aspartato aminotransferase
  - **TB (#1)**: Proteínas totais
  - **ALB**: Albumina
  - **CHOL**: Colesterol
  - **A/G Ratio**: Relação albumina/globulina
  - **BILmg**: Bilirrubina em miligramas por decilitro no sangue
- Características Demográficas: Além das informações bioquímicas, o conjunto de dados também inclui características demográficas, como idade e género do paciente.
- Problemas nos Dados: Algumas inconsistências nos dados são observadas, como valores ausentes representados por "00/00"em datas de nascimento e variações no formato de género (por exemplo, "female"e "Female", "Masculine"e "Male").
- Classe Alvo: A variável alvo é representada pela coluna *Selector*, que indica se o paciente sofre de doença hepática (valor "1=liver disease") ou não (valor "2=no liver disease").

Compreender estes aspetos dos dados é fundamental para o processo de preparação e análise subsequente. As etapas seguintes envolverão a limpeza dos dados, a seleção de características relevantes e a exploração mais aprofundada para extrair insights significativos para o problema em questão.

2.2.1 Preparação dos Dados

Durante a etapa inicial de preparação dos dados, foi realizada uma análise minuciosa dos atributos do conjunto de dados, seguida por ajustes necessários para garantir que os valores dos atributos estivessem em conformidade com as expectativas estabelecidas. Uma série de procedimentos de tratamento de dados foi aplicada aos seguintes atributos, visando otimizar a sua qualidade e consistência. Destaca-se que, através do nodo *CSV Reader*, fomos capazes de realizar as alterações.

- **TB (#1)**: Dado que este atributo deve representar *Total Proteins*, optamos por convertê-lo para **TP**. Para alcançar este objetivo, utilizamos o nodo *Column Renamer*, com as seguintes configurações:

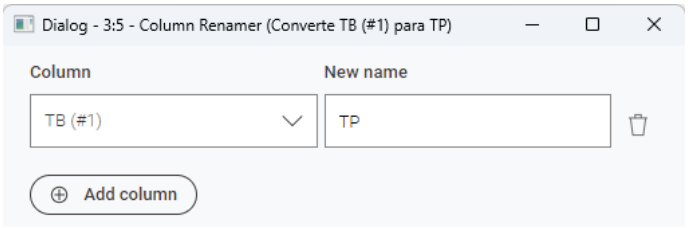

Figura 1: Configurações no nodo *Column Renamer*

- **Gender**: Com o intuito de homogeneizar este atributo, dado que existem variações no formato de gênero (por exemplo, "female"e "Female"), optamos por transformar estes dados para um simples ***M*** para *Male* e ***F*** para *Female*. Para tal, utilizamos o nodo *String Manipulation*, com as seguintes configurações:

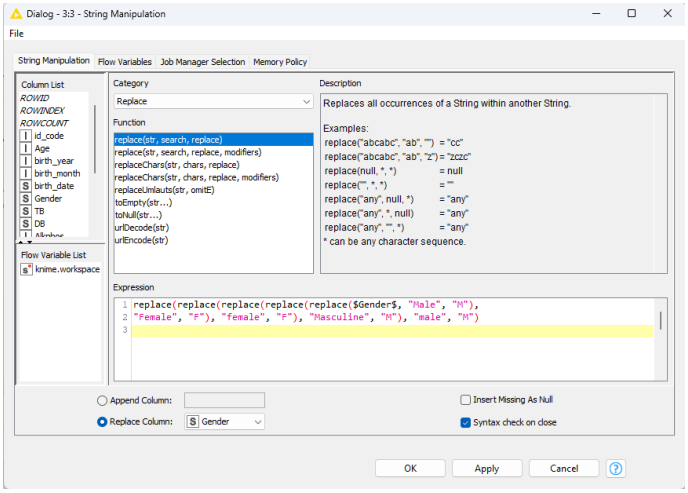

Figura 2: Configurações no nodo *String Manipulation*

- **birth_date, id_code e birth_month**: Os atributos correspondentes à data de nascimento foram removidos, dada a sua natureza inútil para este estudo, não acrescentando absolutamente nenhuma nova informação ao *dataset*. Por motivos semelhantes, também removemos os atributos *id_code* e *birth_month*. Para remover estes atributos, utilizamos o nodo *Column Filter*, com a seguinte configuração:

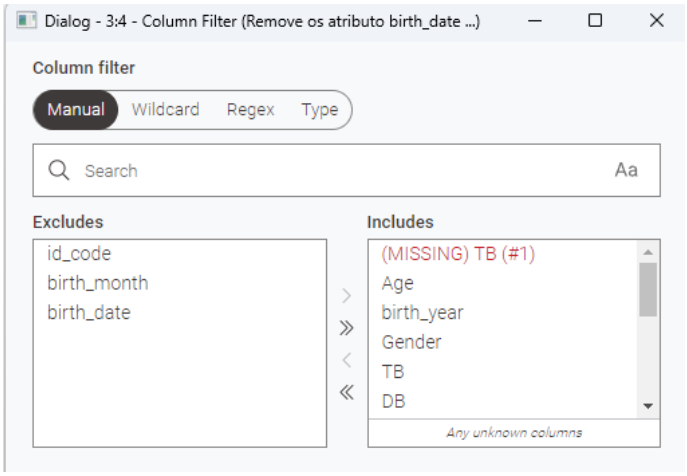

Figura 3: Configurações no nodo *Column Filter*

- **Selector**: Inicialmente, este atributo possui os seguintes valores:
  - 1=liver disease

- 2=no liver disease
- 2=without liver disease

De modo a simplificar o estudo, utilizamos o nodo *String Manipulation* para transformar o atributo em apenas dois simples valores: 1 ou 2. Eis a configuração no respectivo nodo:

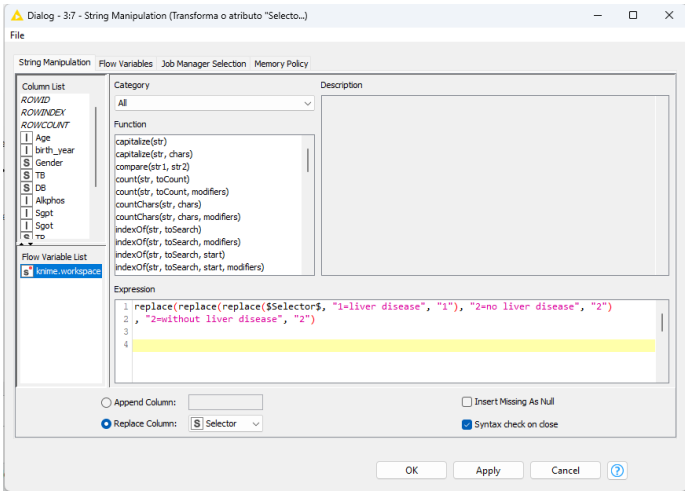

Figura 4: Configurações no nodo *String Manipulation*

- **Selector, TP, DB, TP, ALB, AG_Ratio, BILmg**: Este tipo de atributos, maioritariamente relacionados com marcadores bioquímicos (com exceção do atributo *Selector*), têm o tipo de dados definido como *String*. Sendo valores numéricos, devem ser identificados como tal. Utilizamos o nodo *String to Number* para essa finalidade. Eis a configuração:

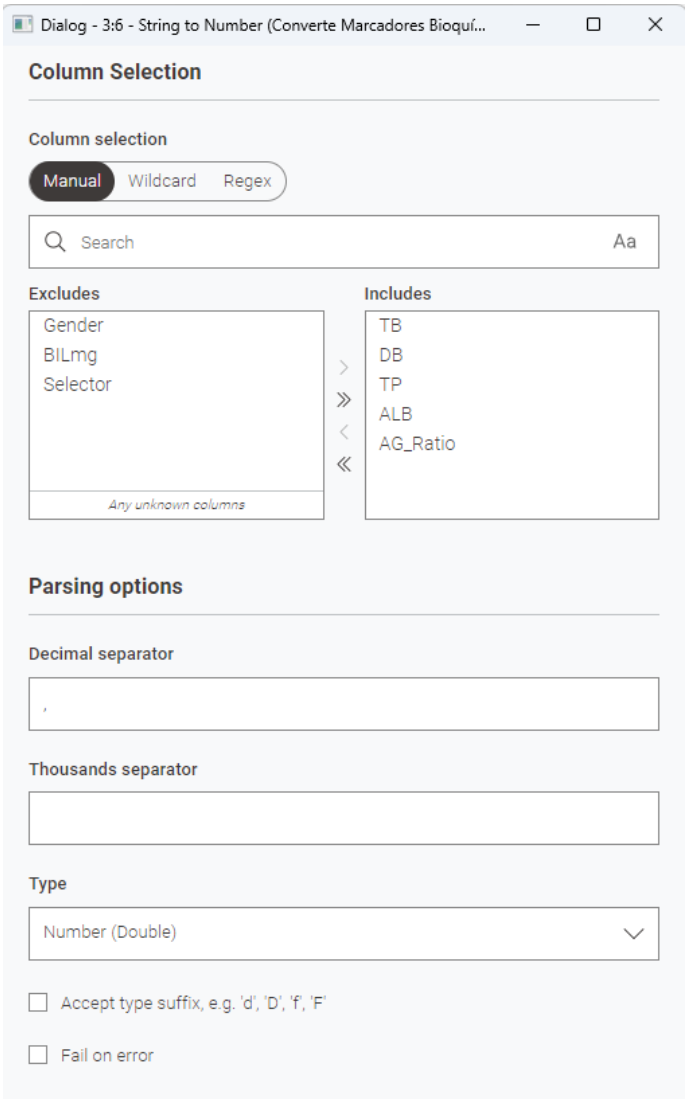

Figura 5: Configurações no nodo *String to Number* de *String* para *Double*

Nota: Foi promovida uma tentativa de conversão do atributo *BILmg* de *String* para *Long*, porém não foi possível. Assim, foi convertido para *double*. Consequentemente, os valores de *BILmg* foram arredondados às milésimas. Este arredondamento não deve afetar o estudo em questão.

2.3 Modelação

2.4 Avaliação

2.5 Implementação