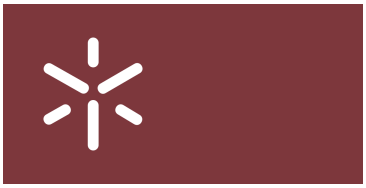


Universidade do Minho
Escola de Engenharia
Licenciatura em Engenharia Informática

Unidade Curricular de Aprendizagem e Decisão Inteligentes

Ano Letivo de 2023/2024

Conceção de Modelos de Aprendizagem e Decisão

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

Maio, 2024

Data da Receção	
Responsável	
Avaliação	
Observações	

Conceção de Modelos de Aprendizagem e Decisão

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

Maio, 2024

Índice

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 *dataset* escolhido pelo nosso grupo, seguido por uma exploração, análise, preparação, modelação e avaliação. Já a segunda atividade requer a exploração, análise, preparação de um conjunto de dados fornecido pelos professores da disciplina, seguido pela conceção de modelos de *machine learning* de classificação, além de uma análise crítica dos resultados obtidos.

2. Tarefa Dataset Grupo

O grupo selecionou um *dataset* sobre vendas de videojogos para a sua análise. Este conjunto de dados oferece informações abrangentes sobre uma variedade de videojogos, incluindo detalhes como nome, plataforma, ano de lançamento, género, editora e vendas em diferentes regiões do mundo. O objetivo principal da análise é identificar os principais atributos que influenciam as vendas de videojogos e avaliar a força dessas relações.

O mercado de videojogos é uma indústria em constante crescimento, com uma base de fãs global e diversificada. Compreender quais fatores contribuem para o sucesso de um videojogo em termos de vendas é crucial para desenvolvedores, editores e investidores. Estas informações podem orientar estratégias de desenvolvimento de jogos, campanhas de *marketing* e decisões de investimento.

O grupo optou por seguir a metodologia **CRISP-DM**, que é um modelo padrão composto por seis etapas: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelação, Avaliação e Desenvolvimento. As cinco primeiras etapas serão abordadas.

2.1. Compreensão do Negócio

Na primeira fase da metodologia **CRISP-DM**, é crucial 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 encontra-se na necessidade de compreender os padrões e factores que influenciam as vendas de videojogos, num mercado que está em constante evolução e crescimento. Este sector enfrenta desafios decorrentes da diversidade de plataformas e do gosto variado dos consumidores. O *dataset* fornecido contém registos detalhados de vendas de videojogos, permitindo análises aprofundadas para identificar padrões de sucesso e tendências de mercado.

2.1.2. Objetivo

Os objetivos estabelecidos pela equipa de trabalho são os seguintes:

1. Desenvolver um modelo que procure prever as vendas globais de videojogos com base em atributos como plataforma, género, editora, entre outros.
2. Investigar os principais fatores que influenciam as vendas de videojogos e identificar padrões de sucesso em diferentes regiões do mundo.
3. Avaliar o impacto de variáveis como ano de lançamento, género e plataforma na performance de vendas dos videojogos.

2.1.3. Critério de Aceitação

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

- O modelo desenvolvido deve apresentar um coeficiente de determinação acima de um limite estabelecido (por exemplo, 0.7) na previsão das vendas globais de videojogos.
- A análise dos fatores que influenciam as vendas deve ser realizada com sucesso, identificando os principais impulsionadores de sucesso no mercado de videojogos.
- A avaliação do impacto de variáveis como ano de lançamento, género e plataforma nas vendas de videojogos deve ser conduzida e documentada, fornecendo *insights* valiosos para *stakeholders* do sector.

2.2. Compreensão dos Dados

A fase de compreensão dos dados é essencial para entender a estrutura e a natureza das informações disponíveis. Ao analisar o *dataset* fornecido sobre vendas de videojogos, destacamos os seguintes pontos importantes:

- **Formato dos Dados:** O conjunto de dados é apresentado em formato *.csv*, onde cada linha representa um jogo e cada coluna representa uma característica específica. Este *dataset* possui um total de 16598 linhas e 11 colunas (11 atributos).
- **Variáveis Disponíveis:**

  -**Rank**: Classificação do jogo com base nas vendas globais.

  -**Name**: Nome do jogo.

  -**Platform**: Plataforma em que o jogo foi lançado.

  -**Year**: Ano de lançamento do jogo.

  -**Genre**: Género do jogo.

  -**Publisher**: Editora responsável pelo jogo.

  -**NA_Sales**: Vendas na América do Norte (em milhões de unidades).

  -**EU_Sales**: Vendas na Europa (em milhões de unidades).

  -**JP_Sales**: Vendas no Japão (em milhões de unidades).

  -**Other_Sales**: Vendas em outras regiões (em milhões de unidades).

  -**Global_Sales**: Vendas globais totais (em milhões de unidades).
- **Problemas nos Dados:** Alguns problemas podem estar presentes nos dados, como *missing values* ou inconsistências nos formatos das colunas.
- **Variável Alvo:** A variável alvo é indicada pelas colunas *x_Sales*, onde x representa uma região. Neste contexto, focamo-nos nas regiões da América do Norte (NA), Europa (EU), Japão (JP) e global (Global).

2.2.1. Preparação dos Dados para Exploração

Durante a fase inicial de preparação dos dados para análise, realizamos uma investigação detalhada dos atributos do *dataset*, seguida por ajustes necessários para garantir que os valores dos atributos estivessem em conformidade com as expectativas estabelecidas. Diversos procedimentos de limpeza e tratamento de dados foram aplicados a esses atributos específicos, com o objetivo de melhorar a sua qualidade e consistência.

- **Remoção da Acentuação**: Utilizamos o nodo *Java Snippet* para remover a acentuação dos nomes, normalizando assim as *strings* associadas.
- **Remoção de Valores Ausentes**: Empregamos o nodo *Row Filter* para eliminar todas as linhas do *dataset* que continham *missing values*.

A partir destes dois procedimentos, seguimos quatro abordagens distintas, cada uma com objetivos específicos: analisar estatisticamente as vendas por franquia, editora (*publisher*), género e ano.

Para analisar estatisticamente as vendas **por franquia**, realizamos o seguinte tratamento:

1. **Clonagem do Atributo Nome**: Utilizamos o nodo *string manipulation* para clonar toda a coluna de nomes (exceto aqueles que continham números).
2. **Separação da Coluna Clonada**: Empregamos o nodo *cell splitter* para dividir a coluna clonada anteriormente em palavras (Exemplo: *New Super Mario Bros Wii* → *New* | *Super* | *Mario* | *Bros* | *Wii*).
3. **Filtragem de Palavras**: Utilizamos o nodo *groupBy* para selecionar as quatro primeiras palavras obtidas anteriormente (Exemplo: *New* | *Super* | *Mario* | *Bros* | *Wii* → *New* | *Super* | *Mario* | *Bros*).
4. **Concatenação de Palavras**: Usamos o nodo *string manipulation* para concatenar as quatro primeiras palavras obtidas anteriormente (Exemplo: *New* | *Super* | *Mario* | *Bros* → *New Super Mario Bros*).
5. **Agregação de Franquias**: Empregamos o nodo *groupBy* para agregar a coluna de franquias e somar as vendas (Exemplo: *Super Mario Bros* → Vendas Globais = 73.64).
6. **Utilização de Expressões Regulares**: Utilizamos o nodo *string manipulation* e o método *regexReplace( )* para identificar as franquias de uma forma mais eficiente (Exemplo: *regexReplace(" .*Mario.* ", "Super Mario")*).
7. **Nova Agregação de Franquias**: Empregamos o nodo *groupBy* para agregar a coluna *new column* (que continha a identificação aprimorada das franquias) e somar as vendas (Exemplo: *Super Mario* → Vendas Globais = 526.28).
8. **Ordenação de Atributos**: Utilizamos o nodo *sorter* para organizar as linhas por vendas globais.

Para analisar estatisticamente as vendas **por editora** (publisher), realizamos o seguinte tratamento:

1. **Agregação de Editoras**: Utilizamos o nodo *groupBy* para agregar a coluna de editoras e somar as vendas (Exemplo: *Nintendo* → Vendas Globais 1678.82).
2. **Ordenação de Atributos**: Empregamos o nodo *sorter* para organizar as linhas por vendas globais.

Para analisar estatisticamente as vendas **por género**, realizamos o seguinte tratamento:

1. **Agregação por Género**: Utilizamos o nodo *groupBy* para agregar a coluna de género e somar as vendas (Exemplo: *Sports* → Vendas Globais 1257.57).

Para analisar estatisticamente as vendas **por ano**, realizamos o seguinte tratamento:

1. **Agregação por Ano**: Utilizamos o nodo *groupBy* para agregar a coluna *year* e somar as vendas (Exemplo: 2003 → Vendas Globais 348.37).

2.2.2. Exploração Inicial dos Dados

Na fase de exploração de dados, o nosso grupo optou por analisar estatisticamente as ramificações mencionadas anteriormente. Todas as análises seguiram uma estrutura consistente:

Top 10 Baseado na Região: Utilizamos o nodo *top k row filter* para selecionar apenas as linhas cujos atributos em estudo estivessem na maior faixa de valores de vendas de uma determinada região.

Ordenação de Valores: Empregamos o nodo *sorter* para classificar os 10 valores obtidos anteriormente.

Visualização Gráfica: Utilizamos o nodo *bar chart* para visualizar os resultados obtidos de forma gráfica.

2.2.2.1. Analise das Vendas por *Franchise*

Região Global: A franquia de jogos mais vendida globalmente, com um destaque interessante, é *Super Mario*, seguida por *Pokemon*. Ambos são jogos que têm as suas raízes no continente asiático, mais precisamente no Japão, o que confere ao continente um papel de liderança no setor de videojogos. Além destes,

também merecem destaque os jogos da franquia *Wii*, que ocupam a quarta posição em termos de vendas globais.

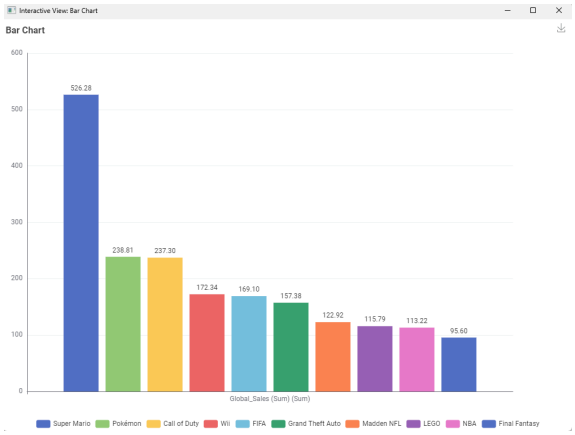


Figura 1: *Vendas a Nível Global*

Região da Europa: Na região europeia, destaca-se o *FIFA*, um jogo desportivo de futebol que é lançado anualmente desde 1993. Contrariamente ao que as estatísticas globais possam sugerir, os jogos de futebol são altamente procurados na Europa. No entanto, é interessante notar que tanto *Super Mario* quanto *Wii* mantêm as suas posições. Para além disso, *Call of Duty* continua a ocupar a terceira posição, conforme observado no estudo anterior. Por outro lado, Pokémon acaba por ser uma "desilusão" em termos de vendas na Europa, descendo do 2º lugar para o 5º comparativamente ao caso anterior.

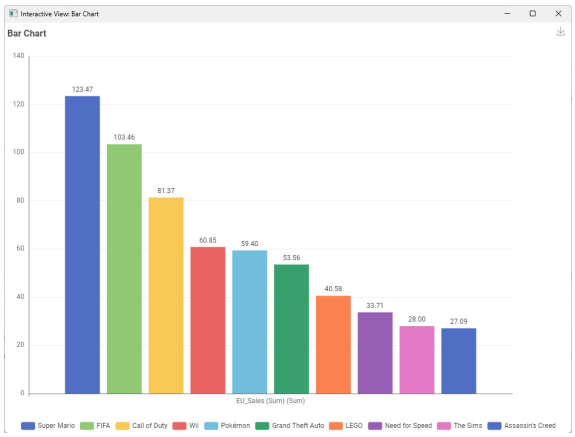


Figura 2: *Vendas a Nível Europeu*

Região da América do Norte: Na região da América do Norte, *Super Mario* mantém-se novamente como a franquia mais vendida, seguido por *Call of Duty* no segundo lugar. No entanto, à semelhança do que se verificou no caso europeu, os jogos desportivos ocupam o terceiro lugar. Neste caso, *Madden* assume essa posição. *Madden* é uma franquia de videojogos de futebol americano, lançada pela primeira vez em 1988 e nomeada em homenagem ao ex-treinador da *NFL*, John Madden. Para além disso, é digno de nota que os jogos da *NBA* (franquia de jogos de *basketball*) ocupam o quinto lugar em termos de vendas na região da América do Norte.

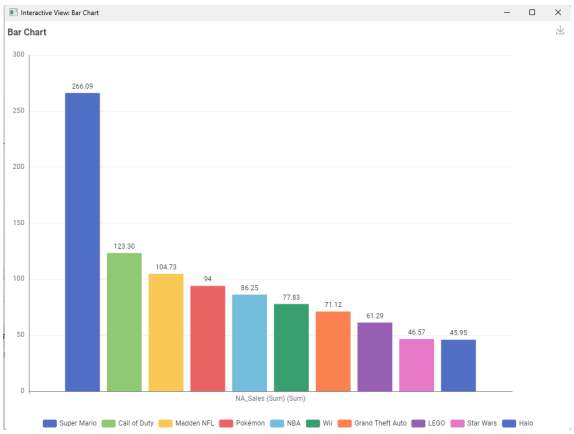


Figura 3: *Vendas a Nível Norte-Americano*

Região Japonesa: Finalmente, na região nipónica, como seria de esperar, *Super Mario* e *Pokemon* destacam-se claramente dos outros jogos devido à sua origem no próprio país. Além disso, é digno de nota que *Final Fantasy* surge em terceiro lugar, seguido por *Dragon Quest* em quarto e *Monster Hunter* em quinto. Estes jogos ocupam posições de destaque devido à sua natureza japonesa e à forte presença na cultura de jogos do Japão. Notemos, também, que ao contrário do verificado anteriormente nas regiões europeias e norte-americanas, o Japão não é conhecido por investir em jogos desportivos.

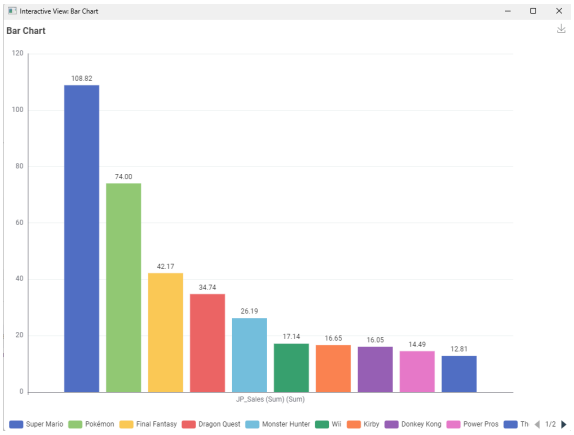


Figura 4: *Vendas a Nível Japonês*

2.2.2.2. Analise das Vendas por *Publisher*

Região Global: Em termos globais e no cenário dos videojogos, a *Nintendo* destaca-se de forma incontestável, principalmente devido ao sucesso das suas franquias de jogos icónicas, como *Super Mario* e *Wii*. Com décadas de história e um legado duradouro, a *Nintendo* conquistou uma base de fãs leal e uma posição de destaque na indústria.

Em segundo lugar, mas não menos significativo, encontra-se a *Electronic Arts* (EA), impulsionada principalmente pelo sucesso dos seus jogos desportivos anteriormente referidos: *FIFA* e *Madden*. Estes títulos, que recriam fielmente experiências desportivas populares, conquistaram uma grande base de fãs e contribuíram significativamente para o reconhecimento e sucesso da EA como editora de jogos.

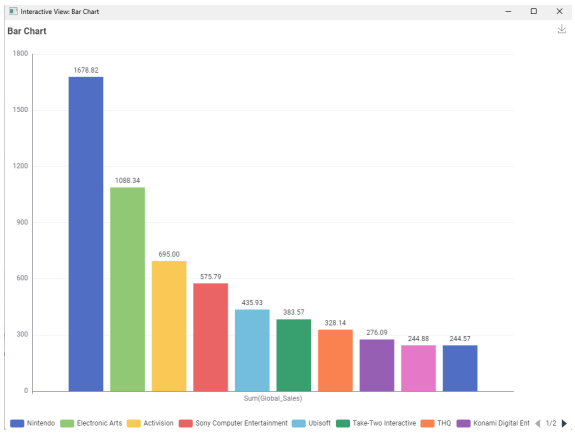


Figura 5: *Vendas a Nível Global*

Região da Europa: A nível europeu, a competição entre a *Nintendo* e a *Electronic Arts* é intensa, com ambas as empresas a se destacar devido à grande paixão dos europeus, principalmente, pelo jogo da EA - o *FIFA*. Esta franquia de jogos desportivos conquistou uma enorme popularidade na Europa, contribuindo significativamente para a posição competitiva da EA na região. Como resultado, a *Nintendo* e a EA estão muito próximas em termos de reconhecimento e sucesso no mercado europeu.

No entanto, em comparação com o mercado global, o *ranking* é praticamente o mesmo.

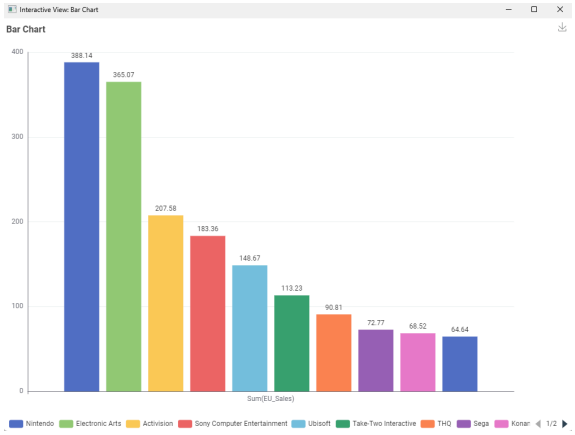


Figura 6: *Vendas a Nível Europeu*

Região da América do Norte: No mercado norte-americano, tanto a *Nintendo* quanto a *Electronic Arts* mantêm a sua competitividade, embora haja uma ligeira queda para a *Electronic Arts* em comparação com o mercado europeu. Isto deve-se em parte ao facto de que o jogo mais vendido da EA na América do Norte, o *Madden*, não é tão procurado quanto o *FIFA* na Europa. Apesar disso, ambas as empresas continuam a ser jogadoras importantes no mercado norte-americano de jogos eletrónicos.

Adicionalmente, há um crescimento mais evidente da *Activision* neste mercado, impulsionado pela paixão dos americanos pela franquia *Call of Duty*. A popularidade duradoura e o apelo dos jogos da série *Call of Duty* têm contribuído significativamente para o sucesso da *Activision* na América do Norte, tornando-a uma força a ser reconhecida e aumentando a sua participação no mercado de jogos eletrónicos da região.

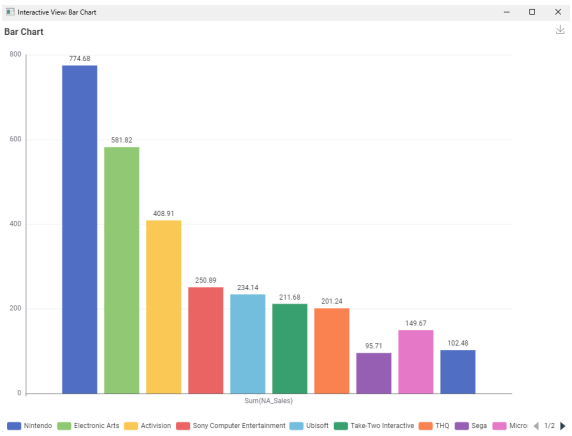


Figura 7: *Vendas a Nível Norte-Americano*

Região Japonesa: Mais uma vez, como concluímos na análise anterior, na região japonesa os dados são completamente diferentes dos europeus, norte-americanos e globais.

Por um lado, como seria de se esperar, a *Nintendo* mantém-se em primeiro lugar, mas de uma maneira completamente absurda, dominando de forma avassaladora o mercado japonês (um verdadeiro monopólio estrondoso). De seguida, surge de forma surpreendente (comparativamente ao verificado anteriormente) a *Bandai Namco*, responsável por franquias de jogos como *Tekken*, que conquistou uma posição significativa no mercado japonês.

Para além disso, destacamos também a *Konami*, graças a jogos como *Yu-Gi-Oh!*, que mantém uma presença notável no cenário dos jogos japoneses. Não podemos deixar de mencionar a *Capcom*, que também se destaca com as suas próprias franquias de sucesso e continua a ter uma forte influência na região japonesa.

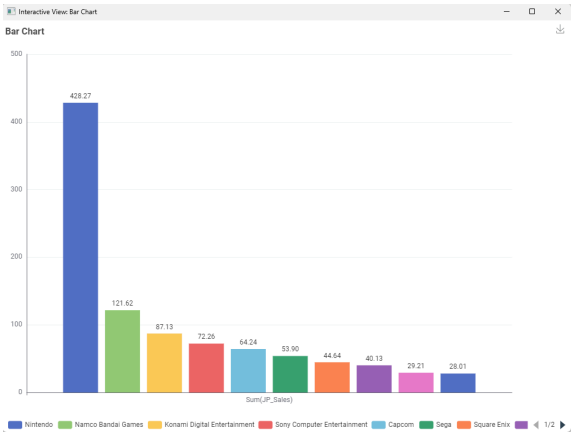


Figura 8: *Vendas a Nível Japonês*

2.2.2.3. Analise das Vendas por *Genre*

Na análise global dos gráficos, uma tendência bastante evidente é o domínio dos jogos de ação em termos de vendas. Estes jogos geralmente concentram-se na aventura e na exploração de vastos mapas. Entre os exemplos mais proeminentes estão títulos como *Grand Theft Auto V*, *The Legend of Zelda: Ocarina of Time* e *Assassin's Creed III*. Além disso, os jogos desportivos também desfrutam de uma popularidade significativa, incluindo aqueles mencionados anteriormente, assim como os *shooters*, que se destacam pelo seu foco em armas de fogo e combate. Algumas séries bem conhecidas nessa categoria incluem *Call of Duty* e *Battlefield*.

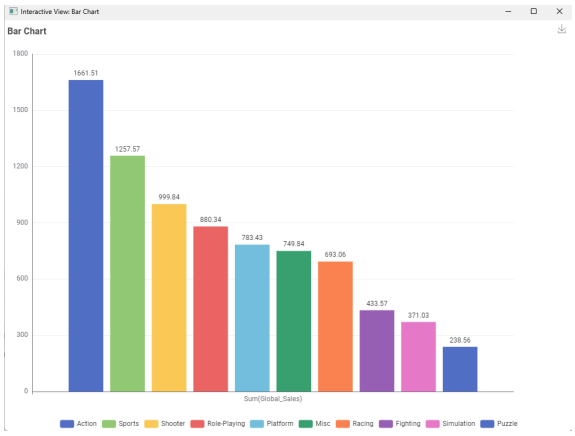


Figura 9: *Vendas a Nível Global*

Já nos gráficos da Europa e da América do Norte, notamos que não existem diferenças muito significativas. Os jogos de Ação, Desporto e *Shooter* continuam a ocupar o top 3, mas é interessante notar que enquanto que o público europeu parece preferir estilos de jogos como Corrida e de estilos sortidos (*Misc*), os americanos preferem *platformers* acima dos anteriores. Jogos de Plataforma (ou *platformers*) são jogos em que os jogadores controlam um personagem que deve atravessar níveis cheios de obstáculos, inimigos e quebra-cabeças. Destacam-se nesse meio jogos como: *Super Mario Bros.* e *Sonic The Hedgehog*.

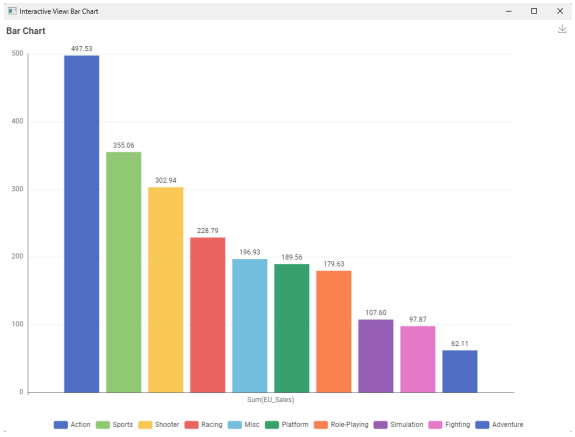


Figura 10: *Vendas a Nível Europeu*

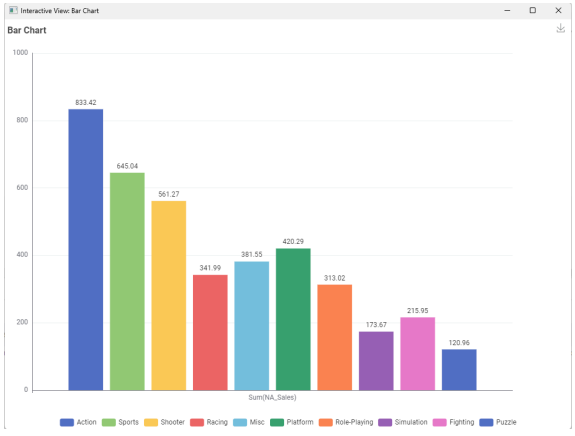


Figura 11: *Vendas a Nível Norte-Americano*

Por fim, o Japão apresenta novamente um gráfico muito diferente do resto do mundo, com uma preferência especial por jogos *Role-Playing* ou RPGs. Estes jogos colocam os jogadores em mundos de fantasia, fazendo-os explorar ambientes, interagir com personagens não jogáveis, completar missões e combater inimigos. Geralmente, os jogadores podem personalizar os seus personagens, adquirir habilidades e equipamentos, e tomar decisões que afetam a história do jogo. Entre jogos destes estilos alguns exemplos são:

Pokémon, *Final Fantasy* e *Dragon Quest* (séries que já vimos que têm forte presença na "terra do sol nascente"). Além disto nota-se também uma falta de jogos *Shooters* (populares em outras regiões do mundo) e que mesmo não havendo nenhuma franquia de desporto dominante no Japão, jogos desportivos ainda conseguem vender mais do que *Platformers*.

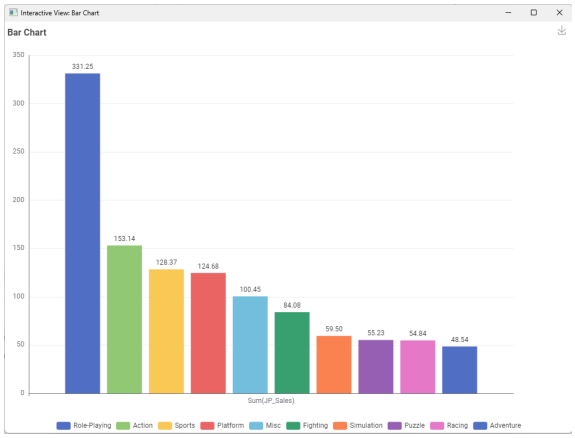


Figura 12: *Vendas a Nível Japonês*

2.2.2.4. Analise das Vendas por *Year*

Em termos de vendas globais, o ano de 2008 destacou-se como o mais significativo, seguido por 2009 e 2010. Durante esse período, a *Nintendo* (estatisticamente, como verificado, a maior *publisher*) lançou títulos de destaque como *Super Smash Bros. Brawl* (2008) e *New Super Mario Bros. Wii* (2009), enquanto que a *Electronic Arts* (segunda "maior" empresa) trouxe *FIFA 09* (2008) e *FIFA 10* (2009), contribuindo para o sucesso desses anos em termos de vendas de jogos. Adicionalmente, a *Activision* (que se encontra sempre no pódio) lançou *Call of Duty: World at War* em 2008 e *Call of Duty: Modern Warfare 2* em 2009, ampliando ainda mais o impacto desses anos no mercado global de videojogos.

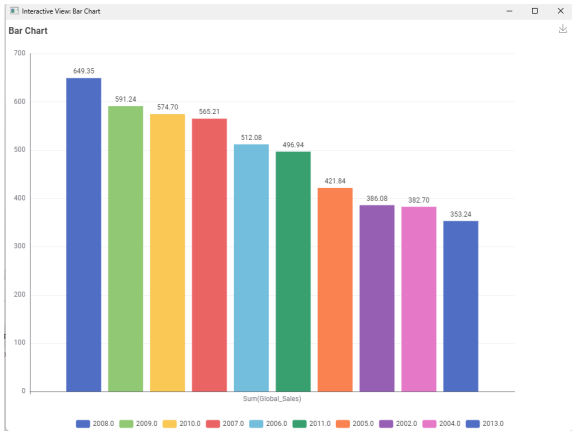


Figura 13: *Vendas a Nível Global*

Esta tendência de vendas significativas verifica-se não apenas a nível global, mas também em regiões específicas como Europa e América do Norte. Em todos estes mercados, os anos de 2008, 2009, 2010 destacaram-se como períodos significativos em termos de vendas de jogos, impulsionados pelo lançamento de títulos de sucesso.

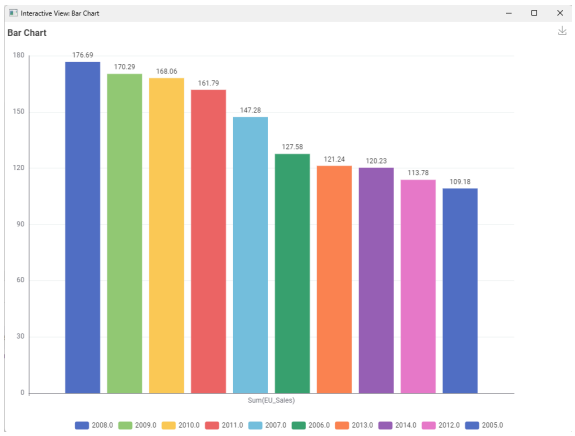


Figura 14: *Vendas a Nível Europeu*

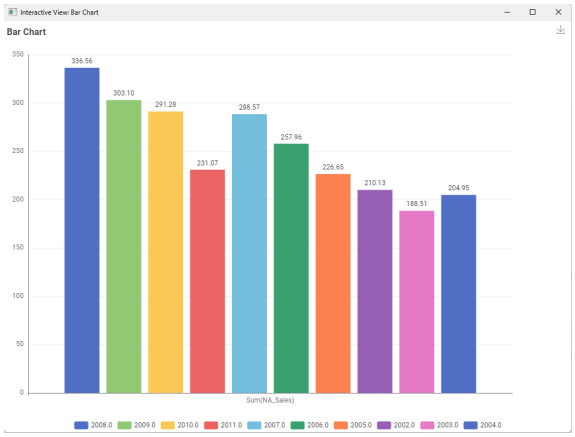


Figura 15: *Vendas a Nível Norte-Americano*

Já a nível japonês, verificamos que os anos onde foram vendidos mais videojogos foram os de 2006, 2010 e 2007. Durante esses anos, títulos como *Wii Sports* (2006), *Super Mario Galaxy 2* (2010) e *Wii Fit* (2007) da *Nintendo* foram particularmente populares no mercado japonês.

Em conclusão, 2010 foi, para todos os mercados, o melhor ano em termos de vendas de jogos, com lançamentos de sucesso e alto impacto tanto da *Nintendo*, *Electronic Arts* quanto da *Activision*.

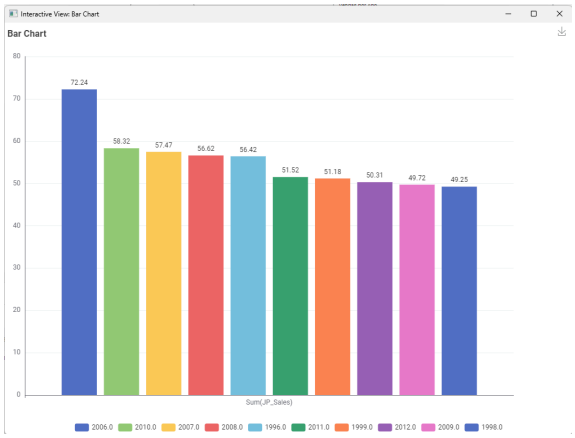


Figura 16: *Vendas a Nível Japonês*

2.3. Preparação de dados

Na etapa de preparação dos dados, realizamos uma análise detalhada dos atributos contidos no *dataset*, efetuando as modificações necessárias para garantir que os valores estejam em conformidade com as expectativas. Durante este processo, realizamos as seguintes alterações:

- **Rank e Other_Sales**: O *rank* existe para mostrar os jogos mais vendidos a nível global. Este tipo de informação é desnecessária, e portanto foi removida. Por outro lado, como não estamos a estudar o *Other_Sales*, também o removemos. Assim, deixamos apenas as vendas a nível global, europeu, americano e japonês.
- **Categorias para número** : Para facilitar a modelação em regressão todos os atributos que não eram valores numéricos foram convertidos para tal. Isto inclui: *Name*, *Platform*, *Genre* e *Publisher*.
- **Normalização** : Dado que os valores de vendas podem variar significativamente de uma região para outra e entre diferentes jogos, aplicamos uma normalização aos dados das vendas. Isto é importante para não distorcer os modelos de regressão com valores extremamente altos ou baixos. Utilizamos a técnica de normalização Min-Max para escalar os dados de vendas, de modo a que todos os valores fiquem entre 0 e 1. Esta abordagem ajuda a manter a estrutura proporcional dos dados enquanto reduz a sensibilidade dos algoritmos de regressão a *outliers*.

Além disso, é importante ressaltar que linhas com valores vazios foram removidas, garantindo que o modelo de regressão seja treinado apenas com dados completos e confiáveis. A preparação de dados é uma etapa crucial para assegurar a precisão das previsões futuras.

2.4. Modelação

2.4.1. Modelação com todos os Atributos

A modelação do problema começou pelo uso de nodos de aprendizagem de regressão sobre os atributos presentes no *dataset*. Infelizmente, não foi possível utilizar o atributo **Franchise** criado na preparação de dados para exploração, pelo o facto de não conseguirmos garantir um *franchise* para cada grupo de jogos presentes no *dataset*.

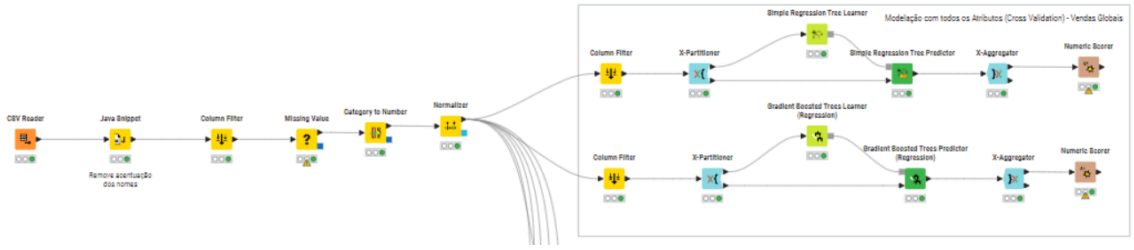


Figura 17: *Modelação com cross-validation*

Figura 18: *Modelação sem cross-validation*

Para avaliar os modelos, empregamos tanto *Cross-Validation* quanto *Hold-out Validation*. Estas abordagens permitem-nos avaliar o desempenho do modelo de diferentes maneiras e determinar a mais adequada. Em ambas as técnicas, mantivemos todos os atributos, exceto o *rank*.

Os algoritmos de aprendizagem utilizados foram **Simple Regression Tree** e **Gradient Boosted Trees (Regression)**. Para os algoritmos com *Cross-Validation* obtivemos os seguintes resultados:

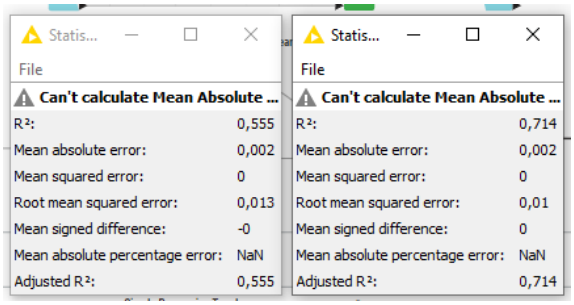


Figura 19: *Simple Regression Vs. Gradient Boosted Trees (Global)*

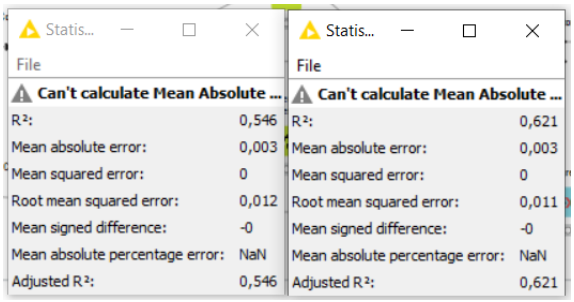


Figura 20: *Simple Regression Vs. Gradient Boosted Trees (Europa)*

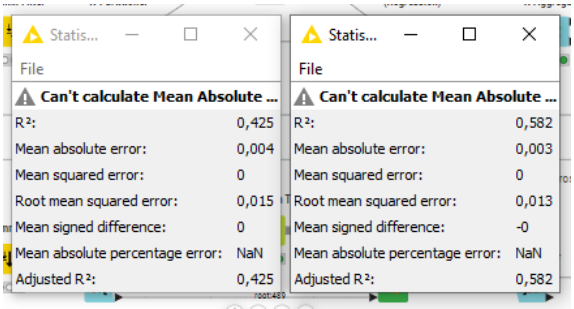


Figura 21: *Simple Regression Vs. Gradient Boosted Trees (America)*

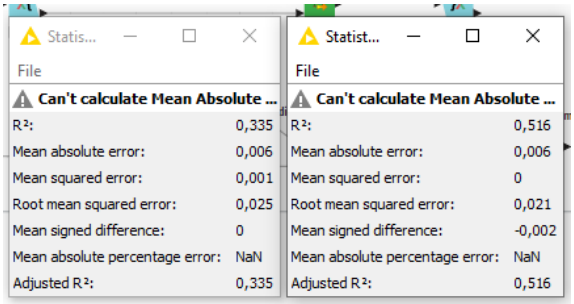


Figura 22: *Simple Regression Vs. Gradient Boosted Trees (Japão)*

Por outro lado, os resultados que obtivemos com *Hold-out Validation* foram os seguintes:

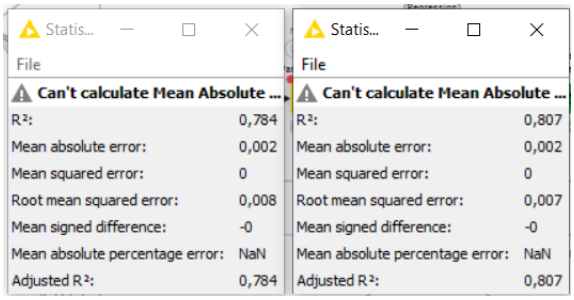


Figura 23: *Simple Regression Vs. Gradient Boosted Trees (Global)*

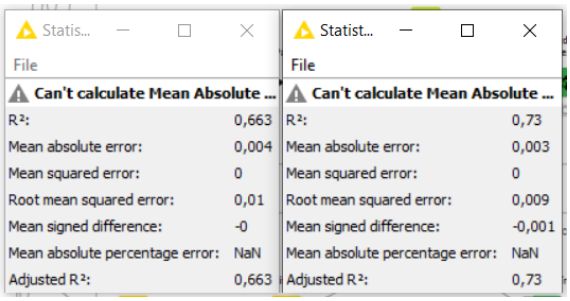


Figura 24: *Simple Regression Vs. Gradient Boosted Trees (Europa)*

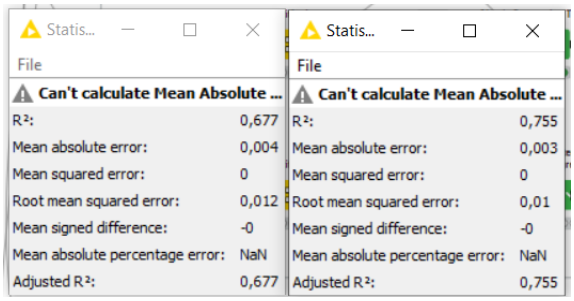


Figura 25: *Simple Regression Vs. Gradient Boosted Trees (America)*

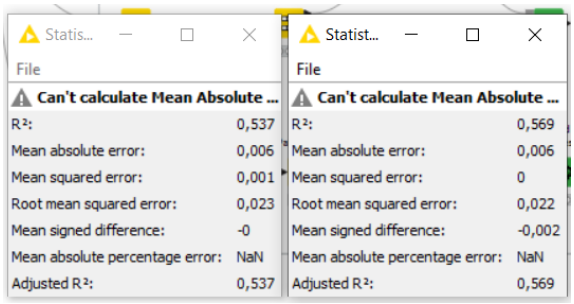


Figura 26: *Simple Regression Vs. Gradient Boosted Trees (Japão)*

O algoritmo mais eficaz é o **Simple Regression Tree**, obtendo os melhores valores em todas as métricas calculadas pelo *Numeric Scorer*. Adicionalmente, os valores obtidos em *Hold-out* são muito melhores que os de *Cross-Validation*, apesar de bastante próximos entre si.

Analisando estes resultados, em especial o $\mathbf{R^2}$, verificamos que a nível global existe uma variação entre 0.7 e 0.8. Isto sugere que o modelo de análise está bem ajustado e é capaz de prever com precisão as tendências de vendas em escala global.

A nível europeu e americano, os valores encontram-se entre 0.5 e 0.6 utilizando *cross-validation* e entre 0.6 a 0.7 utilizando *hold-out*. Embora não seja tão alto quanto o R^2 global, o modelo revela capacidade razoável de prever as vendas de jogos nessas regiões.

Por fim, a nível nipónico, obtivemos uma gama de valores entre 0.4 e 0.5, podendo indicar que existem outros fatores importantes não considerados pelo modelo ou que a relação entre os atributos e as vendas de jogos no mercado japonês é mais complexa do que nas outras regiões.

2.4.2. Modelação com *Feature Selection*

A seleção de atributos é uma fase crítica na construção de modelos preditivos eficientes, permitindo identificar as variáveis mais influentes para a previsão das vendas de jogos em diferentes regiões. Para este fim utilizamos uma abordagem iterativa dentro do KNIME, empregando os nodos *Feature Selection Loop Start*, *Feature Selection Loop End* e *Feature Selection Filter*. Estes nodos facilitam a execução de um processo de seleção de atributos robusto, automatizado e orientado por dados.

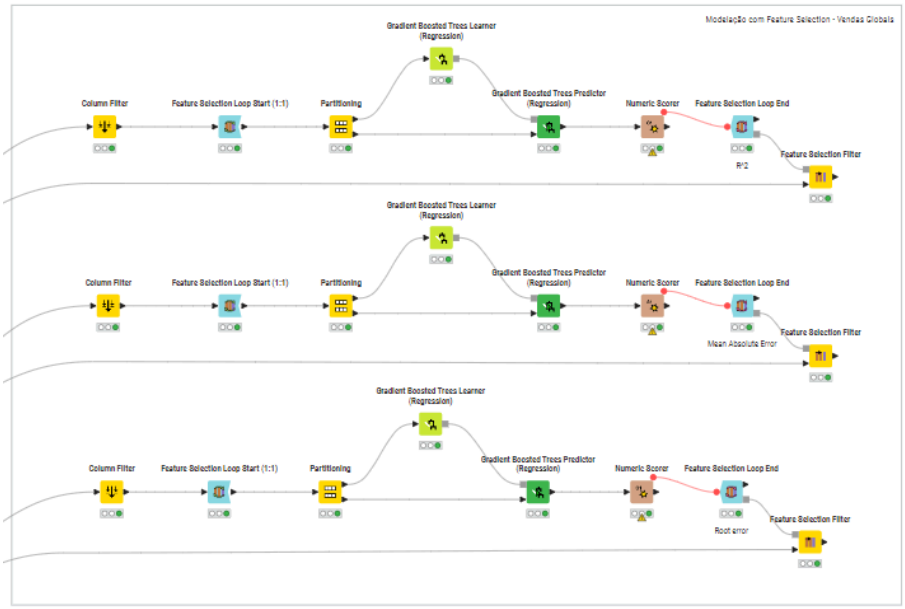


Figura 27: *Feature Selection*

Foram realizados 3 processos de *feature selection* para cada uma das regiões, avaliando 3 métricas diferentes, sendo elas o R^2, o *Mean Absolute Error* (MAE) e o *Root Mean Squared Error* (RMSE).

A nível global, processos onde as métricas de avaliação foram o R^2 e o *Root Mean Squared Error* (RMSE) os atributos que obtiveram um melhor resultado foram: *Name*, *Platform*, *Genre* e *Publisher*, obtendo R^2 o valor de **0.823** (algo ligeiramente superior ao obtido na modelação anterior). Por sua vez o *Mean Absolute Error* e o *Root Mean Squared Error* situam-se nos valores **0.006** e **0.017**, respetivamente, sugerindo que as previsões são muito próximas aos valores reais.

A nível europeu obtivemos um R^2 de **0.742** para os atributos *Name*, *Platform* e *Year*, demonstrando que estas variáveis são as mais influentes. O *Mean Absolute Error* e o *Root Mean Squared Error* foram **0.006** e **0.018**, indicando baixa discrepância.

A nível americano o valor do R^2 foi **0.712** para os atributos *Name*, *Platform*, *Year*, *Genre* e *Publisher* (assim como para as vendas globais). O *Mean Absolute Error* e o *Root Mean Squared Error* foram **0.007** e **0.02**, valores ligeiramente superiores aos anteriores.

No Japão, o R^2 foi **0.569** para os atributos *Name*, *Platform*, *Year*, *Genre* e *Publisher*. O MAE e o RMSE foram, respetivamente, **0.009** e **0.034**. Estes foram os valores menos favoráveis registados até ao momento.

É importante salientar que apesar dos atributos supra referidos terem sido aqueles que nos deram melhores resultados, outros atributos também obtiveram valores bastante próximos a estes, não sendo definitivamente os melhores. Não obstante, em outras circunstâncias, talvez outros atributos tenham um melhor desempenho.

2.4.3. Modelação com Redes Neuronais

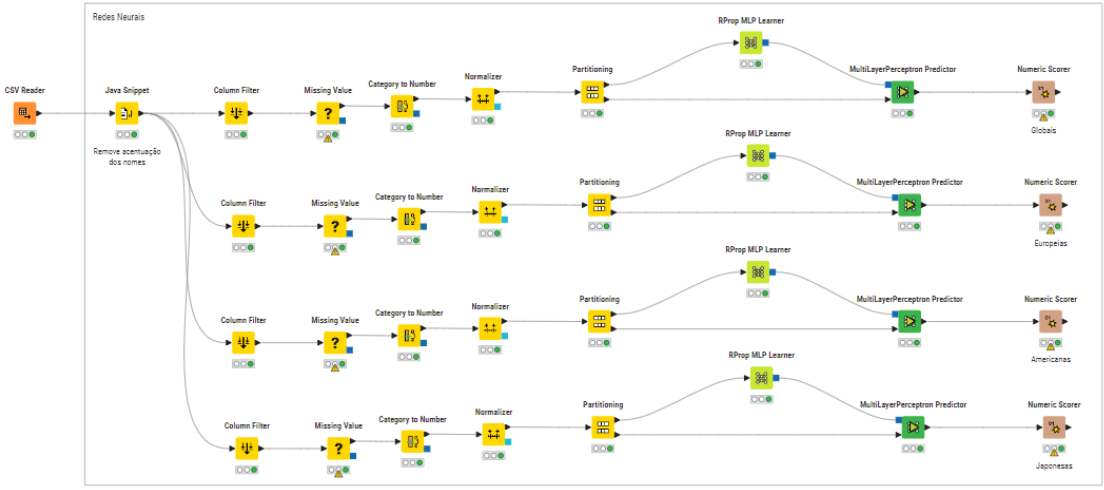


Figura 28: *Redes Neuronais*

Redes Neuronais são um tipo de modelação inspirado no funcionamento de um cérebro humano. São compostas por camadas de neurónios conectados através de sinapses que recebem informação, processam essa informação alterando o valor da sinapse por onde receberam a informação e produzem uma saída que é transmitida para a próxima camada. Assim, as redes neuronais conseguem aprender a partir de exemplos onde identificam padrões e relações entre os atributos que recebem.

Para explorar o potencial das redes neuronais, optamos por utilizar o nodo **RProp MLP Learner** como algoritmo de rede neuronal. Este nodo é especificamente projetado para implementar redes neuronais do tipo **Perceptron Multicamadas** (**MLP**) com o algoritmo de treino **RProp** (*Resilient Propagation*). RProp é uma variante do *backpropagation* que ajusta os pesos dos neurónios eficientemente, sendo menos susceptível a ficar preso em mínimos locais do que métodos tradicionais de gradiente descendente.

O nodo foi configurado para utilizar um número máximo de iterações de 100 e 1 camada oculta, com 10 neurónios contidos. Os resultados obtidos foram os seguintes:

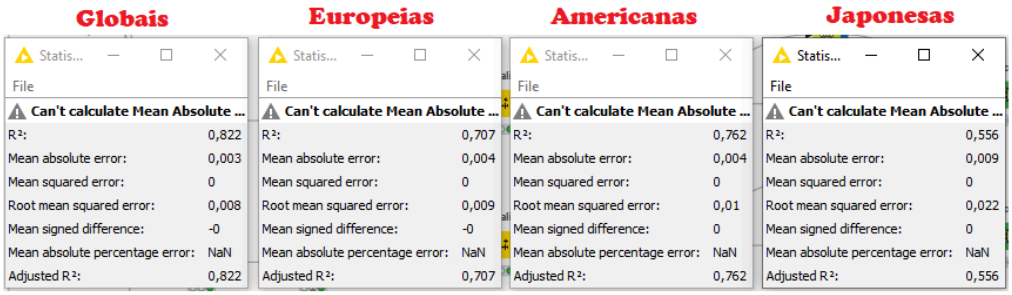


Figura 29: *Redes Neuronais - resultados*

Após análise dos resultados, percebe-se que os valores de R^2 para as vendas globais e americanas são superiores aos obtidos anteriormente, excetuando os das europeias e japonesas, levemente inferiores. Enquanto isso, os valores dos erros continuam praticamente inalterados, indicando que os resultados se encontrem próximos dos valores reais.

2.5. Avaliação

Através dos processos de seleção de atributos foi possível identificar as variáveis mais significativas que influenciam as vendas em diferentes mercados. Esta etapa foi crucial, pois permitiu a simplificação dos modelos, através da redução da quantidade de dados processados, o que permitiu focar os fatores mais

impactantes, melhorando significativamente a precisão das previsões. Os resultados desta abordagem mostraram que atributos como *Name, Platform, Genre, e Publisher* são determinantes para o sucesso das vendas, com a seleção de atributos levando a um aumento substancial nos valores de R^2, alcançando até 0.823, globalmente.

Finalmente, o mercado japonês revelou-se o mais desafiante, onde os valores de R^2 foram mais baixos, ressaltando a possibilidade de variáveis externas ou fatores culturais específicos não completamente capturados pelos modelos atuais. Sugere-se, assim, uma área para investigação futura, que inclua dados adicionais.

3. 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.

3.1. 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. Isto envolve identificar o problema, determinar os objetivos do projeto e estabelecer um critério de aceitação para definir o projeto como entregue.

3.1.1. 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, 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.

3.1.2. Objetivo

O objetivo estabelecido pela equipa de trabalho é o seguinte:

1. Desenvolver um modelo que determine se um paciente sofre de doença hepática com base em marcadores bioquímicos.

3.1.3. 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, 70%) na previsão de doença hepática com base nos marcadores bioquímicos fornecidos.

3.2. 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
  - **Selector**: Indicador de doença hepática
- **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").
- **Variável Alvo:** A variável alvo é representada pela coluna *Selector*, que indica se o paciente sofre de doença hepática ("1=liver disease") ou não ("2=no liver disease" ou "2=without liver disease").

3.2.1. Preparação dos Dados para Exploração

Durante a etapa inicial de preparação dos dados para exploração, foi realizada uma análise minuciosa dos atributos do *dataset*, 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:

- **TB (#1)**: Dado que este atributo deve representar *Total Proteins*, optamos por converter o seu nome para **TP**. Para alcançar este objetivo, utilizamos o nodo *Column Renamer*.
- **Gender**: Com o intuito de homogeneizar este atributo, dado que existem variações no formato de género (por exemplo, "female" e "Female"), optamos por transformar estes dados num simples ***M*** para *Male* e ***F*** para *Female*. Para tal, utilizamos o nodo *String Manipulation*.
- **birth_year,birth_date, id_code , birth_month, CHOL**: Os atributos correspondentes ao ano de nascimento e 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*.

O atributo **CHOL** (que representa o nível de colesterol num paciente) revela-se inútil, por apenas conter valores de 0 (o que biologicamente é impossível). Assim, o grupo resolveu remover o atributo por completo, utilizando o nodo *Column Filter*.

- **TB, DB, TP , ALB, AG_RATIO e BILmg**: Estes atributos foram convertidos para valores numéricos, simplificando a análise estatística e modelação. Esta conversão foi realizada utilizando o nodo *String to Number*.
- **Selector**: Inicialmente, este atributo possuía 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: **Doente** e **Não Doente**.

  Adicionalmente, foi utilizado o nodo *One to Many* para separar a variável alvo em duas colunas diferentes : **Doente** e **Não Doente**.

3.2.2. Exploração inicial dos dados

- **Selector:** A distribuição da variável alvo apresenta um desequilíbrio significativo, com aproximadamente 71% dos pacientes sendo diagnosticados como doentes, enquanto apenas 29% são classificados como não doentes. Esta disparidade pode impactar negativamente a eficácia dos modelos de previsão, especialmente em termos de precisão. O desafio reside no facto de que o modelo pode tender a favorecer a classe maioritária, resultando numa menor capacidade de identificar corretamente os casos da classe minoritária. Este desequilíbrio exige técnicas de balanceamento de dados ou métricas de avaliação alternativas para garantir que o modelo seja capaz de generalizar de maneira adequada e fornecer previsões precisas para ambas as classes.

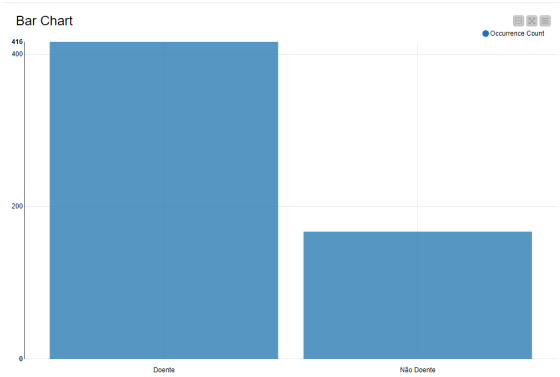


Figura 30: Distribuição das idades

- **Age:** Ao analisar o nosso conjunto de dados, notamos que as idades dos pacientes variam entre 4 e 90 anos, sendo que qualquer paciente com idade superior a 89 é registado como tendo 90 anos (segundo a fonte do *dataset*). Ao representarmos graficamente a presença de doenças hepáticas em relação à idade, observamos um pico significativo na faixa etária de 45 a 60 anos.

  Esta distribuição sugere que a ocorrência de doenças hepáticas é mais frequente em adultos e idosos. Por outro lado, notamos que a presença de doenças hepáticas entre o público mais jovem é menos comum. Esta análise leva-nos a concluir que o tipo de doença em estudo, nesse caso hepática, tem uma maior incidência em pessoas de faixas etárias mais avançadas.

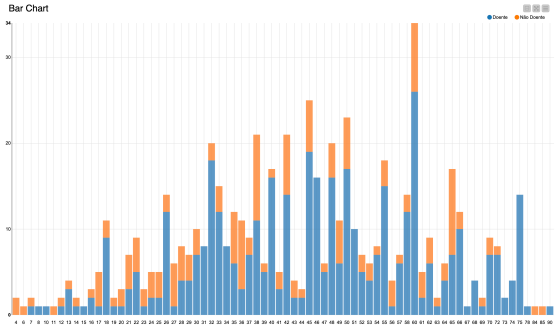


Figura 31: Distribuição das idades

Ao utilizar o nodo *Numeric Binner*, pudemos simplificar a gama de valores no eixo x do gráfico, dividindo as idades em faixas etárias: Jovens (até 19 anos), Adultos (20 a 59 anos) e Idosos (60 a 90 anos). Isto permitiu-nos obter uma visão mais clara da distribuição das doenças hepáticas em diferentes grupos etários.

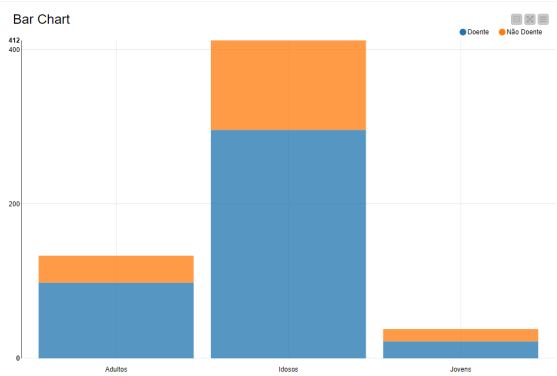


Figura 32: Distribuição das idades em blocos

Com base nos dados da distribuição por faixa etária (Figura 32), podemos inferir que a taxa de pacientes com doença hepática aumenta significativamente na faixa etária dos Idosos. Nota-se que nesse intervalo etário a incidência da doença é consideravelmente mais alta em comparação com outras faixas etárias. Por outro lado, nas faixas etárias dos Jovens e Adultos, a taxa de pacientes com a doença hepática é mais equiparável à taxa daqueles que não a possuem, e o número de pacientes é menor.

- **Gender:** No conjunto de dados, existe um desequilíbrio evidente entre o número de homens e mulheres. Existem 441 homens e apenas 142 mulheres.

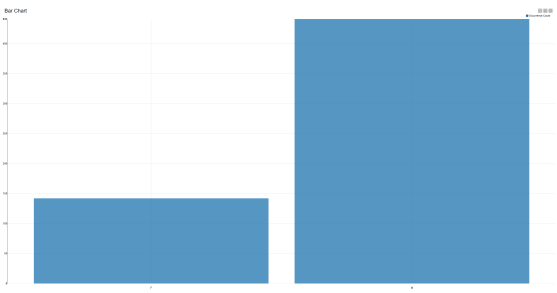


Figura 33: Distribuição dos Géneros analisados

Na Figura 34, é possível concluir que os homens apresentam uma probabilidade maior de ter doenças hepáticas do que as mulheres. Cerca de 324 em 441 homens ($\simeq 73\%$) possuem a condição, enquanto aproximadamente 92 em 142 mulheres, o que corresponde a cerca de 65%, são afetados pela doença.

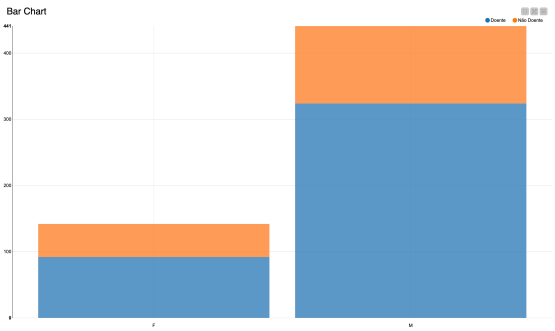


Figura 34: Distribuição dos Géneros

- **Marcadores Bioquímicos:** Até o momento do *checkpoint* realizado internamente na Unidade Curricular, a nossa exploração de dados dos marcadores bioquímicos estava restrita a gráficos de barras, o que dificultava a interpretação de valores significativamente discrepantes. Consequentemente, optamos por incorporar dados reais, cujas fontes serão devidamente citadas na bibliografia, com o intuito de aprimorar a nossa análise e conclusões. Para estabelecer comparações com valores de referência reais, incluindo mínimos, máximos e médios, utilizamos o nodo de estatísticas disponível (*statistics*).

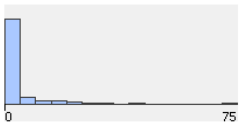
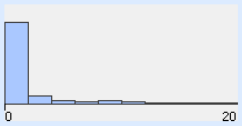
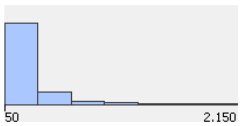
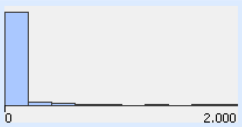
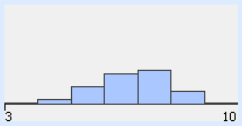
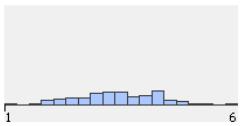
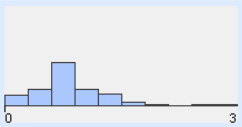
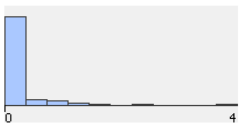
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
TB	0.4	3.2988	?	75	6.2095	4.9075	37.1638	0	0	0	
DB	0.1	1.4861	?	19.7	2.8085	3.2124	11.3525	0	0	0	
Alkphos	63	290.5763	?	2,110	242.938	3.7651	17.7528	0	0	0	
Sgpt	10	80.7136	?	2,000	182.6204	6.5492	50.5794	0	0	0	
Sgot	10	109.9108	?	4,929	288.9185	10.5462	150.9199	0	0	0	
TP	2.7	6.4832	?	9.6	1.0855	-0.2857	0.233	0	0	0	
ALB	0.9	3.1419	?	5.5	0.7955	-0.0437	-0.3879	0	0	0	
AG_Ratio	0.3	0.9491	?	2.8	0.3201	1.0043	3.2785	4	0	0	
BILmg	0.0234	0.1911	?	4.386	0.3581	5.0668	40.1222	22	0	0	

Figura 35: Estatísticas dos Marcadores Bioquímicos

- **TB:** Em relação ao valor da bilirrubina total, os valores que temos como referência são os seguintes :
  - **Valor Mínimo :** 0.3 mg/dL
  - **Valor Máximo :** 1.0 mg/dL
  - **Valores Críticos :** >12 mg/dL

  Após analisarmos o *dataset* e suas estatísticas correspondentes, constatamos que o valor mínimo da bilirrubina total é de 0.4 mg/dL, mantendo-se dentro dos padrões considerados saudáveis. No entanto, o valor máximo é alarmantemente alto, atingindo 75 mg/dL, o que representa uma discrepância significativa em relação ao limite crítico inicial de 12 mg/dL e classifica-se como um *outlier*. Além disso, observamos que o valor médio está acima do esperado, situando-se em cerca de 3.3 mg/dL. O desvio padrão é próximo de zero, indicando uma baixa dispersão dos valores deste atributo e sugerindo que os pacientes apresentam níveis pouco saudáveis de bilirrubina total. É importante ressaltar que este atributo não possui *missing values*.

- **DB:** Em relação ao valor da bilirrubina direta, os valores que temos como referência são os seguintes :
  - **Valor Mínimo :** 0.1 mg/dL

- **Valor Máximo :** 0.3 mg/dL

  O valor mínimo da bilirrubina direta é de 0.1 mg/dL, permanecendo dentro dos padrões considerados saudáveis. No entanto, o valor máximo é novamente alarmantemente alto, alcançando 19.7 mg/dL, distanciando-se significativamente do limite de 0.3 mg/dL e sendo identificado como um *outlier*. Além disso, o valor médio está acima do esperado, situando-se em cerca de 1.5 mg/dL. O desvio padrão, em comparação com o atributo anterior, está ainda mais próximo de zero, indicando uma baixa dispersão dos valores de bilirrubina direta e corroborando a presença predominante de valores possíveis, porém anormais, neste atributo. Vale ressaltar que este atributo também não possui *missing values*.

- **Alkphos:** Em relação ao valor da fosfatase alcalina, os valores que temos como referência são os seguintes :
  - **Valor Mínimo :** 44 IU/L
  - **Valor Máximo :** 147 IU/L

O valor mínimo da fosfatase alcalina é de 63 IU/L, mantendo-se dentro dos padrões considerados saudáveis. No entanto, o valor máximo é novamente alarmantemente alto, atingindo 2110 IU/L, distanciando-se significativamente do limite de 147 IU/L e sendo identificado como um *outlier*. Além disso, o valor médio está acima do esperado, situando-se em cerca de 290 IU/L. O desvio padrão, em comparação com o atributo anterior, é extremamente distante de zero, indicando uma alta dispersão dos valores do marcador. Esta dispersão sugere uma variação significativa nos níveis de fosfatase alcalina entre os pacientes, possivelmente refletindo uma condição clínica heterogénea. É importante ressaltar que este atributo também não possui *missing values*.

- **Sgpt:** Em relação ao valor da alanina aminotransferase, os valores que temos como referência são os seguintes :
  - **Valor Mínimo :** 7 U/L
  - **Valor Máximo :** 56 U/L

O valor mínimo da alanina aminotransferase é de 10 U/L, permanecendo dentro dos padrões considerados saudáveis. No entanto, o valor máximo é novamente alarmantemente alto, chegando a 2000 U/L, o que representa uma discrepância significativa em relação ao limite de 56 U/L e é identificado como um *outlier*. Além disso, o valor médio está acima do esperado, situando-se em cerca de 80 U/L. Embora o desvio padrão esteja menos distante de zero em comparação com o atributo anterior, ainda está consideravelmente acima de 0, indicando uma alta dispersão dos valores do marcador. Esta dispersão sugere uma variação significativa nos níveis de alanina aminotransferase entre os pacientes, possivelmente refletindo uma condição clínica heterogénea. É importante ressaltar que este atributo também não possui *missing values*.

- **Sgot:** Em relação ao valor da aspartato aminotransferase, os valores que temos como referência são os seguintes :
  - **Valor Mínimo :** 0 U/L
  - **Valor Máximo :** 35 U/L

O valor mínimo da aspartato aminotransferase é de 10 U/L, permanecendo dentro dos padrões considerados saudáveis. No entanto, o valor máximo é novamente alarmantemente alto, chegando a 4929 U/L, o que representa uma discrepância significativa em relação ao limite de 35 U/L e é identificado como um *outlier*. Além disso, o valor médio está acima do esperado, situando-se em cerca de 110 U/L. O desvio padrão, em comparação com o atributo anterior, é extremamente distante de zero, indicando uma alta dispersão dos valores do marcador. Esta dispersão sugere uma variação significativa nos níveis de aspartato aminotransferase entre os pacientes, possivelmente refletindo uma condição clínica heterogénea. É importante ressaltar que este atributo também não possui *missing values*.

- **TP:** Em relação ao valor de proteínas totais, os valores que temos como referência são os seguintes :
  - **Valor Mínimo :** 6 g/dL
  - **Valor Máximo :** 8.3 g/dL

O valor mínimo de proteínas totais é de 2.7 g/dL, situando-se abaixo dos padrões considerados saudáveis. Além disso, o valor máximo é bastante próximo ao valor máximo saudável, alcançando os 9.6 g/dL. Adicionalmente, o valor médio encontra-se nos 6.5 g/dL situando-se numa zona perfeitamente saudável. O desvio padrão, encontra-se muito próximo de 0, indicando uma baixíssima dispersão dos valores do marcador. É importante ressaltar que este atributo também não possui *missing values*.

- **ALB:** Em relação ao valor de albumina, os valores que temos como referência são os seguintes :
  - **Valor Mínimo :** 3.4 g/dL
  - **Valor Máximo :** 5.4 g/dL

O valor mínimo de albumina é de 0.9 g/dL, situando-se abaixo dos padrões considerados saudáveis. Além disso, o valor máximo é bastante próximo ao valor máximo saudável, alcançando os 5.5 g/dL. Adicionalmente, o valor médio encontra-se nos 3.12 g/dL situando-se numa zona próxima à saudável. Novamente, o desvio padrão, encontra-se muito próximo de 0, indicando uma baixíssima dispersão dos valores do marcador. É importante ressaltar que este atributo também não possui *missing values*.

- **AG_Ratio:** Em relação ao valor da proporção de albumina e globulina, os valores que temos como referência são os seguintes :
  - **Valor Mínimo :** 1.1
  - **Valor Máximo :** 2.5

O valor mínimo da proporção de albumina e globulina é de 0.3, encontrando-se abaixo dos padrões considerados saudáveis. Além disso, o valor máximo é bastante próximo ao valor máximo saudável, alcançando os 2.8. Adicionalmente, o valor médio encontra-se nos 0.95 situando-se numa zona saudável. Novamente, o desvio padrão, encontra-se muito próximo de 0, indicando uma baixíssima dispersão dos valores do marcador. É importante ressaltar que este atributo possui 4 *missing values*.

- **BILmg:** Não pudemos obter informações confiáveis sobre este atributo, o que nos impede de realizar qualquer comparação significativa. No entanto, observamos a presença de 22 *missing values* e uma dispersão relativamente baixa nos valores disponíveis.

3.3. Preparação dos Dados

Na etapa inicial de preparação dos dados, realizamos uma análise detalhada dos atributos contidos no *dataset*, efetuando as modificações necessárias para garantir que os valores estejam em conformidade com as expectativas. Durante este processo, identificamos e corrigimos os seguintes atributos que exigiam ajustes:

- **Gender**: Convertido de *String* para *Integer* para permitir o uso subsequente na modelação, especialmente em algoritmos de *clustering*, que requerem valores numéricos. Para essa conversão, utilizou-se o nodo *Category to Number*.
- **Alkphos, Sgpt, Sgot e BILmg** : Por apresentarem o maior desvio padrão (maior dispersão de valores) com a exceção do atributo BILmg, foi necessário aplicar o nodo *Numeric Outliers* para lidar com valores que estavam significativamente fora da faixa esperada (ou seja, valores máximos). Todos os *outliers* foram substituídos por *missing values*.
- **Alkphos, Sgpt, Sgot, AG_Ratio e BILmg** : Sendo os únicos atributos com *missing values*, foi necessário resolver esse problema. Por anteriormente terem sido (maioritariamente) valores extremamente fora do esperado, optou-se por substituir todos os *missing values* pelo valor máximo observado em cada atributo. No caso do atributo *AG_Ratio*, que tinha 4 missing values, as linhas correspondentes foram removidas.

  Finalmente, no caso do atributo BILmg, que possuía cerca de 22 *missing values*, as linhas correspondentes foram substituidas pelo valor médio do atributo.

Para a modelação com atributos agrupados, optamos por criar *bins* em alguns atributos do *dataset*. Os atributos que foram agrupados são os seguintes:

- **Age** : O atributo, tal como na exploração de dados, foi dividido em três *bins* :
  - **Jovens** : $[4, 19]$ anos
  - **Adultos** : $[20, 59]$ anos
  - **Idosos** : $[60, 90]$ anos

  Para tal efeito, foi utilizado o nodo *Numeric Binner*.
- **TB, DB, Alkphos, Sgpt, Sgot, TP, ALB, AG_Ratio** : Todos os marcadores bioquímicos (com a exceção do atributo BILmg) foram organizados em *Healthy* e *Unhealthy*, com base nos intervalos que consideramos na fase de exploração de dados. Para isso foi necessário utilizar o nodo *Numeric Binner* e *Category to Number*.

Além disso, é importante ressaltar que a preparação de dados incorpora a etapa de preparação para exploração de dados. Esta fase também é responsável por garantir que os dados estejam prontos para interpretação, incluindo a limpeza de dados e padronização de formatos, facilitando assim a análise e extração de *insights*.

3.4. Modelação

3.4.1. Modelação de Controlo

Com o objetivo de realizar uma modelação inicial (de controlo), o grupo recorreu a uma modelação com apenas os dados resultantes da preparação de dados.

Foi realizada uma modelação com *cross-validation* (recorrendo aos nodos *X-Partitioner* e *X-Aggregator*) e uma com *hold-out validation*, *i.e*, recorrendo a um nodo *partitioner*.

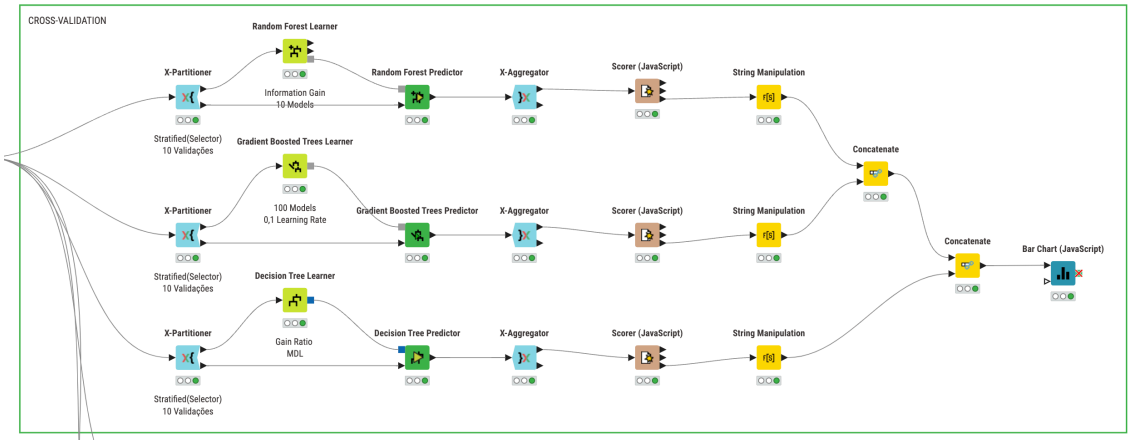


Figura 36: Modelação de Controlo com *Cross-Validation*

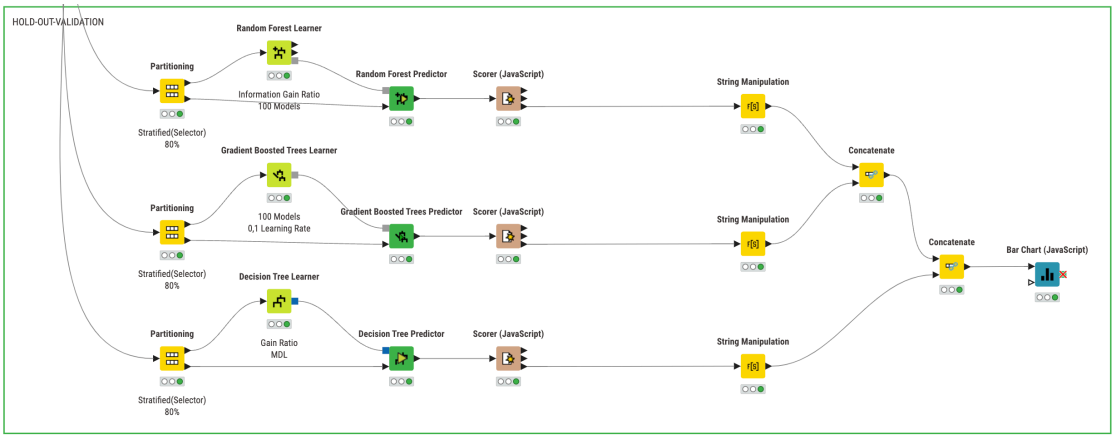


Figura 37: Modelação de Controlo com *Hold-out validation*

Para abordar o problema de classificação presente no *dataset*, exploramos o poder de três algoritmos de aprendizagem: *Random Forest*, *Gradient Boosting* e *Decision Tree*.

No nodo *X-Partitioner*, foi empregada a opção de *stratified sampling* na coluna alvo *selector* para assegurar proporções consistentes de cada tipo (Doente ou não Doente).

Scorer View

Confusion Matrix

	Doente (Predicted)	Não Doente (Predicted)	
Doente (Actual)	71	12	85.54%
Não Doente (Actual)	17	16	48.48%
	80.68%	57.14%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
Doente	71	17	16	12	85.54%	80.68%	85.54%	48.48%	83.04%
Não Doente	16	12	71	17	48.48%	57.14%	48.48%	85.54%	52.46%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
75.00%	25.00%	0.357	87	29

Figura 38: *Scorer* Modelação de Controlo (*Gradient Boosted*)

Os resultados mais favoráveis foram alcançados através do uso da validação *Hold-Out*, isto é, empregando o nodo *Partitioner*. O algoritmo *Gradient Boosted* apresentou uma precisão modesta de 75%, superando os outros dois algoritmos avaliados. No entanto, o coeficiente de *Cohen's kappa* registou um valor de 0.357, consideravelmente abaixo do ideal. Esta discrepância pode ser atribuída, em parte, à distribuição altamente desequilibrada entre os casos de doença hepática e os não casos no *dataset*.

3.4.2. Modelação com Dados *Binned*

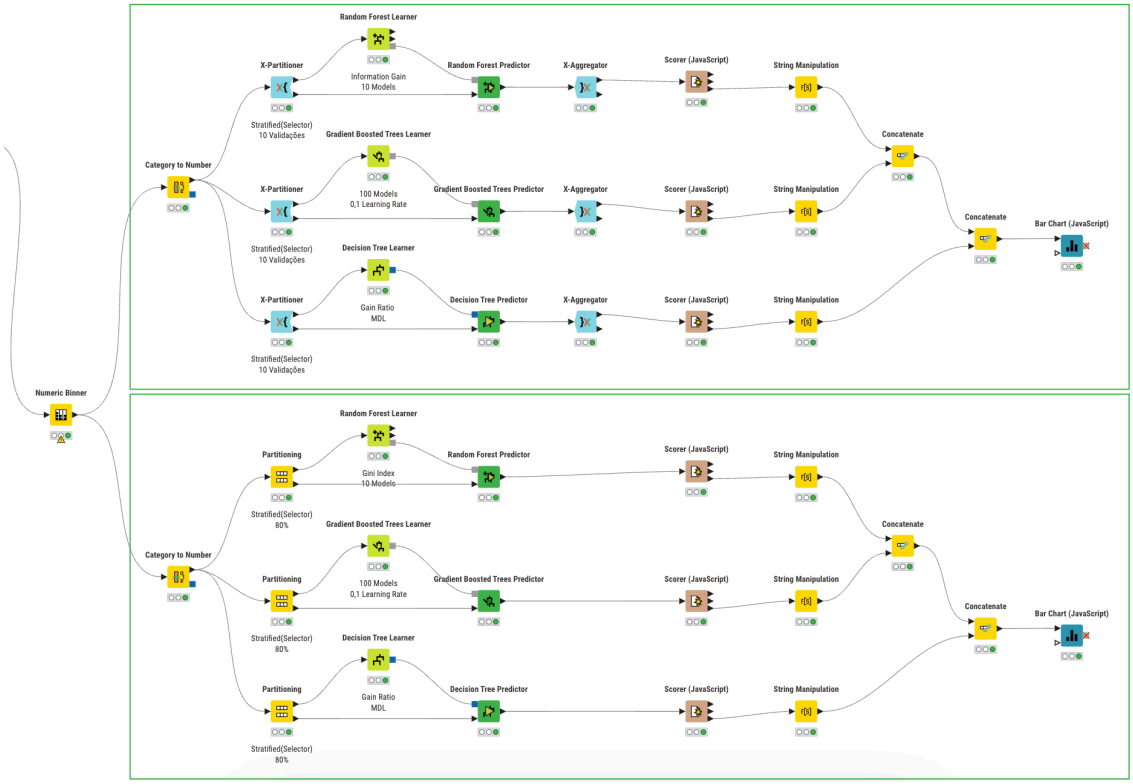


Figura 39: Modelação com Dados *Binned* com *Hold-out validation* e *Cross-Validation*

Tal como referido na fase de preparação de dados, optamos por agrupar em *bins*, dados relativos a idade e marcadores bioquímicos. Utilizando os mesmos algoritmos assim como *cross-validation* e *Hold-out validation* (separadamente), eis o melhor resultado obtido :

Scorer View

Confusion Matrix

	Doente (Predicted)	Não Doente (Predicted)	
Doente (Actual)	80	3	96.39%
Não Doente (Actual)	27	6	18.18%
	74.77%	66.67%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
74.14%	25.86%	0.187	86	30

Figura 40: *Scorer* Modelação com Dados *Binned* (*Random Forest*)

Como podemos ver pela Figura 40, esta abordagem para o nosso problema resultou numa *accuracy* pior do que a obtida anteriormente.

3.4.3. Modelação com *Feature Selection*

Para otimizar a seleção de atributos no nosso modelo de previsão, empregamos uma técnica conhecida como *Feature Selection*. Esta abordagem permite-nos identificar os atributos mais relevantes que contribuem significativamente para a precisão do modelo.

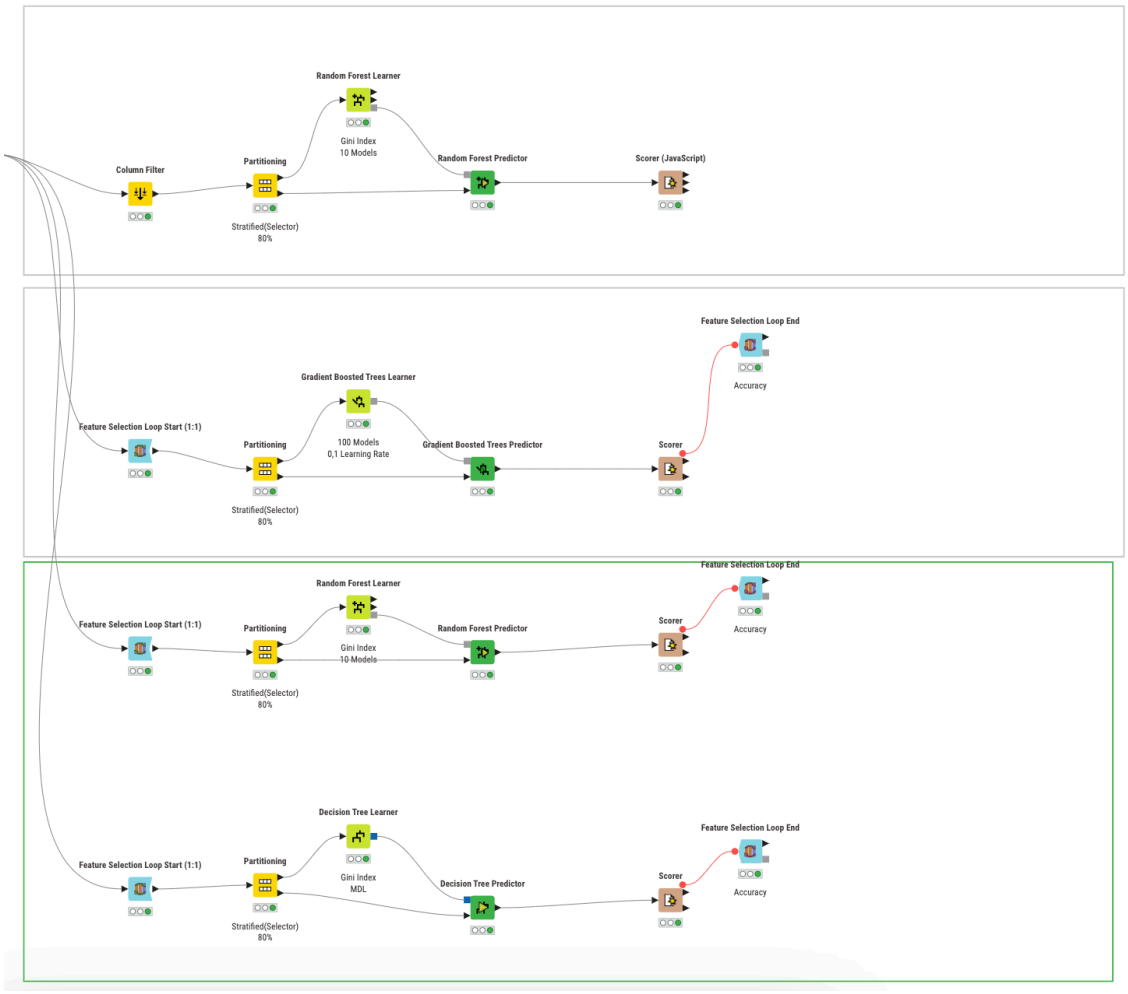


Figura 41: Modelação com *Feature Selection* com *Hold-out validation*

Durante o processo de *Feature Selection*, os atributos **TP** e **ALB** foram excluídos (pela melhor solução). De seguida, executamos o modelo mais promissor, *Random Forest*, utilizando apenas os atributos selecionados. O resultado obtido foi o mais alto registado até o momento em termos de precisão e coeficiente de *Cohen's kappa*, atingindo 80,17% e 0.436, respectivamente.

Scorer View

Confusion Matrix

	Doente (Predicted)	Não Doente (Predicted)	
Doente (Actual)	79	4	95.18%
Não Doente (Actual)	19	14	42.42%
	80.61%	77.78%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
80.17%	19.83%	0.436	93	23

Figura 42: *Scorer* Modelação com *Feature Selection* (*Random Forest*)

A análise sugere que a os atributos removidos não contribuem significativamente para a previsão de doenças hepáticas. Isto leva-nos a concluir que a probabilidade de um paciente desenvolver uma doença hepática não é substancialmente influenciada pelos marcadores bioquímicos desprezados.

3.4.4. Modelação com *Clustering*

Decidimos avançar com a análise deste problema investigando como a técnica de agrupamento (*clustering*) poderia ser aplicada ao *dataset*. Utilizamos o nodo *SOTA Learner* neste modelo. Mais uma vez, seguimos uma estratégia de validação com *Hold-Out* e *Cross-Validation*.

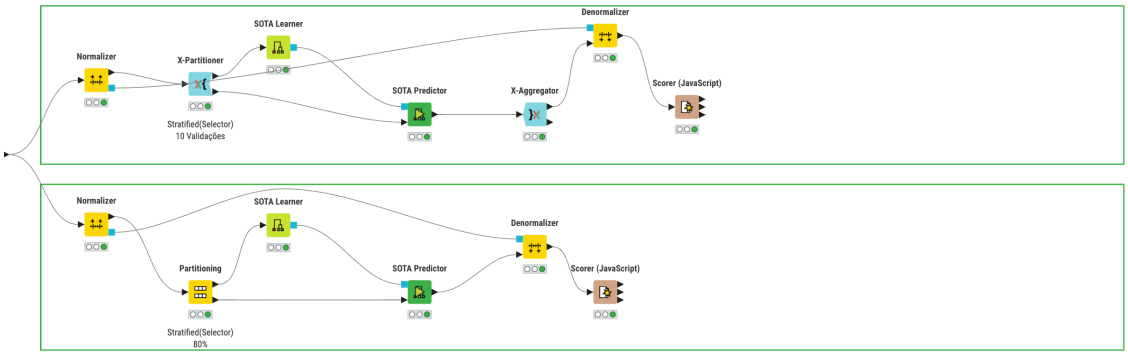


Figura 43: Modelação com *Clustering* com *Hold-out validation* e *Cross-Validation*

Os resultados mais promissores foram alcançados através da *Cross-Validation*. Para melhorar o desempenho, os dados foram normalizados, uma prática comum neste tipo de abordagem. Eis os resultados obtidos :

Scorer View

Confusion Matrix

	Doente (Predicted)	Não Doente (Predicted)	
Doente (Actual)	327	87	78.99%
Não Doente (Actual)	102	63	38.18%
	76.22%	42.00%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
67.36%	32.64%	0.176	390	189

Figura 44: *Scorer* Modelação com *Clustering* (*SOTA*)

Os resultados alcançados revelaram-se significativamente inferiores em comparação com os obtidos na modelação de referência (controlo). Esta disparidade sugere que a abordagem com *clustering* pode não ser a mais indicada para este contexto específico.

3.4.5. Modelação com *Downsampling*

Considerando o desequilíbrio extremo no *dataset* em relação à variável alvo (ou seja, há significativamente mais exemplos de não-doentes do que de doentes, com aproximadamente 29% dos exemplos sendo não-doentes em comparação com 71% de doentes), o grupo optou por aplicar o nodo *Equal Size Sampling*. Esta técnica visa equilibrar o número de observações para cada classe (doente e não-doente), proporcionando uma análise mais justa e precisa.

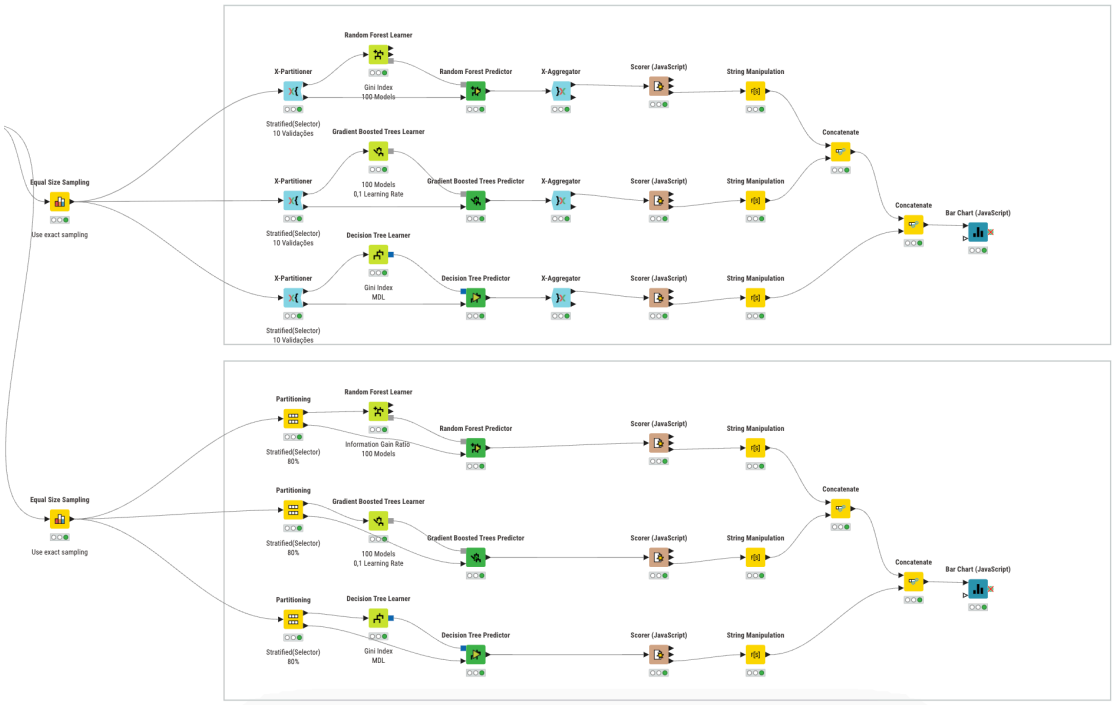


Figura 45: Modelação com *Downsampling* com *Hold-out validation* e *Cross-Validation*

No nodo em questão, foi utilizada a opção *use exact sampling*, para que o número de doentes e não doentes fosse exatamente o mesmo. Eis o melhor resultado obtido:

Scorer View

Confusion Matrix

	Doente (Predicted)	Não Doente (Predicted)	
Doente (Actual)	105	60	63.64%
Não Doente (Actual)	37	128	77.58%
	73.94%	68.09%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
70.61%	29.39%	0.412	233	97

Figura 46: *Scorer* Modelação com *Downsampling* (*Random Forest*)

Os resultados alcançados após a aplicação do *downsampling* foram superiores em comparação com os melhores resultados obtidos nas modelações anteriores utilizando *clustering*. No entanto, esta melhoria não necessariamente reflete uma maior precisão do modelo. Em vez disso, é provável que o aumento da *accuracy* seja atribuível principalmente à redução significativa do tamanho do *dataset*. Com menos exemplos para classificar, a probabilidade de acertar aleatoriamente aumenta, o que pode resultar numa pontuação aparentemente mais alta de *accuracy*.

3.4.6. Modelação com *Oversampling*

Para lidar com o desequilíbrio extremo no *dataset* em relação à variável alvo, é comum recorrer à técnica de *Oversampling*, que visa aumentar o número de exemplos da classe minoritária. Uma abordagem popular dentro do *Oversampling* é o **SMOTE** (*Synthetic Minority Over-sampling Technique*), que gera novos exemplos sintéticos da classe minoritária com base nos exemplos existentes.

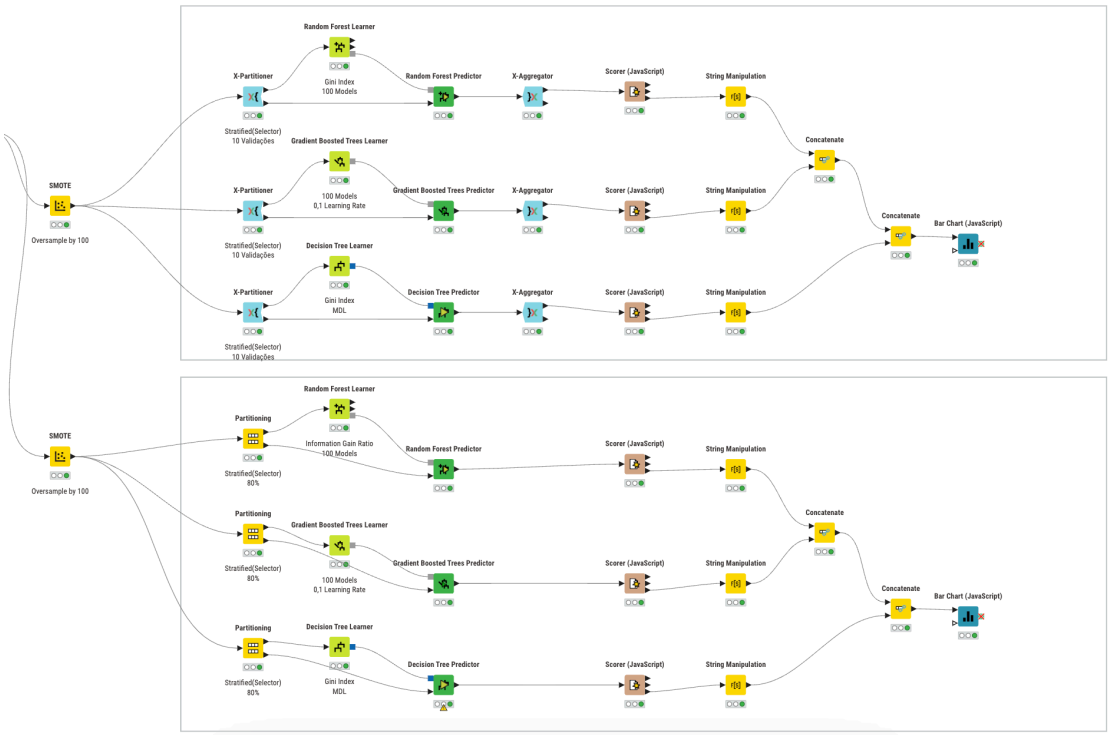


Figura 47: Modelação com *Oversampling* com *Hold-out validation* e *Cross-Validation*

Ao utilizar o algoritmo **SMOTE** com um fator de aumento (oversample by) igual a 100, obtivemos o seguinte resultado:

Scorer View

Confusion Matrix

	Doente (Predicted)	Não Doente (Predicted)	
Doente (Actual)	41799	15	99.96%
Não Doente (Actual)	65	16600	99.61%
	99.84%	99.91%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
99.86%	0.14%	0.997	58399	80

Figura 48: *Scorer* Modelação com *Oversampling* (*Random Forest*)

A modelação com *Oversampling* (utilizando *Random Forest* e *Cross-Validation*) resultou no valor mais alto e quase perfeito ($\simeq 100\%$) em termos de precisão (*accuracy*) alcançado até o momento, destacando-se significativamente em relação ao melhor resultado anterior. No entanto, é importante ressaltar que não podemos concluir que o modelo prevê melhor simplesmente com base nesta métrica. Isto porque **SMOTE** deve ser aplicado apenas aos dados de treino; ou seja, podemos criar *rows* artificiais apenas para treinar o modelo, sem influenciar os dados de teste.

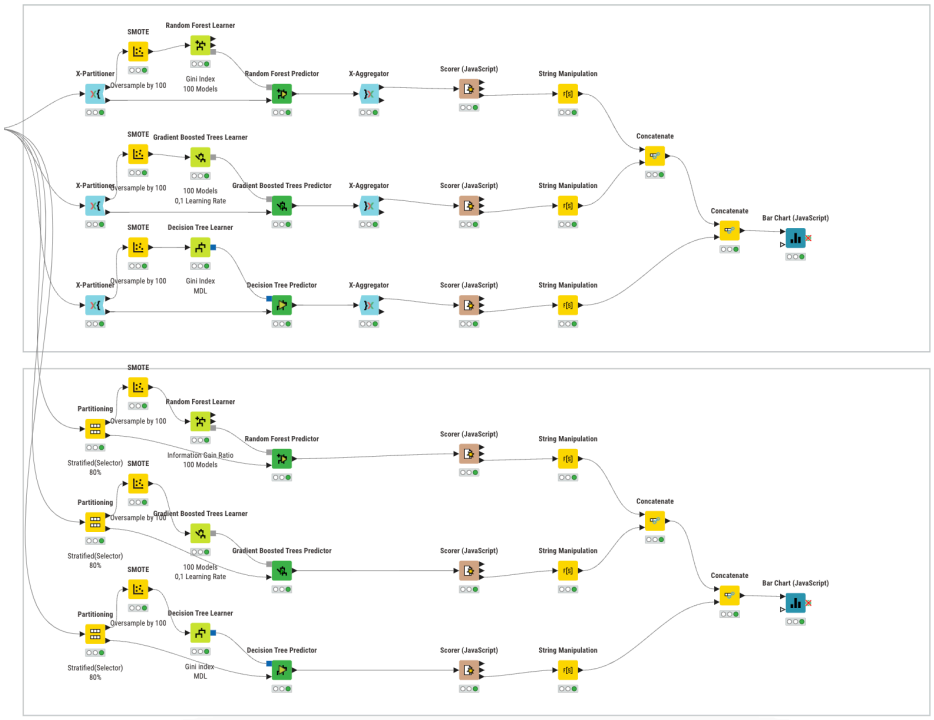


Figura 49: Modelação com *Oversampling* (apenas para treino) com *Hold-out validation* e *Cross-Validation*

Aplicando **SMOTE** apenas para criar observações artificiais para o treinamento (com um fator de aumento de 100), obtivemos o seguinte melhor resultado:

Scorer View

Confusion Matrix

	Doente (Predicted)	Não Doente (Predicted)	
Doente (Actual)	78	5	93.98%
Não Doente (Actual)	20	13	39.39%
	79.59%	72.22%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
78.45%	21.55%	0.387	91	25

Figura 50: *Scorer* Modelação com *Oversampling* apenas para treino (*Random Forest* e *Hold-out validation*)

Esta abordagem acaba por ser a que apresenta o segundo melhor valor de *accuracy* (perde apenas para *Feature Selection*) e garante que a avaliação do desempenho do modelo seja realizada de forma justa e realista, evitando qualquer influência artificial que possa distorcer os resultados. Uma razão possível para ter alcançado tal sucesso deve-se ao facto de poder ter tido muitas mais observações (artificiais) para treinar o modelo, aumentando a probabilidade de estar mais bem preparado para prever dados reais.

3.5. Avaliação

3.5.1. Modelação de Controlo

Na modelação de controlo, foram utilizados três algoritmos de aprendizagem para abordar o problema de classificação: *Random Forest*, *Gradient Boosting* e *Decision Tree*. Uma análise baseada em comparações revelou que o *Gradient Boosting* obteve a maior precisão entre os algoritmos testados, alcançando 75%. No entanto, o coeficiente de *Cohen's kappa* de 0.357 indica uma concordância apenas moderada entre as previsões do modelo e os valores reais. Isto pode ser atribuído à distribuição altamente desequilibrada entre os casos de doença hepática e não casos no *dataset*.

3.5.2. Modelação com Dados *Binned*

Ao agrupar dados de idade e marcadores bioquímicos em *bins*, a precisão do modelo diminuiu em comparação com a modelação de controlo, sugerindo que este tipo de processo possa ter causado perda de informação relevante para a previsão da doença hepática.

3.5.3. Modelação com *Feature Selection*

A aplicação da técnica de *Feature Selection* resultou numa melhoria significativa na precisão do modelo, alcançando uma precisão de 80,17% e um coeficiente de *Cohen's kappa* de 0.436 com o algoritmo *Random Forest*. A exclusão de certos atributos demonstrou que nem todos os marcadores bioquímicos têm influência significativa na previsão da doença hepática.

3.5.4. Modelação com *Clustering*

Os resultados da modelação com *clustering* foram inferiores aos da modelação de controlo e das abordagens anteriores. Isto indica que a técnica de *clustering* pode não ser adequada para este *dataset* ou que a configuração do modelo pode precisar de ajustes adicionais para melhorar o desempenho.

3.5.5. Modelação com *Downsampling*

A aplicação da técnica de *downsampling* resultou numa melhoria na precisão do modelo (em comparaão com a técnica de *clustering*), mas é importante notar que esta melhoria pode ser atribuída principalmente à redução do tamanho do *dataset*.

3.5.6. Modelação com *Oversampling*

A modelação com *oversampling* usando o algoritmo **SMOTE** resultou numa precisão quase perfeita. No entanto, ao aplicar **SMOTE** apenas aos dados de treino, a precisão permaneceu alta, mas mais realista, evitando potenciais distorções devido ao aumento artificial da classe minoritária.

3.5.7. Conclusão

A aplicação de diferentes técnicas de modelação revelou *insights* importantes sobre a previsão de doenças hepáticas. A seleção de atributos mostrou-se crucial para melhorar a precisão do modelo, enquanto a o uso de *bins*, clustering e *downsampling* não produziram resultados tão favoráveis. O *oversampling* com **SMOTE** resultou em altas precisões, mas é importante aplicá-lo de forma realista, apenas aos dados de treino, para evitar distorções nos resultados. Em geral, a escolha da técnica de modelação depende da natureza dos dados e dos objetivos específicos do projeto.

4. Conclusão

Com a conclusão deste trabalho prático, evidenciamos a nossa proficiência na aplicação de uma gama diversificada de conceitos inerentes ao desenvolvimento de modelos de aprendizagem, tanto aos amplamente discutidos durante o período académico quanto os conceitos mais especializados que emergiram ao longo da investigação.

Ao longo do desenvolvimento deste projeto, não apenas nos dedicamos à construção de modelos de aprendizagem, como também realizamos uma análise aprofundada dos *datasets*, seguida de um processo meticuloso de pré-processamento. Esta abordagem proporcionou-nos uma compreensão aprofundada do problema em questão, permitindo-nos selecionar as ferramentas mais adequadas para abordar os desafios identificados.

Apesar dos desafios iniciais enfrentados na seleção do *dataset* para a Tarefa A, consideramos que alcançamos resultados notáveis nas fases subsequentes. Os modelos de aprendizagem que desenvolvemos demonstraram um desempenho satisfatório para os conjuntos de dados investigados, ao passo que mantivemos um registo detalhado e sistemático de todo o processo de pesquisa e desenvolvimento do projeto.

Bibliografia

Alp - Blood Test. Mount Sinai Health System. (n.d.-a). https://www.mountsinai.org/health-library/tests/alp-blood-test

James Myhre & Dennis Sifris, M. (2023, November 29). What does a high or low A/G ratio mean?. Verywell Health. https://www.verywellhealth.com/a-g-ratio-8405376

Mohammad Wehbi, M. (2020, December 5). Bilirubin. Reference Range, Interpretation, Collection and Panels. https://emedicine.medscape.com/article/2074068-overview?form=fpf

Moriles, K. E. (2022, December 10). Alanine amino transferase. StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK559278/

Total protein. Mount Sinai Health System. (n.d.-b). https://www.mountsinai.org/health-library/tests/total-protein

GregorySmith. (2016a, October 26). Video game sales. Kaggle. https://www.kaggle.com/datasets/gregorut/videogamesales

ILPD (Indian liver patient dataset). UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset