Universidade do Minho
Licenciatura em Engenharia Informática

Unidade Curricular de Aprendizagem e Decisões Inteligentes

19 de maio de 2025

Trabalho Prático de Aprendizagem e Decisões Inteligentes

Dataset Atribuído

Grupo 31:

- **João Ferreira, nº89497;**
- **Bruno Campos, nº98639;**
- **Tiago Alves, nº80872;**
- **Gustavo Castro, nº100482.**

ADI

Índice

Índice de Figuras

1. Introdução

Este relatório foi elaborado no âmbito da unidade curricular de Aprendizagem e Decisões Inteligentes cujo objetivo é aplicar diferentes paradigmas de aprendizagem e recorrer a técnicas de *machine learning*.

Este projeto é dividido em 2 tarefas, sendo uma delas aplicar as técnicas a um dataset atribuído pelos professores e outra onde o dataset a ser usado é escolhido pelo grupo.

Para melhor dividir o trabalho, foram criados 2 relatórios, sendo que neste estamos a tratar do dataset atribuído pelos professores.

2. Metodologia

Para a metodologia entendemos usar o **CRISP-DM** (Cross Industry Standard Process for Data Mining) por oferecer uma estrutura clara para o desenvolvimento de projetos. Sendo este processo foi dividido em seis fases:

- **Estudo do Negócio:** O principal objetivo do projeto é explorar um conjunto de dados da área da saúde, com o intuito de desenvolver modelos que possam permitam estimar resultados clínicos (tal como o resultado de testes médicos) ou valores de faturação, a partir das variáveis disponíveis.

- **Estudo dos Dados:** Através da análise do dataset disponibilizado, foi possível identificar as principais características dos pacientes, condições médicas, dados administrativos e resultados de exames. Este passo permitiu compreender a estrutura e a qualidade dos dados.

- **Preparação dos Dados:** Foram aplicadas técnicas de limpeza, transformação e codificação das variáveis, de modo a preparar os dados para a fase de modelação. Para este tratamento de dados, utilizamos a plataforma KNIME.

- **Modelação:** Diversos algoritmos de aprendizagem automática foram testados com o objetivo de encontrar o modelo mais adequado ao problema em análise.

- **Avaliação:** Os modelos foram avaliados, sendo realizada uma comparação entre os diferentes modelos desenvolvidos a partir da sua eficácia.

- **Desenvolvimento:** Visto que se trata de um projeto a nível académico, esta fase não está presente, mas poderia, eventualmente, vir a ser implementada em sistemas clínicos para que estes tenham um apoio à decisão.

3. Análise do Dataset

O dataset atribuído pelos professores foi escolhido através do número do grupo, ficando o nosso grupo (31) com um dataset de dados clínicos.

O dataset tem a seguinte estrutura:

- **Name**: Representa o nome do paciente;
- **Age**: Idade do doente à data de admissão, expressa em anos;
- **Gender**: Indica o género do doente, "Masculine" ou "Feminine";
- **Blood Type**: O tipo sanguíneo do paciente ("A+", "O-", etc.);
- **Medical Condition**: Especifica a condição médica / diagnóstico associado ao paciente;
- **Date of Admission**: Data em que o doente deu entrada no estabelecimento de saúde;
- **Doctor**: Nome do médico responsável pelo atendimento do doente;
- **Hospital**: Estabelecimento de saúde ou hospital onde o doente foi internado;
- **Insurance Provider**: Indica o fornecedor de seguros do paciente;
- **Billing Amount**: Valor cobrado pelos serviços de saúde do doente;
- **Room Number**: Número do quarto onde o doente foi acomodado;
- **Admission Type**: Especifica o tipo de admissão (“Emergency", "Urgent", etc.);
- **Discharge Date**: Data em que o doente teve alta do estabelecimento de saúde;
- **Medication**: Identifica um medicamento prescrito ou administrado ao doente durante a sua admissão;
- **Test Results**: Descreve os resultados de um teste médico realizado.

3.1. Estatísticas Iniciais e Target Value

Após uma análise inicial com os dados, verificamos que o dataset contém 50.000 registos, onde 50.09% das entradas são de pacientes masculinos, a média de idades é 51.50 anos, o tipo de sangue menos presente é o “O-” e o atributo “Test Results” é a nossa variável alvo (target), visto que se trata de um problema de classificação.

3.2. Outliers e Dados em Falta

Apesar de numa primeira análise os dados parecerem muito equilibrados, através do nodo “Data Explorer” conseguimos ver várias incongruências, tal como o “Billing Amount” ter valores negativos precisando de usar o módulo para poderem estar positivos outra vez, ou outros atributos tal como a idade que através de um “scatter plot” descobrimos uma afluência atípica de pessoas com 37 anos, sendo que estes deviam ter outra idade segundo a sua data de nascimento.

Notamos também vários atributos que continham valores nulos, e para não estar a retirar registos à toa, decidimos usar o valor mais frequente para o substituir.

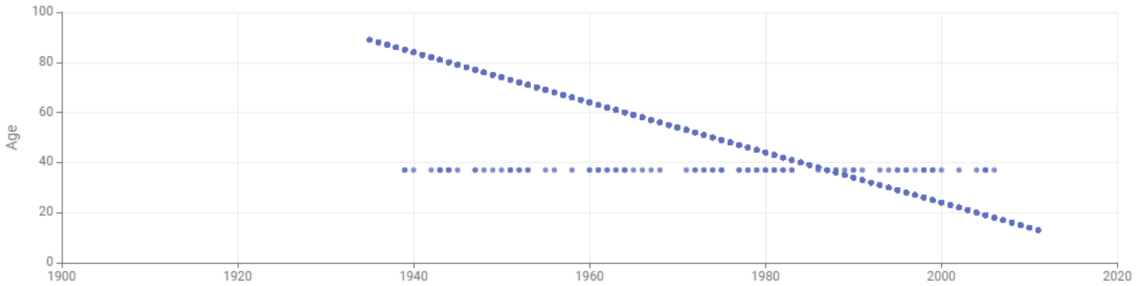

Figura 1: Scatter Plot com as idades dos pacientes por ano.

3.3. Visualizações

Foram utilizadas ferramentas de visualização no KNIME, como histogramas, boxplots e gráficos de barras, para explorar a distribuição dos dados e relações entre variáveis. Por norma todos os atributos estavam distribuídos equivalentemente, com exceção dos atributos que tinham valores “repetidos” como o género que tinha os valores “Female”, “Girl” e “Feminine” que se referem ao mesmo, e do “Admission Type” que era irregular.

Figura 2: Data Explorer com valores e gráfico de frequência do género.

4. Processamento de Dados

Todo o pré-processamento foi realizado na plataforma KNIME, utilizando os seguintes procedimentos:

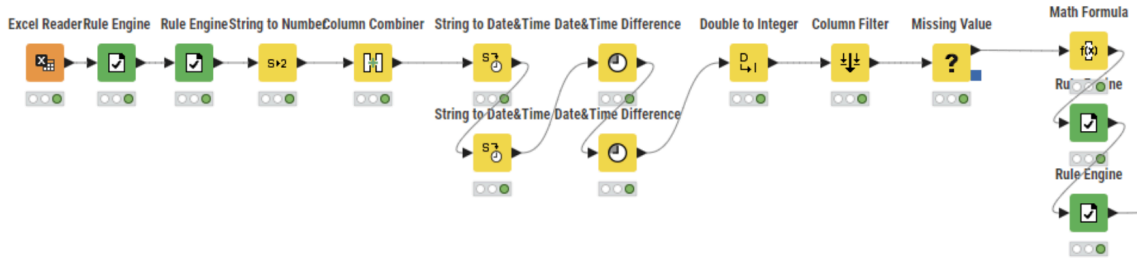

Figura 3: Tratamento dos dados na plataforma KNIME.

- **Excel Reader:** Serve para ler e extrair o dataset;
- **1º Rule Engine:** Serve para converter os meses que têm dias que não deviam. Cada mês que tem um registo com dias a mais que não devia, sendo convertidos para o verdadeiro fim do mês;
- **2º Rule Engine:** Converte as “repetições” do atributo “Gender” para os inteiros 1 e 0, sendo 1 referente ao género masculino e 0 para o género feminino;
- **String to Number:** Muda o “Billing Amount” de string para double;
- **Column Combiner:** Combina os atributos “Year”, “Month” e “Day” de modo a tornar numa data denominada “Date of Birth”;
- **1º String to Date&Time:** Utilizado para converter a “Discharge Date” numa data;
- **2º String to Date&Time:** Converte a “Date of Birth” criada numa data;
- **1º Date&Time Difference:** Utiliza a “Discharge Date” e a “Admission Date” para criar um novo atributo “Dias no Hospital” que mostra o número de dias que o paciente passou no hospital;
- **2º Date&Time Difference:** Devido a haver erros na idade dada, utilizamos a “Date of Birth” para calcular a idade dos pacientes, tendo como data final para o cálculo, o dia 19-05-2025;
- **Double to Integer:** Converte “Billing Amount”, “Dias no Hospital” e “Idade” para inteiros, sendo os valores arredondados normalmente;
- **Column Filter:** Devido à pouca correlação, achamos melhor retirar da equação os seguintes atributos “Age”, “Year”, “Month”, “Day”, “Name”, “Floor” e “Date of Birth”;
- **Missing Value:** Decidimos que não devíamos remover nenhuma entrada, tendo assim usado a média para os números inteiros e a moda para as strings;
- **Math Formula:** Serve para fazer o módulo dos valores negativos do “Billing Amount”;
- **3º Rule Engine:** Converte as “repetições” do atributo “Medication” de modo a eliminar os medicamentos mal escritos;
- **4º Rule Engine:** Converte as “repetições” do atributo “Medical Condition” de modo a eliminar as condições mal escritas.

5. Avaliação dos Modelos

A validação dos modelos foi realizada através de *Hold-Out cross-validation* já que neste dataset havia, já separados, ficheiros distintos, sendo um para treino e outro para teste, onde cada modelo é treinado no primeiro ficheiro sendo posteriormente posto à prova usando o ficheiro de teste.

Para este dataset decidimos usar *Decision Trees*, *Random Forest*, *Gradient Boost* e *k-Means* como modelos para testar o nosso tratamento de dados, usando sempre uma *seed* fixa (99099).

Para o *Random Forest* usamos como critério o *Information Gain Ratio* visto que foi o que nos deu melhores resultados.

Já para a *Decision Tree* usamos o "Gini index" para medir a qualidade de como o *split* vai ser calculado, não usamos pruning nenhum, e usamos a opção de "average split point" já que são as que nos deram melhores resultados.

Nas *Gradient Boosted Trees* apenas mexemos no número de modelos e na curva de aprendizagem, vendo pouca variação.

Tentamos usar o k-Means e k-Medoids como método de clustering onde mudamos o número de iterações, porém não muito efetivo tal como dá para ver pela tabela. Para isto normalizamos os atributos e calculamos as distâncias numéricas destes, tendo no fim de utilizar um rule engine para adaptar o resultado de cada um.

Tipo de Modelo	Percentagem de *Accuracy*	Tempo Registado
Random Forest com 100 árvores	45,62%	8,43s
Random Forest com 300 árvores	45,36%	25,64s
Gradient Boost com 2 árvores 0.1 rate	42,88%	1,91s
Gradient Boost com 1 árvore 0.1 rate	42,80%	1,32s
Gradient Boost com 1 árvore 0.2 rate	42,48%	1,33s
Decision Tree com 1 nodo mínimo	40,40%	0,68s
Decision Tree com 2 nodo mínimo	39,72%	0,99s
Random Forest com 1 árvore	38,72%	1,18s
k-Medoids com 1 iteração	38,46%	1,04s
k-Medoids sem restrições	36,04%	2,19s
k-Means com 100 iterações	34,42%	0,47s
k-Means com 1 iteração	33,94%	0,18s

6. Conclusões

A análise revelou que o modelo de *Random Forest* destacou-se como o mais eficaz para este problema de classificação. Apesar de menos preciso, o modelo *Gradient Boost* destaca-se pela próxima precisão ao *Random Forest* mas em muito menos tempo que este último.

Os resultados não são promissores em nenhum dos modelos, visto que nenhum deles consegue sequer passar dos 50%, ficando entendido que o será preciso uma reforma no tratamento de dados, potencialmente acrescentando novos atributos, bem como retirar outros que não são tão relevantes.

Este projeto deixou-nos aprofundar conhecimentos relativos à resolução de problemas com base na classificação e permitiu que explorássemos um conjunto de dados clínicos com diferentes tipos de atributos, compreendendo melhor o impacto que a inteligência artificial pode ter no mundo real, como o resultado de testes médicos com base em dados anteriores.