

Relatório de Projeto

Modelo Preditivo: Dataset de Ortopedia sobre artroplastia total do
joelho

Unidade Curricular: Machine Learning

Docente: Professora Andreia Nunes

Realizado por:

Ana Ripado, n.º	2023136
Raquel Quintão n.º	2023134
Rita Varandas n.º	2023112
Paulo Nunes n.º	2023129

Lisboa, Maio de 2024

Índice

1. Introdução	4
2. Análise Exploratória de Dados	5
2.1. Importação e Verificação Inicial dos Dados	5
2.2 Criação da Coluna Target e Validação da sua Distribuição	5
2.3 Tratamento de Valores Ausentes	6
2.4 Análise da Variância e Valores Duplicados	7
2.5 Outliers.....	7
2.6 Análise Estatística	7
2.8 Análise Bivariada.....	9
2.9 Análise Multivariada	9
2.10 Conclusão da Análise Exploratória	11
3. Modelação.....	11
3.1 Separação de target e features (df_orto_y, df_orto_X)	11
3.2 Técnicas de Balanceamento de Dados	12
3.3. Aplicação dos modelos e geração de dados de classificação.....	13
4. Avaliação	14
4.1 Interpretação das Métricas de Avaliação e Seleção do Melhor Modelo	14
4.1.1 Modelo Global (Variáveis Pré e Pós-Cirúrgicas)	14
4.1.2 Modelo “Momento 0” (Variáveis pré-cirúrgicas).....	16
4.2. Análise das Variáveis Mais Importantes	18
4.2.1. Modelo Global	18
4.2.2. Modelo Momento 0.....	19
5. Conclusão	21

1. Introdução

No âmbito da Pós-Graduação em *Health Data Science* da Escola Superior de Tecnologia da Saúde de Lisboa, fomos desafiados a desenvolver um modelo preditivo para avaliar as mudanças na classificação CPAK, decorrente da cirurgia de artroplastia do joelho, que reflete o alinhamento do joelho antes e após a cirurgia.

O projeto foi estruturado em três fases principais:

1. *Exploratory Data Analysis (EDA)*: esta fase inicial envolveu uma análise detalhada dos dados disponíveis, identificando padrões, anomalias e correlações fundamentais. O objetivo foi compreender as variáveis envolvidas e preparar o dataset para a modelação.

2. *Modelling*: Com base nos *insights* obtidos na fase de EDA, dois modelos preditivos foram testados e ajustados para melhor se adequarem às especificidades dos dados. Esta fase focou-se na seleção e aplicação de técnicas de *machine learning* apropriadas para prever a mudança de CPAK.

3. *Evaluation*: A última fase do projeto consistiu na avaliação rigorosa dos modelos através de métricas estatísticas mais sofisticadas, como a sob a curva ROC (AUC-ROC), precisão, *recall* (*sensibilidade*) e *F1-score*. O propósito era validar a eficácia do modelo preditivo e assegurar sua confiabilidade e aplicabilidade.

Este projeto considera o desenvolvimento de 2 modelos preditivos de classificação. O primeiro modelo, Modelo Global, engloba todas as variáveis da base de dados, onde se incluem as variáveis pré e pós cirurgia. O segundo modelo, denominado de Momento 0, inclui somente as variáveis pré-cirúrgicas

Consideramos que a aplicação destes 2 modelos terá objetivos e aplicações distintas na prática clínica e poderão forneceram *insights* relevantes.

Este trabalho foi desenvolvido com base no Modelo Global e, posteriormente, replicado e ajustado para o Momento 0. A análise e raciocínio serão abordados ao longo do desenvolvimento do projeto.

2. Análise Exploratória de Dados

A análise Exploratória de Dados é uma etapa fundamental para entender os dados, captar insights iniciais que podem influenciar a construção dos modelos e preparar o *dataset* para a fase de modelação. Faz parte de qualquer projeto de *Machine Learning*.

2.1. Importação e Verificação Inicial dos Dados

Iniciámos o projeto importando o *dataset* (*df_orto*) e verificando as suas dimensões para confirmar a importação bem-sucedida.

O *dataset* inclui 265 registos de utentes submetidos a uma cirurgia de substituição do joelho, com 22 variáveis que incluem métricas pré e pós-cirúrgicas. A variável CPAK varia entre os níveis de classificação de 1 a 9, sabendo que o nível 5 traduz o alinhamento “perfeito” do joelho.

2.2 Criação da Coluna Target e Validação da sua Distribuição

Definimos a variável *target* baseada na mudança ou não dos grupos CPAK entre o período pré e pós-cirúrgico:

- *Target 0* – Grupo pré = Grupo Pós
 - CPAK mantém-se: 236 registos (89%)
- *Target 1* – Grupo Pré ≠ Grupo Pós
 - CPAK altera-se: 29 registos (11%)

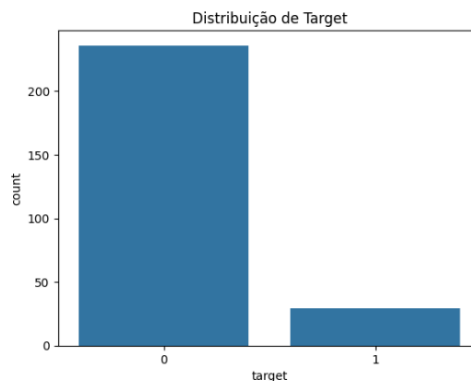


Gráfico 1 – Distribuição variável *target* (*df_orto*)

Observámos a distribuição inicial da variável alvo que criámos, o que nos ajudou a entender o desequilíbrio entre as classes.

Dado que temos um desbalanceamento significativo, sabemos de antemão que, mais à frente no processo, teremos de implementar técnicas de preparação dos dados (*undersampling* e *oversampling*).

As colunas usadas para criar a *target* foram removidas para evitar viés no modelo preditivo.

O passo seguinte envolveu a pesquisa de registos duplicados no *dataset*, o que não se verificou. Posto isto, procedemos à validação da existência de *missing values* e subsequente tratamento.

Todas as variáveis foram verificadas para confirmar que estavam no formato numérico, eliminando a necessidade de transformações adicionais para este tipo de dados.

2.3 Tratamento de Valores Ausentes

Embora não sejam reportados *missing values* pela biblioteca *pandas*, sabemos de antemão que estão a ser representados por 999. Como 999 não tem significado para nenhuma das colunas, foi aplicado o *replace* por NaN em todo o *dataframe* e validado novamente.

Os *missing values* identificados nas seguintes variáveis, por ordem foram os seguintes:

- Satisfação: 10
- EVA_0: 6
- EVA_90: 4
- Fle_0: 3
- Peso: 2
- Altura_cm: 2
- IMC: 2
- PM6_0: 2
- Fle_90: 2
- PM6_90: 2
- WATotal_0: 1

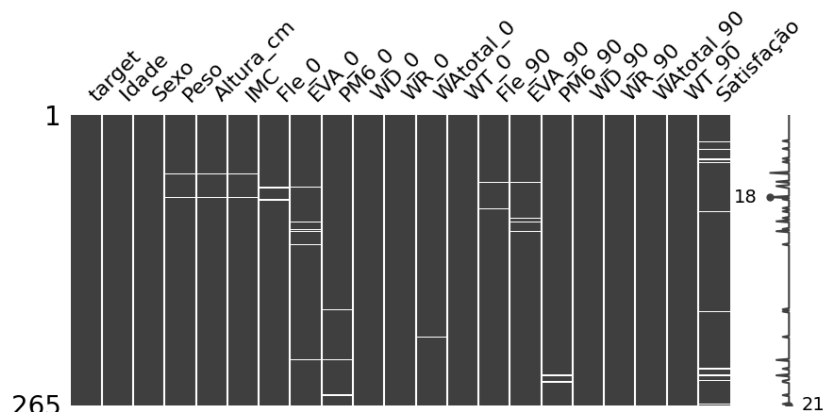


Gráfico 2 – Identificação de missing values nas variáveis

Tendo em conta a baixa prevalência de *missing values* em cada coluna/variável, não houve necessidade de remoção de nenhuma coluna. Neste sentido e dada a dimensão do *dataframe*, optamos pelo tratamento em vez da eliminação desses registos.

Em relação ao tratamento destes valores, começámos por classificar os diferentes tipos de variáveis (variáveis contínuas, discretas, ordinais, etc.) e criámos um ciclo *'for'* para fazer a substituição dos mesmos.

As variáveis com valores ausentes foram tratadas conforme o tipo:

- Variáveis contínuas e discretas: Substituição pela média
- Variáveis ordinais e categóricas: Substituição pela moda

Para validar o procedimento, voltámos a procurar e confirmámos a inexistência de valores ausentes.

2.4 Análise da Variância e Valores Duplicados

Na análise exploratória, avaliámos também as variáveis do *dataset* para identificar aquelas com variância praticamente nula, ou seja, com 99% dos valores idênticos.

Variáveis com alta homogeneidade são geralmente consideradas para remoção porque oferecem pouca informação que possa contribuir para o poder preditivo dos modelos.

A nossa revisão indicou que todas as variáveis do *dataset* apresentam variabilidade adequada, garantindo assim que continuamos a análise com um conjunto de dados que possui diversidade suficiente para suportar análises preditivas eficazes.

Verificou-se ainda a eventual presença de valores duplicados, mas estão ausentes do nosso *dataset*. O foco centrou-se então na identificação e tratamento dos *outliers*.

2.5 Outliers

Identificámos *outliers* nas variáveis numéricas e optámos por substituir os valores extremos (superiores e inferiores) pelos limites calculados através do método IQR, preferindo esta abordagem à eliminação completa para preservar o tamanho da amostra.

Para aferir a melhor opção face ao nosso *dataset*, fizemos uma comparação entre os dois métodos para perceber o impacto de cada um deles. Para isso, começámos por criar 2 novos *datasets* (*df_orto2*, *df_orto3*).

Analisámos o número e a percentagem de valores perdidos em cada método.

Concluímos que ao utilizar o método de eliminação, iríamos perder 20,38% dos dados, o que era significativo, sobretudo dado que o *dataset* já é limitado em termos do número de registos.

Escolhemos, assim, a segunda opção, ou seja, a substituição dos *outliers* pelos máximos e mínimos. A partir deste momento, passámos então a trabalhar com o *dataset* *df_orto3*.

2.6 Análise Estatística

Antes de proceder à análise estatística, importa clarificar algumas variáveis para efeitos de uma melhor compreensão do *dataset* e respetiva análise.

WOMAC:

1. WD_0 e WD_90 | Dor: avalia a dor durante atividades como caminhar, subir escadas, etc.
 2. WR_0 e WR_90 | Rigidez: avalia a rigidez após acordar e mais tarde durante o dia.
 3. WATotal_0 e WATotal_90 | Função Física: avalia a capacidade de realizar diversas atividades físicas.
- Pontuações mais altas no WOMAC indicam pior dor, rigidez e limitações funcionais.

PM6 (prova de marcha): medida da distância percorrida em seis minutos, em metros, refletindo a capacidade física do utente.

Escala Visual Analógica (EVA) da dor: utilizada para medir a perceção da intensidade da dor numa escala de 0 (sem dor) a 10 (dor máxima), sendo uma ferramenta simples, mas eficaz para avaliar o bem-estar do utente.

CPAK

A classificação CPAK é uma maneira de categorizar o alinhamento do joelho. Compreende nove fenótipos diferentes, que vão de I a IX.

Feita uma breve descrição das variáveis que podem suscitar mais questões em termos de interpretação, seguimos para a análise dos resultados do estudo.

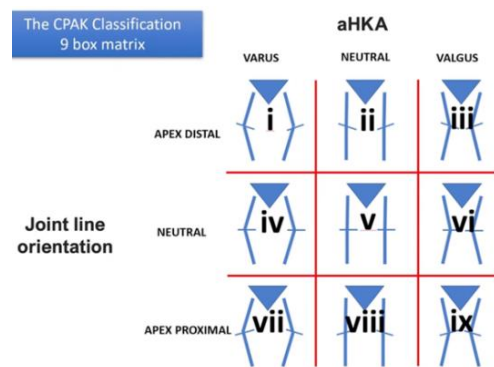


Fig. 1 – Identificação do grupo CPAK

2.7 Análise Univariada

Qualitativa | var_int = variáveis ordinais + categóricas

- **Satisfação Pós-Cirúrgica:** Notavelmente alta após a cirurgia, com a maioria dos utentes a reportar níveis de satisfação de 10/10, indicando sucesso cirúrgico.
- **Dor (EVA):** Observamos uma diminuição significativa nos valores de EVA de pré para pós-cirurgia, refletindo melhoria na condição de dor dos utentes.

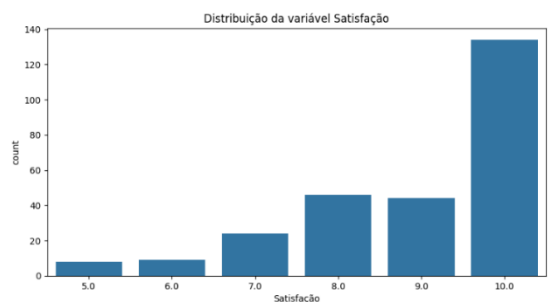


Gráfico 3 – Tabela de frequências nível satisfação

- **Distribuição de Sexo:** Predominância feminina no *dataset*, o que é relevante para considerar possíveis diferenças de resposta ao tratamento entre géneros.

Quantitativa | $var_num = \text{variáveis contínuas} + \text{discretas}$

- **Distribuições de Peso e Altura:** Normais, sem evidências de *outliers* significativos, fornecendo uma base sólida para a análise demográfica.
- **Prova de Marcha (PM6):** A melhoria da capacidade de marcha após a cirurgia é evidente, com um aumento na distância percorrida, o que sugere uma recuperação funcional efetiva.
- **Flexibilidade do Joelho (Fle_0 e Fle_90):** Melhoria notável na flexibilidade do joelho pós-cirurgia, com menos utentes apresentando restrições severas de movimento.

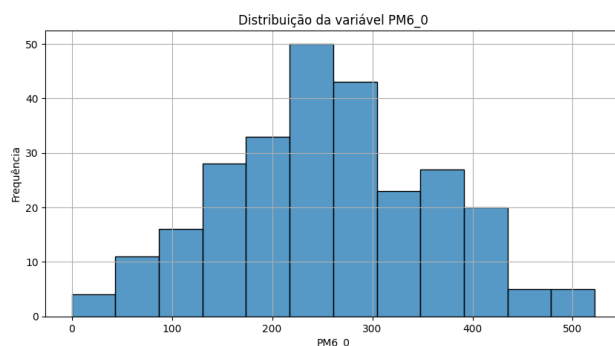


Gráfico 4 – Histograma PM6_0

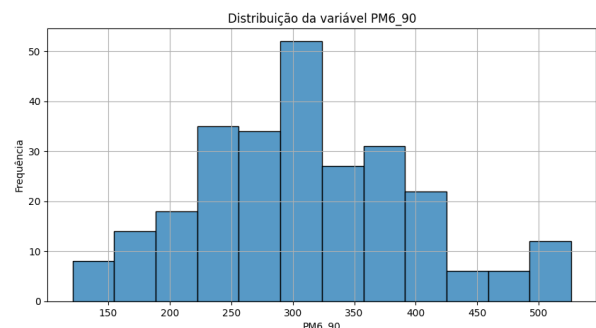


Gráfico 5 – Histograma PM6_90

2.8 Análise Bivariada

Analisámos de forma bivariada todas as variáveis. Explorámos como cada variável se comporta em relação à mudança nos grupos CPAK, comparando os utentes que mudaram de grupo com aqueles que permaneceram no mesmo grupo após a cirurgia(target).

Observámos que certas variáveis exibem padrões distintos, o que sugere a sua relevância para a predição de mudanças no grupo CPAK.

2.9 Análise Multivariada

Procedemos à análise multivariada para entender a correlação entre as variáveis e remover as que são fortemente correlacionadas (próximas de -1 ou 1), uma vez que não trazem valor, sendo assim consideradas redundantes para uma classificação preditiva.

Correlação de Spearman e Pearson

Inicialmente, aplicamos a correlação de *Spearman* para captar as relações que se evidenciam, entre as variáveis. Esta análise reflete-se através de tabelas e um heatmap que facilitaram a visualização das correlações mais intensas, incluindo relações monotônicas entre duas variáveis.

Identificámos correlações especialmente fortes entre:

- **WAtotal_0 e WT_0:** Esta forte correlação sugere uma redundância significativa entre estas variáveis, indicando que ambas partilham informações similares sobre a condição física dos utentes no momento pré-cirurgia.
- **WAtotal_90 e WT_90:** Da mesma forma, esta correlação destaca uma sobreposição nas informações capturadas por essas medidas no momento pós-cirurgia.

Complementando a análise com a correlação de *Pearson*, procurámos confirmar as correlações identificadas e entender o seu impacto direto na variável alvo (target).

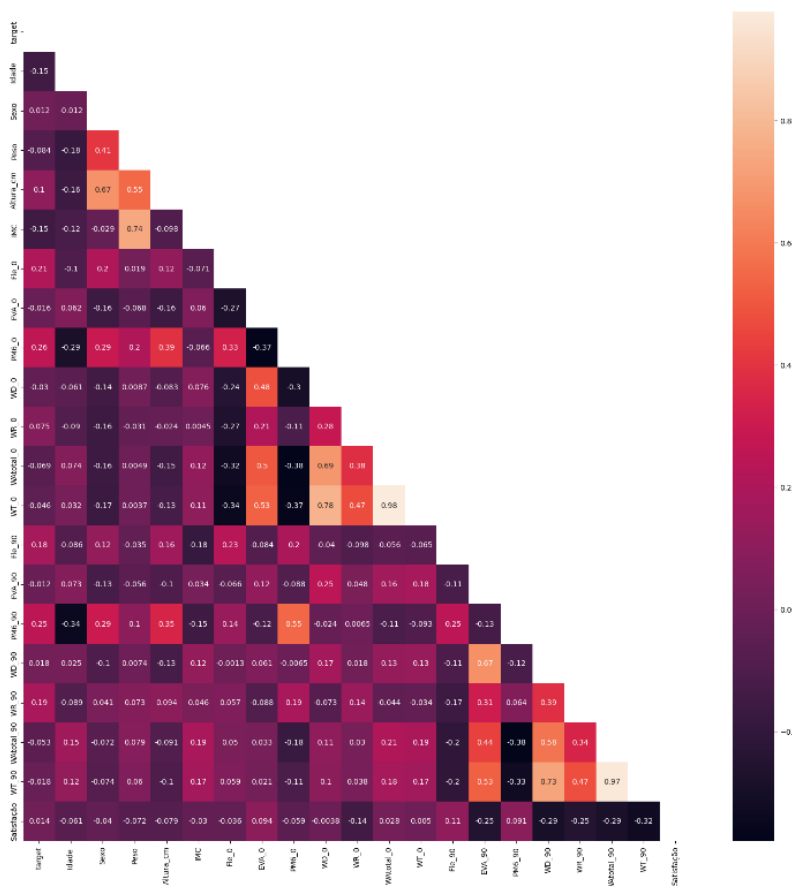


Gráfico 6 – Heatmap Correlação de Spearman

	level_0	level_1	Corr_Spearman
0	WAtotal_0	WT_0	0.977849
1	WT_0	WAtotal_0	0.977849
2	WT_90	WAtotal_90	0.965177
3	WAtotal_90	WT_90	0.965177
4	WD_0	WT_0	0.777303
5	WT_0	WD_0	0.777303
6	Peso	IMC	0.744383
7	IMC	Peso	0.744383
8	WT_90	WD_90	0.730253
9	WD_90	WT_90	0.730253

Tabela 1 – Correlação de Spearman

	level_0	level_1	Corr_Pearson
0	WAtotal_0	WT_0	0.978113
1	WT_0	WAtotal_0	0.978113
2	WT_90	WAtotal_90	0.964167
3	WAtotal_90	WT_90	0.964167
4	WT_0	WD_0	0.807959
5	WD_0	WT_0	0.807959
6	WT_90	WD_90	0.767937
7	WD_90	WT_90	0.767937
8	IMC	Peso	0.736511
9	Peso	IMC	0.736511
10	WAtotal_0	WD_0	0.706992
11	WD_0	WAtotal_0	0.706992

Tabela 2 – Correlação de Pearson

Decisão sobre a Remoção de Variáveis

Ao avaliar qual variável de cada par correlacionado deveria ser removida, consideramos a força e a relevância de cada uma para a previsão da variável alvo. A partir da análise dos *heatmaps*, decidimos eliminar:

- **WT_90**: Mostrou-se menos correlacionada com a target comparativamente a WAtotal_90.
- **WT_0**: Apresentou menor correlação com a target em relação a WAtotal_0.

Este processo de eliminação foi fundamentado pela necessidade de reduzir a complexidade do modelo sem comprometer a integridade das informações essenciais.

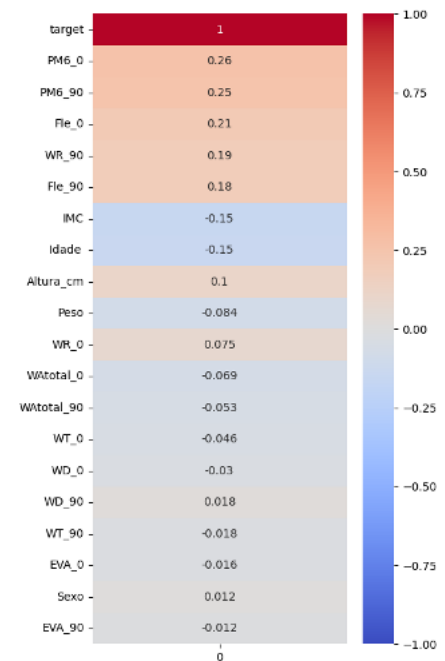


Gráfico 7 – Correlações das variáveis com a Target

Validação Final das Correlações

Realizámos uma última verificação através de um novo *heatmap* (anexo) para garantir que não restassem correlações excessivamente fortes entre as variáveis remanescentes. Esta etapa confirmou a adequação do conjunto de dados revisto, estabelecendo um equilíbrio ideal para avançar para a modelação.

2.10 Conclusão da Análise Exploratória

Com a conclusão da análise multivariada, garantimos que o *dataset* final está livre de redundâncias e pronto para a próxima fase. Este cuidado em preparar os dados garante não só a eficiência, mas também a precisão do modelo a ser desenvolvido.

3. Modelação

3.1 Separação de target e features (df_orto_y, df_orto_X)

Iniciamos o processo de modelação com a separação clara entre as variáveis *Features* (df_orto_X) e a *Target* (df_orto_y).

Este passo é essencial para garantir que o modelo seja treinado para prever a variável de interesse (mudança de grupo CPAK) com base em características relevantes dos utentes (como idade, peso, e outras avaliações clínicas pré e pós-cirúrgicas, etc.).

3.2 Técnicas de Balanceamento de Dados

Para enfrentar o desequilíbrio na distribuição das classes na nossa variável *target*, adotámos técnicas de *undersampling* e *oversampling*.

Estas abordagens garantem que o modelo seja robusto e justo, adaptando-se à realidade de um conjunto de dados desequilibrado e fornecendo insights valiosos para o contexto clínico.

Undersampling (df_orto_under)

Utilizámos *undersampling* para mitigar os desequilíbrios significativos na distribuição das classes.

Esta técnica envolve a redução aleatória da classe maioritária, prevenindo o viés do modelo em favor das classes mais frequentes. É particularmente benéfica para aprimorar a capacidade do modelo em prever de forma precisa a classe minoritária, que é essencial para o nosso estudo.

O *dataset* resultante foi estruturado com 70% pertencendo à *target* 0 e 30% à *target* 1, mantendo uma proporção que reflete a realidade dos dados disponíveis.

Oversampling (df_orto_over)

De modo similar, implementámos o *oversampling* para aumentar a representatividade da classe minoritária, ampliando os seus exemplos.

Este método contribui para um melhor equilíbrio no treino do modelo, permitindo-lhe aprender de forma mais eficiente as características associadas às mudanças no grupo CPAK. A divisão ajustada ficou em 60% para a *target* 0 e 40% para a *target* 1.

3.3. Aplicação dos modelos e geração de dados de classificação

Após a preparação e divisão dos *datasets*, procedemos à etapa de aplicação dos modelos de *machine learning*.

Utilizamos três tipos de modelos de *machine learning* — *Random Forest*, Regressão Logística e *Decision Tree* — com o objetivo de analisar e prever a mudança de grupo CPAK.

Esta fase envolveu treinar os modelos selecionados em todos os *datasets* previamente preparados (original, *undersampling* e *oversampling*).

Métodos de Divisão de Dados

Para cada configuração de modelo e *dataset*, usamos duas técnicas principais de divisão de dados, através da biblioteca '*sklearn*':

- **Split 70/30** - Os dados foram divididos em 70% para treino e 30% para testes de performance sendo esta uma prática padrão que permite ao modelo aprender com uma vasta gama de dados e, simultaneamente, reserva uma parcela significativa para testes que avalia a sua capacidade de generalizar em dados novos.
- **Cross-Validation** - Implementámos a validação cruzada para garantir uma avaliação mais robusta e menos enviesada da performance do modelo. Esta técnica ajuda a verificar a capacidade do modelo em generalizar para novos dados. É ideal para *datasets* limitados, o que é o caso da nossa base de dados.

Estas técnicas de divisão pretendem garantir a consistência e validade comparativa nos resultados dos testes.

Análise Detalhada dos Resultados dos Modelos

O nosso processo de modelação incluiu a geração de matrizes de confusão (visuais e numéricas) e de tabelas de classificação detalhadas, que destacam as métricas essenciais de avaliação do desempenho dos modelos, como precisão, *recall*, *F1-Score* e *AUC*.

Exportação de Dados de Classificação

Após a modelação e validação, os resultados de classificação para cada modelo e configuração de *dataset* foram exportados para arquivos Excel. Foi realizado um script para consolidar todos estes arquivos individuais num único documento para cada uma das abordagens, que foram posteriormente formatados (Avaliação_modelos_global.xlsx, Avaliação_modelos_m0.xlsx).

4. Avaliação

4.1 Interpretação das Métricas de Avaliação e Seleção do Melhor Modelo

4.1.1 Modelo Global (Variáveis Pré e Pós-Cirúrgicas)

Este modelo utiliza tanto dados recolhidos antes como após a cirurgia para prever se houve mudança no grupo CPAK. Ele é um modelo de classificação que pode ser considerado como **preditivo** e **analítico**, pois ajuda a entender quais os fatores/variáveis (antes e depois da cirurgia) que estão associados à alteração do grupo CPAK.

- *Target 0* – Não há mudança do CPAK entre o momento pré e pós cirúrgico
- *Target 1* – Há mudança do CPAK entre o momento pré e pós cirúrgico

Para atingir esta meta, analisámos a performance de diferentes modelos de *machine learning* aplicados aos 3 *datasets* anteriormente referidos.

Métricas Essenciais

- **Precisão (*Precision*):** essencial para minimizar os falsos positivos, garantindo que as previsões de mudança no grupo CPAK sejam confiáveis. Uma precisão alta significa que o modelo oferece previsões confiáveis, reduzindo a probabilidade de alertas falsos.
- **Sensibilidade (*Recall*):** métrica vital para reduzir os falsos negativos, assegurando que todas as mudanças reais de grupo são capturadas pelo modelo. Um alto *recall* na *target 1* é especialmente significativo no nosso contexto, pois indica que o modelo

efetivamente identifica a maioria dos utentes que realmente mudam de grupo, um aspeto crucial para a tomada de decisão.

- **F1-Score:** a métrica fornece um equilíbrio entre precisão e *recall*, refletindo efetivamente a precisão geral do modelo em identificar mudanças verdadeiras sem gerar excesso de alertas falsos. Um alto *F1-score* indica que o modelo é eficaz tanto em precisão quanto em sensibilidade, tornando-o uma ferramenta confiável para a tomada de decisão.
 - Perfeito = 1
 - Razoável > 0.7
 - Excelente > 0.8
- **AUC-ROC (Área Sob a Curva - Característica Operacional do Recetor):** Esta métrica avalia a capacidade do modelo em diferenciar corretamente os utentes que alteram o grupo CPAK, daqueles utentes mantêm o grupo CPAK no período pós-cirúrgico. Um alto valor de AUC indica uma excelente capacidade de discriminação, validando a eficácia do modelo em generalizar bem para novos dados.
 - AUC = 1: discriminação de classes perfeita.
 - AUC = 0.7: capta discriminação, mas não é excelente.
 - AUC = 0.5: O modelo não está a prever e a distinguir as classes, é uma classificação aleatória (sem poder preditivo).
 - AUC = 0: O modelo faz classificação inversa. Na prática classifica, mas classifica erradamente, ou seja, reais positivos prevê como negativos e reais negativos prevê como positivos.

Análise e Seleção do Modelo Global - Random Forest com Oversampling

Entre os modelos testados, o **Random Forest com Oversampling**, foi o considerado mais adequado para o nosso contexto, conforme indicado pelas métricas-chave que foram avaliadas em ambos os métodos de divisão de dados (Split 70/30 e CV):

- *Recall* elevada para a *target* 1 - indicando que quase todas as mudanças reais de grupo são identificadas.
- Alta Precisão para ambas as *targets* - mostrando que o modelo é altamente confiável tanto em prever mudanças quanto em confirmar não-mudanças.
- *F1-Score* robusto para as duas *targets* - refletindo um excelente equilíbrio entre precisão e sensibilidade.

- AUC superior a 0.9 - evidencia uma forte capacidade de discriminação em variados contextos de teste.

Estes resultados confirmam que o modelo de Random Forest com Oversampling não só atende às necessidades do projeto de prever mudanças de grupo CPAK de forma precisa e rápida, como também se destaca na sua capacidade de generalização, um aspeto fundamental para aplicação em cenários clínicos reais.

Original/Pre-processado					Oversampling					Undersampling				
Decision Tree					Random Forest					Regressão Logística				
Split 70/30					Split 70/30					Split 70/30				
	precisio	recall	f1-scor	support		precisio	recall	f1-scor	support		precisio	recall	f1-scor	support
0	0.875	0.913	0.8936	69	0	0.98551	0.932	0.9577	73	0	0.76	1	0.8636	19
1	0.25	0.1818	0.2105	11	1	0.89796	0.978	0.9362	45	1	0.4	0.5714	10	
accuracy	0.8125	0.8125	0.8125	0.8125	accuracy	0.94915	0.949	0.9492	0.9492	accuracy	0.7931	0.793	0.7931	0.7931
macro avg	0.5625	0.5474	0.5521	80	macro avg	0.94173	0.955	0.947	118	macro avg	0.88	0.7	0.7175	29
weighted avg	0.78906	0.8125	0.7997	80	weighted avg	0.95212	0.949	0.9495	118	weighted avg	0.84276	0.793	0.7629	29
AUC	0.54743				AUC	0.95464				AUC	0.72368			
Cross-Validation					Cross-Validation					Cross-Validation				
	precisio	recall	f1-scor	support		precisio	recall	f1-scor	support		precisio	recall	f1-scor	support
0	0.91453	0.9068	0.9106	236	0	0.93939	0.919	0.9293	236	0	0.82353	0.836	0.8296	67
1	0.29032	0.3103	0.3	29	1	0.88272	0.911	0.8966	157	1	0.60714	0.586	0.5965	29
accuracy	0.84151	0.8415	0.8415	0.8415	accuracy	0.91603	0.916	0.916	0.916	accuracy	0.76042	0.76	0.7604	0.7604
macro avg	0.60243	0.6086	0.6053	265	macro avg	0.91105	0.915	0.9129	393	macro avg	0.71534	0.711	0.7131	96
weighted avg	0.84622	0.8415	0.8438	265	weighted avg	0.91675	0.916	0.9162	393	weighted avg	0.75816	0.76	0.7592	96
AUC	0.60856				AUC	0.91516				AUC	0.56049			

Imagem 2 – Tabelas de Classificação dos modelos seleccionados para cada dataset – Modelo Global

4.1.2 Modelo “Momento 0” (Variáveis pré-cirúrgicas)

Razão para Excluir Variáveis Pós-90 Dias

Este modelo usa exclusivamente variáveis recolhidas antes da cirurgia. Consideramos um modelo preditivo puro focado em estimar a mudança de CPAK antes da realização da cirurgia.

Ao excluir as variáveis registadas após os 90 dias, permite focar a análise em características/variáveis acessíveis antecipadamente, permitindo que o modelo seja utilizado em previsões preditivas antes de qualquer intervenção cirúrgica e, assim, prever a alteração do grupo CPAK antes de ter acesso aos dados pós-operatórios.

Vantagens desta Abordagem

Fornece insights valiosos sobre a possível eficácia da intervenção, potencialmente permitindo decisões médicas mais rápidas e fundamentadas.

- *Target 0* – Não há mudança do CPAK entre o momento pré e pós cirúrgico
- *Target 1* – Há mudança do CPAK entre o momento pré e pós cirúrgico

Análise e Seleção do Modelo “Momento 0” - Random Forest com Oversampling

Entre os modelos testados, o **Random Forest com Oversampling**, foi o considerado mais adequado para o nosso contexto, conforme indicado pelas métricas-chave, que foram avaliados em ambos os métodos de divisão de dados (Split 70/30 e CV):

- *Recall* elevado para a target 1: indica que o modelo é eficaz em identificar utentes que realmente experimentarão uma mudança no grupo CPAK. Esta capacidade é essencial para minimizar os casos de falsos negativos, ou seja, utentes que o modelo indicaria como não tendo mudado de grupo quando, na verdade, haverá mudança.
- Alta precisão para ambas as classes (0 e 1): assegura que as previsões do modelo sobre mudanças no grupo CPAK são confiáveis, minimizando os riscos de falsos positivos que poderiam, por exemplo, levar a cirurgias a utentes que, na realidade, não vão mudar de grupo CPAK.
- Elevado AUC (0.87686 no *split 70/30* e 0.90134 no *cross-validation*): mostra uma forte capacidade de discriminação entre as classes, indicando que o modelo pode diferenciar eficazmente entre utentes que mudarão de grupo e aqueles que permanecerão no mesmo.
- F1-Score: reflete uma boa precisão geral do modelo em identificar mudanças verdadeiras, sem gerar alertas falsos, isto é, capacidade do modelo de detetar de forma precisa os verdadeiros positivos e negativos.

Escolha do Modelo

O modelo Random Forest com Oversampling foi selecionado como o mais adequado devido ao seu bom desempenho em todas as métricas-chave face aos restantes modelos.

Esta variável aparece como a mais importante no modelo. Isto sugere que o desempenho na prova de marcha no pós-operatório é decisivo para prever a mudança no grupo CPAK. A resposta ao tratamento cirúrgico, em termos de melhoria na mobilidade, é um forte indicador de mudança positiva.

2. PM6_0

A segunda variável mais importante é a "PM6" medida no momento zero (pré-cirurgia).

Isto indica que o estado inicial do utente também é essencial para prever a mudança de grupo. A combinação das medições pré e pós-operatórias proporciona uma visão abrangente de como os utentes respondem à cirurgia em relação a esta métrica específica.

3. Sexo

De forma surpreendente, o sexo do utente é a terceira variável mais importante. Esta ordem de importância pode refletir diferenças biológicas ou de tratamento entre géneros que influenciam a probabilidade de mudança de grupo após a cirurgia.

4. Fle_0 e Fle_90

"Fle_0" e "Fle_90" representam medições da flexão do joelho no momento zero e 90 dias, respetivamente.

Ambas são importantes, mas a "Fle_90" tem uma importância ligeiramente menor que a "Fle_0". Esta ligeira discrepância pode indicar que as condições iniciais de flexibilidade são mais preditivas de mudança do que as mudanças pós-operatórias.

4.2.2. Modelo Momento 0

1. PM6_0
2. Fle_0
3. Peso
4. Sexo
5. EVA_0

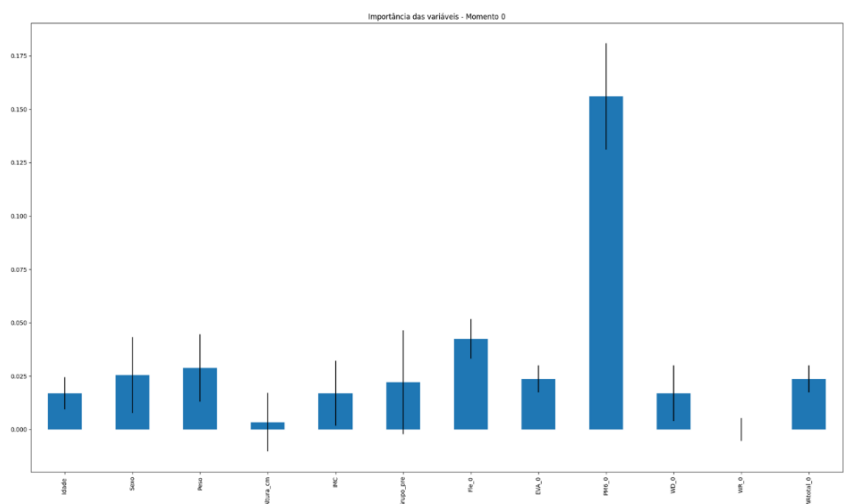


Gráfico 8 – Gráfico de importância das variáveis no Modelo Momento_0

1. **PM6_0:**

A "PM6_0", no modelo Momento_0, é a medida crucial para prever a mudança de grupo CPAK. É indicativo de que o estado físico inicial do utente, em especial a sua capacidade de marcha antes da cirurgia, é um forte indicativo de que mudará de grupo ou não.

2. **Fle_0:**

Esta variável é consistente na sua importância tanto no Modelo_0 quanto no Modelo_Global. Ela indica que a flexibilidade do joelho antes da cirurgia é um preditor significativo da mudança de grupo CPAK. A "Fle_90" também é considerada no Modelo_Global, mas aqui, a "Fle_0" tem uma relevância maior, destacando a importância das condições pré-operatórias para a mudança de grupo.

3. **Peso:**

O peso do utente é uma variável importante no Modelo_0, o que sugere que fatores físicos podem influenciar os resultados, ou seja, a mudança de grupo CPAK.

4. **Sexo:**

Assim como no Modelo_Global, o sexo do utente aparece como uma variável importante. Isto pode refletir que o género pode influenciar as mudanças de grupo CPAK.

5. **EVA_0 (Escala de Dor):**

A intensidade da dor pré-operatória é um indicador da urgência e gravidade da condição que está a ser tratada, sendo um preditor a valorizar de mudanças no grupo CPAK.

De destacar que a variável WR_0, que reflete e classifica a rigidez do joelho, no momento pré-operatório, não apresentou impacto na capacidade do modelo fazer previsões precisas na mudança de CPAK.

5. Conclusão

Consideramos que concluímos com êxito o desenvolvimento e a avaliação de dois modelos preditivos de classificação destinados a prever mudanças na classificação CPAK decorrentes da cirurgia de artroplastia do joelho.

A abordagem detalhada em três fases permitiu não apenas a análise e a compreensão aprofundada dos dados através da Análise Exploratória de Dados, mas também o ajuste fino de modelos preditivos durante a fase de Modelação, resultando em previsões robustas e confiáveis.

A decisão de gerar um segundo modelo, momento 0, foi uma necessidade que surgiu ao longo do desenvolvimento do trabalho pois consideramos que, caso replicássemos este projeto na prática clínica, tornar-se-ia um modelo preditivo puro se conseguisse prever a alteração do grupo CPAK mesmo antes da cirurgia.

Simulando um contexto real, isto tornar-se-ia útil para apoiar a decisão dos profissionais de saúde na sua prática clínica e, conseqüentemente, trazer ganhos para os utentes.

Os modelos "Global" e "Momento 0", embora centrados em conjuntos de variáveis distintos — pré e pós-operatórias, e exclusivamente pré-operatórias, respetivamente — demonstraram ser ferramentas valiosas, fornecendo *insights* que são críticos para a tomada de decisão.

As métricas de desempenho destacaram a capacidade dos modelos em identificar com precisão as mudanças de CPAK, reforçando sua aplicabilidade prática.

Acreditamos que a esquematização no nosso raciocínio e o desenvolvimento do projeto por etapas demonstrou-se um método essencial para a conclusão do trabalho e alcance dos objetivos.

No entanto, consideramos que, no contexto dos modelos de *machine learning*, uma análise *multitarget* poderia oferecer *insights* mais aprofundados. Este potencial para descobertas mais ricas deixou-nos curiosos e entusiasmados para explorar mais técnicas.

O grupo mostrou-se unido e empenhado na realização deste projeto pois consideramos que o desenvolvimento do trabalho se demonstrou desafiante, mas bastante enriquecedor para a consolidação dos conhecimentos base em *machine learning*.