INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Dados de utilização de telefones móveis

Dados na Ciência, Gestão e Sociedade

1º Semestre - 2024/2025

AUTORES	Nº DE CARTÃO
Leandro Bernardo	125293
Tomás Freire	130651
Tomás Bonzinho	130641

1 Introdução

A evolução da tecnologia e das redes moveis transformou os dispositivos moveis em componentes essenciais do dia a dia, não só como um dispositivo de entretenimento

Estes dispositivos são altamente complexos e enviam grandes volumes de dados pessoais através de tecnologias sofisticadas. A análise desses dados permite identificar padrões de uso, prever comportamentos e categorizar os utilizadores como por exemplo podemos responder à pergunta "Quais são as tendências no uso de dispositivos móveis para certas profissões?" e podemos orientar o desenvolvimento de aplicações moveis baseado nos padrões de utilizador e consumo.

Este relatório analisa um dataset de 1049 amostras de dados simulados através de padrões realistas de consumo móvel, estes dados foram retirados do site Kaggle (khorasani, s.d.).

Este relatório faz parte do projeto final da unidade curricular dados na ciência, gestão e sociedade coordenada por Fernando Batista, Ana de Almeida e Vitor Bastos-Fernandes.

O projeto implementa a metodologia do CRISP-DM, sendo as suas fases:

- Business Understanding: Entendimento do contexto do projeto, os seus objetivos e os critérios de sucesso.
- Data Understanding: Compreensão dos dados que são recolhidos.
- Data Preparation: Pré-processamento dos dados de forma que estejam prontos para a modelação.
- Modeling: Aplicação de diversas técnicas de modelação de forma a manipular os dados e obter conclusões.
- Avaliação: Determinação da qualidade do modelo desenvolvido e a sua eficiência na utilização no projeto.
- Deployment**:** A fase final, envolve o início da utilização do modelo que foi criado, produzindo os resultados para os utilizadores finais.

2 Descrição do conjunto de dados

O dataset do projeto tem 12 variáveis: 10 features, 1 metadado e 1 target.

Na figura 1, estão expostas a informação sobre todas as variáveis do dataset, o seu significado, o tipo de dado, a sua função e a sua respetiva variação de valores.

Variáveis	**Significado**	**Tipo de dado**	**Função**	**Variação de valores**
User ID	Identificador único de cada utilizador	Numérico	Metadado	[1 - 1049]
Device Model	Modelo do telemóvel do utilizador	Categórico	Feature	[iPhone 12, Xiaomi Mi 11, Samsung Galaxy S21, OnePlus 9, Google Pixel 5]
Operating System	Sistema operativo do telemóvel (IOS ou Android)	Categórico	Feature	[Android, IOS]
App Usage Time (min/day)	Tempo diário gasto nas aplicações (minutos)	Numérico	Feature	[30 - 598]
Screen On Time (hours/day)	Tempo diário gasto no telemóvel (horas)	Numérico	Feature	[1 - 12]
Battery Drain (mAh/day)	Bateria consumida (mAh)	Numérico	Feature	[7.3 - 3010]
Number of Apps Installed	Total de aplicações instaladas no dispositivo	Numérico	Feature	[10 - 99]
Data Usage (MB/day)	Dados consumidos diariamente no dispositivo	Numérico	Feature	[102 - 2497]
Age	Idade do utilizador	Numérico	Feature	[18 - 59]
Gender	Género do utilizador	Categórico	Feature	[Male, Female]
Profession	Profissão do utilizador	Categórico	Feature	[Unemployed, Student, Professor, Law, Health
User Behavior Class	Classificação de padrões de utilização móvel	Categórico	Target	[1, 2, 3, 4, 5]

Figura 1 - Descrição dos dados

Nós escolhemos a variável user behavior class como target porque esta variável reflete os padrões de consumo móvel dos utilizadores, pode ser utilizada para prever os comportamentos dos utilizadores, isso permite criar estratégias para a personalização de experiências, economizar recursos e produzir funcionalidades direcionadas a essas classes.

A variável "user behavior class" é composta por 5 categorias [1-5], para identificar as variáveis relevantes para o target, utilizamos o Chi-Square (Chi2), este método avalia a dependência entre uma feature e o target com base no teste de qui-quadrado.

Como mostra a figura 1, as variáveis relevantes para o target são o número de aplicações instaladas, tempo de utilização, uso de bateria, uso de dados e profissão do utilizador.

Ranks

		#	χ^2
1	C Profession	5.0	663.681
2	N App Usage Time (min/day)		447.849
3	N Number of Apps Installed		447.264
4	N Battery Drain (mAh/day)		429.390
5	N Screen On Time (hours/day)		420.113
6	N Data Usage (MB/day)		417.455
7	C Gender	2.0	4.100
8	C Device Model	5.0	3.273
9	N Age		1.305
10	C Operating System	4.0	0.983

Figura 2 – Rank de Chi2

Além disso, através da correlação Pearsons encontrámos dependências entre as variáveis numéricas, encontrámos especificamente a existência de uma forte interdependência entre as variáveis numéricas que categorizam o target (número de aplicações instaladas, tempo de utilização, uso de bateria, uso de dados) e mostra irrelevância da idade como mostra a Figura 3.

Pearson correlation

1	+0.756	App Usage Time (min/day)	:	Number of Apps Installed
2	+0.751	App Usage Time (min/day)	:	Battery Drain (mAh/day)
3	+0.745	Battery Drain (mAh/day)	:	Number of Apps Installed
4	+0.732	Number of Apps Installed	:	Screen On Time (hours/day)
5	+0.732	App Usage Time (min/day)	:	Screen On Time (hours/day)
6	+0.718	Battery Drain (mAh/day)	:	Screen On Time (hours/day)
7	+0.704	App Usage Time (min/day)	:	Data Usage (MB/day)
8	+0.699	Data Usage (MB/day)	:	Number of Apps Installed
9	+0.683	Data Usage (MB/day)	:	Screen On Time (hours/day)
10	+0.664	Battery Drain (mAh/day)	:	Data Usage (MB/day)
11	-0.021	Age	:	Battery Drain (mAh/day)
12	+0.018	Age	:	Number of Apps Installed
13	+0.013	Age	:	Screen On Time (hours/day)
14	+0.009	Age	:	App Usage Time (min/day)
15	-0.003	Age	:	Data Usage (MB/day)

Figura 3 - Outras correlações

A Figura 4 apresenta o tratamento dos dados estatísticos, mostra as métricas relevantes como a média, a moda, mediana e a sua respetiva dispersão.

Nome	Média	Moda	Mediana	Dispersão
Screen On Time	5.565	1.6	5.2	0.570

Number of Apps	51.47	10	51	0.51
App Usage Time	283.66	64	257	0.62
Data Usage	1017.17	281	910	0.66
Battery Drain	1541.456	2447	1523.5	0.528

Figura 4 - Tratamento de dados estatísticos

Tratamento dos dados categóricos (Moda, frequência absoluta, dispersão, frequência relativa)

O dataset tem 10 features com 9% dos dados em falta e 1 target com 2% dos dados em falta, no total existem 123 dados em falta (99 dados da features + 24 dados do target) como está representado na tabela em baixo.

Tipo de dados	**Quantidade dos dados omissos (%)**
Age	25 (2%)
Device model	10 (1%)
Operating System	38 (4%)
Profession	16 (2%)
User behavior class	24 (2%)
App Usage Time	2 (~0%)
Screen On Time	2 (~0%)
Battery Drain	1 (~0%)
Gender	5 (~0%)

Figura 5 - Dados omissos

Olhando para a matriz de dados na figura 6, é possível verificar que existem modelos de dispositivos sem sistema operativo definido, como esses modelos são sistemas embebidos é facilmente alterável.

	User Behavior Class	Device Model	Operating System
1	1	iPhone 12	?
2	5	Google Pixel 5	?
3	1	Google Pixel 5	?
4	5	OnePlus 9	?
5	1	Google Pixel 5	?
6	3	Samsung Galaxy S21	?
7	5	OnePlus 9	?
8	4	Google Pixel 5	?
9	1	Google Pixel 5	?
10	?	Google Pixel 5	?
11	3	Google Pixel 5	?
12	5	Samsung Galaxy S21	?
13	1	Samsung Galaxy S21	?
14	2	OnePlus 9	?

Figura 6 - modelos com sistemas operativos omissos

Além disso, é também possível verificar a existência de inconsistências entre modelos e o respetivo sistema operativo, como está demonstrado na figura 6 onde é possível visualizar sistemas embebidos que utilizam o sistema operativo Android a utilizar o iOS.

User Behavior Class	Device Model	Operating System
3	Samsung Galaxy S21	iOS
3	Samsung Galaxy S21	iOS
4	Samsung Galaxy S21	iOS
1	Samsung Galaxy S21	iOS
5	Samsung Galaxy S21	iOS
5	Samsung Galaxy S21	iOS
5	Samsung Galaxy S21	iOS
1	Samsung Galaxy S21	iOS
2	Samsung Galaxy S21	iOS
5	Samsung Galaxy S21	iOS
5	Samsung Galaxy S21	iOS
3	Samsung Galaxy S21	iOS
1	Samsung Galaxy S21	iOS
4	Samsung Galaxy S21	iOS
3	Xiaomi Mi 11	iOS
4	Xiaomi Mi 11	iOS

Figura 7 - Modelo de telemóvel com sistema operativo incorreto

3 Preparação dos dados

3.1 Seleção de dados

O primeiro passo da preparação dos dados é selecionar os dados relevantes para a modelação, para concretizar este passo utilizamos a técnica de Chi-Square como demonstrado na Figura 2 para determinar as variáveis com maior correlação com o target, decidimos então que as variáveis relevantes são o número de aplicações instaladas, tempo de utilização, uso de bateria, uso de dados e profissão.

3.2 Limpeza de dados

Na Figura 6 e Figura 7 mostram-se dados de modelos de dispositivos móveis com sistema operativo inconsistente com o tipo de modelo ou simplesmente valores omissos no sistema operativo do modelo, para limpar o sistema operativo utilizamos um script de Python que atribui a todos modelos os móveis o sistema operativo correto apropriado.

```
def python_script(in_data):
 1 import numpy as np
 2 from Orange.data import Domain, Table
 3
 4 input_data = in_data
 5
 6
 7 output_data = input_data.copy()
 8
 9 device_model_col = output_data.domain.index("Device Model")
10 operating_system_col = output_data.domain.index("Operating System")
11
12 Android  = ["Google Pixel 5","OnePlus 9","Samsung Galaxy S21","Xiaomi Mi 11"]
13 for row in output_data:
14     if row[device_model_col] == "iPhone 12":
15         row[operating_system_col] = "iOS"
16     if row[device_model_col] in Android:
17         row[operating_system_col] = "Android"
```

Figura 8 – Alteração do sistema operativo

No mesmo script de Python, realizamos a conversão de dados de App Usage Time em minutos por dia para horas por dia para ficar na mesma unidade de Screen On Time (ver figura 1), para isso multiplicamos todas as instâncias do App Usage Time por 0.6 como mostra a figura 9.

```
20 App_usage_time_col = output_data.domain.index("App Usage Time (min/day)")
21 for row in output_data:
22     row[App_usage_time_col] *= 0.6
23
```

Figura 9 - Conversão unitária

Em geral, a limpeza dos dados omissos foi feita baseado na relevância dos dados para o modelo.

Em relação aos dados omissos pertencerem a variáveis relevantes (3.1), nós tomamos a decisão de fazer a remoção desses dados porque de certa maneira iriamos enviesar modelo final e o seu resultado se alterássemos os dados mais correlacionados com o target.

E se os dados omissos pertencentes a variáveis não relevantes para o target, decidimos fazer uma média dos dados totais e substituir os dados omissos porque a remoção implica remover a instância inteira o que implica remover dados não omissos de variáveis relevantes, ou seja, haveria uma perda injustificada de dados uteis para modelação.

A operação sobre limpeza de dados em geral está representada na figura 10.

Variáveis	**Operações**
Device model	Média
Operating System	Média
App Usage Time	Remover
Screen On Time	Remover
Battery Drain	Remover
Age	Média
Gender	Média
Profession	Remover
User Behavior Class	Remover

Figura 10 - Operações sobre dados

4 Analise exploratória de dados

A análise exploratória de dados é a etapa em que se examinam os dados para identificar padrões e compreender melhor a sua estrutura através de técnicas de estatística descritiva e visualizações gráficas.

A figura mostra as profissões em relação à sua frequência e também mostra as categorias de padrões de uso movel relacionadas, podemos concluir que a profissão com maior frequência de utilizadores de categoria 5 é direito (Law).

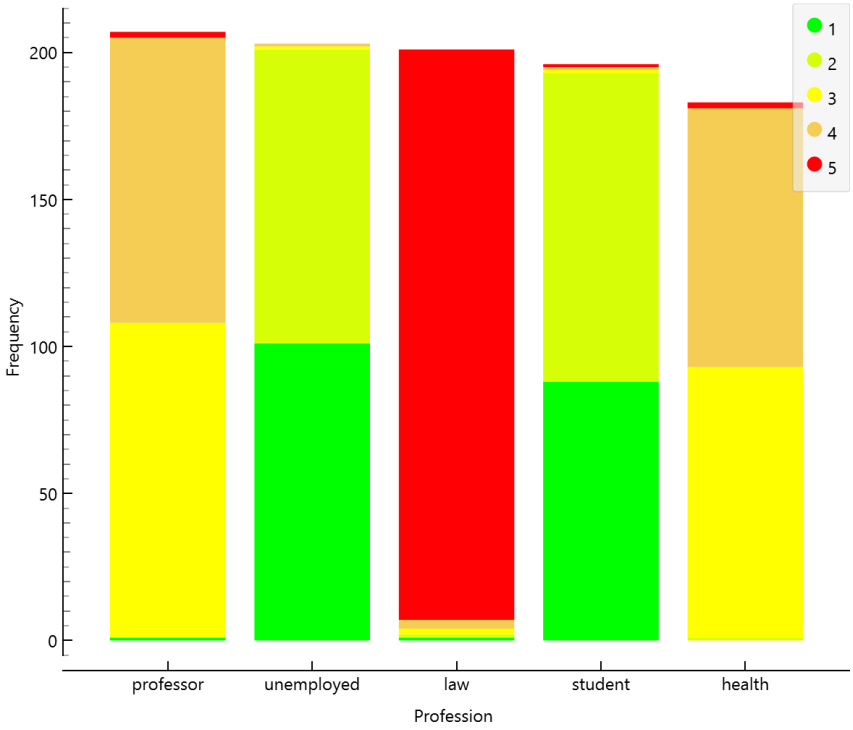

Figura 11 - Histograma de frequência de utilização face à profissão.

Através da figura 13, conseguimos facilmente entender que há claros agrupamentos de utilizadores, diretamente relacionados com os 5 níveis de utilização. Também é notório que a profissão que mais utiliza os dispositivos é direito (Law), informação esta, já obtida na figura 11. Concluímos também que os utilizadores menos frequentes são os desempregados, tendo uma grande concentração destes no nível 1 de utilizador.

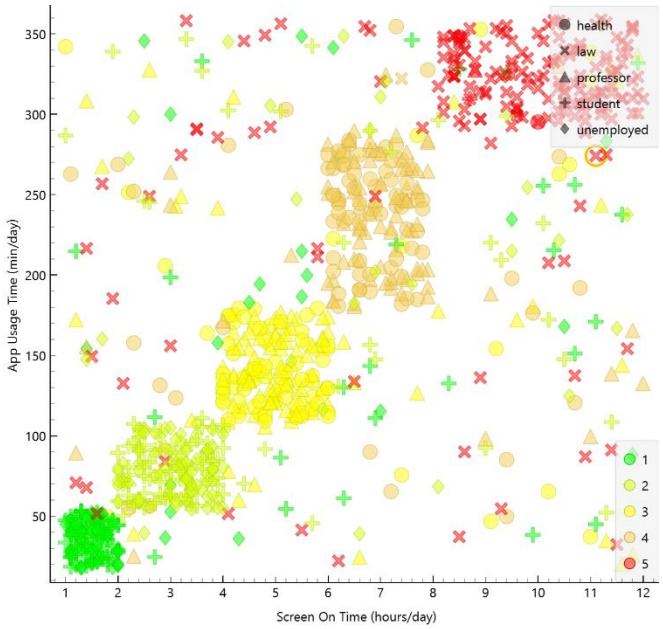

Figura 12 - Scatter Plot, apresenta a profissão e, o tempo de utilização diário em função do tempo de utilização por aplicação, intersetando-os com cada utilizador.

O histograma da figura 13 mostra uma tendência entre a maior utilização de dados moveis e o aumento da categoria, a partir dos ~1500 Mb há um crescimento que tende à linearidade da categoria 5, mas em geral as funções tendem a se expressar como logarítmicas.

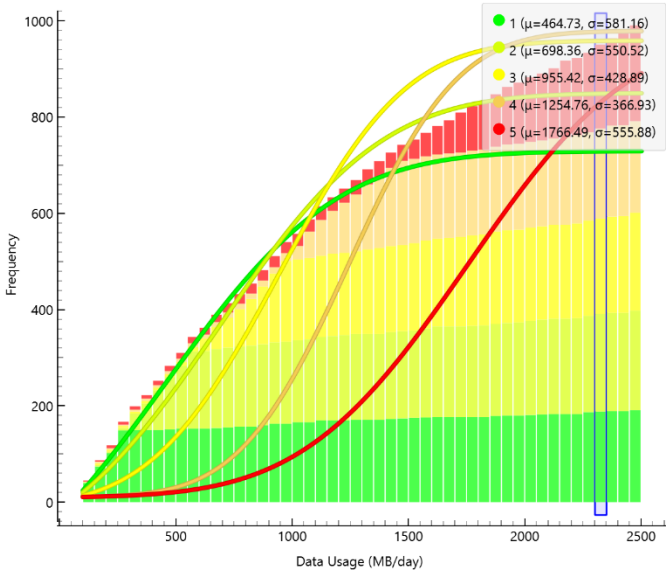

Figura 13 - histograma, data usage e a sua frequencia relacionado com as categorias de padrões de utilização móvel

5 Modelação

O nosso projeto utiliza tarefas supervisionadas baseadas em classificação, nós utilizamos duas técnicas de classificação nomeadamente o CN2 Rule Induction e o Decision Tree.
Optámos por usar o modelo Tree porque este é adequado a datasets com poucos dados, é simples e apresenta elevada precisão, este funciona através estrutura hierárquica onde são criados vários pontos de decisão. O modelo funciona através de uma estrutura hierárquica, em que são criados vários pontos de decisão chamados de "nós", o resultado da decisão em cada nó determina o caminho a ser seguido. Os diferentes caminhos possíveis são os "ramos" da árvore.(Como funciona o algoritmo Árvore de Decisão, s.d.)
Utilizamos também o modelo CN2 Rule Induction porque é um algoritmo eficiente em dados pequenos e ruidosos como o nosso dataset e também para comparar a eficiência com o modelo tree, O CN2 cria uma lista ordenada de regras de acordo com precisão e significância, que separam o target de acordo com as features, ele utiliza uma função heurística para avaliar a qualidade das regras criadas.
Os dados utilizados para o modelo já foram definidos na secção 3.1: número de aplicações instaladas, tempo de utilização, uso de bateria, uso de dados e profissão, os dados foram divididos em um conjunto de treino usando uma amostra com um rácio fixo de 90% e um conjunto teste composto dos dados restantes.

5.1 Discussão dos resultados

No modelo tree, o impacto da variável Screen On Time foi não relevante para o resultado do modelo como mostra a figura 14, ou seja, podia ter sido removida e o resultado seria similar, enquanto no modelo CN2 a variável Screen On Time afeta os resultados.

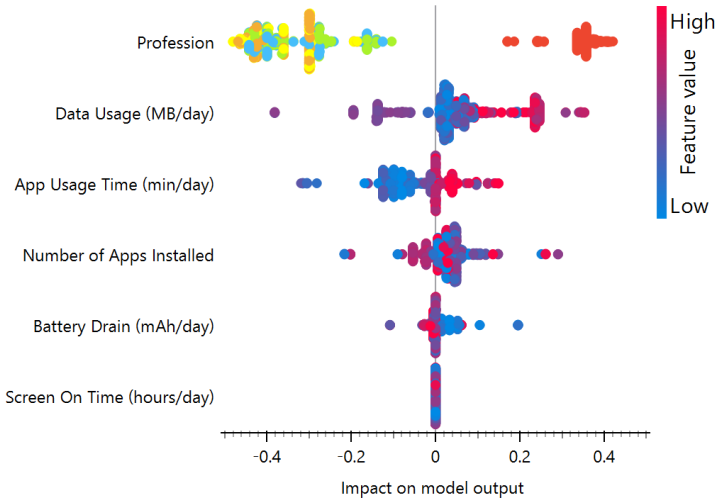

Figura 14 - impacto das variáveis no modelo tree

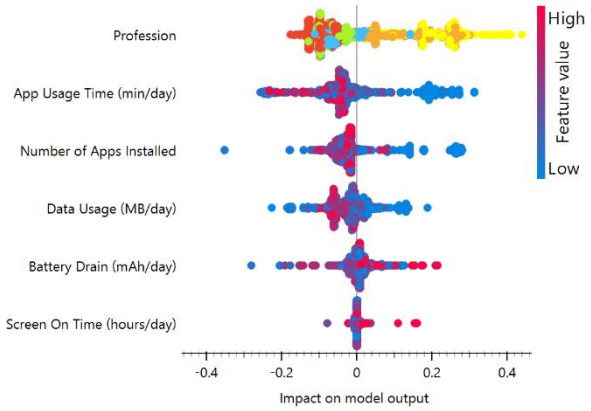

Figura 15 - impacto das variáveis no modelo CN2

Através da matriz de confusão, conseguimos verificar a eficiência dos modelos de previsão. A matriz de confusão mostra que a tree acertou 95,4% quando a categoria esperada é a 1, acertou 96,7% quando a categoria esperada é 2, acertou 96,3% quando a categoria esperada é 4 e acertou 100% quando a categoria esperada é 5 como demonstra a figura 16.

Predicted

Actual	1	2	3	4	5	Σ
1	95.4 %	1.9 %	0.5 %	0.0 %	0.0 %	193
2	3.6 %	96.7 %	0.5 %	0.0 %	0.0 %	216
3	0.5 %	0.5 %	93.4 %	3.7 %	0.0 %	208
4	0.0 %	0.5 %	4.2 %	96.3 %	0.0 %	194
5	0.5 %	0.5 %	1.4 %	0.0 %	100.0 %	199
Σ	197	215	213	191	194	1010

Figura 16 - matriz de confusão da tree

A partir da matriz de confusão do modelo CN2 como mostra a Figura 17, conseguimos concluir que existe uma maior taxa de acerto na categoria 1 com 97,9% de previsões corretas comparado com o modelo tree (95,4%), o mesmo se pode deduzir das próximas 3 categorias (2,3 e 4) com uma taxa de acerto de 97,3%, 98,5% e 99% comparado com o tree (96,7%, 93,4%, 96,3%).

Na categoria 5, acontece o inverso em relação às outras categorias, o modelo tree é superior como uma taxa de 100% enquanto o modelo CN2 só tem 98,5% de previsões corretas.

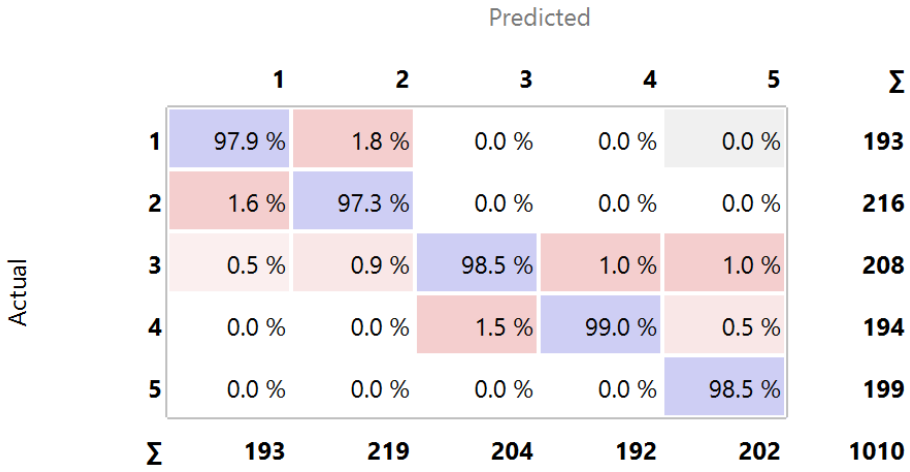

Figura 17 - matriz de confusão da CN2 RULE INDUCTION

6 Avaliação

Os modelos testados foram o CN2 Rule Induction e o Tree, utilizando amostragem aleatória (Random sampling) com um conjunto de treino de 66% dos dados e repetição do processo de treino/teste 100 vezes, para garantir maior confiabilidade dos resultados. Além disso, a validação foi estratificada para assegurar uma distribuição proporcional das classes em cada subconjunto.

O modelo Tree obteve um desempenho superior em relação ao CN2 Rule Induction na maioria das métricas como mostra a Figura 17:

- Em termos de AUC (area under the curve), o modelo CN2 superou com um valor de 0,947 enquanto o tree teve 0.927, isto mostra que o tree tem capacidade inferior em distinguir entre classes neste dataset comparado com o CN2.
- Com resultos opostos, no accuracy (CA), o Decision Tree obteve 0.856 superando o CN2 que só teve accuracy de 0.827 mostrando que o tree teve maior percentagem de acertos de previsões do que o CN2.
- A métrica de F1, Prec (precisão) e recall mostram valores superiores no Tree todos iguais a 0,856 enquanto no CN2 os valores são 0,827, 0,828 e 0,827 respetivamente mostrando que existe um equilíbrio melhor média harmônica de precisão e recall (F1), uma maior proporção de verdadeiros positivos sobre o total de positivos previstos (Prec) e uma maior proporção de verdadeiros positivos sobre o total de positivos reais (Recall) no Tree.
- No MCC (Matthews Correlation Coefficient), o Tree (0,819) superou-se ao CN2(0,784) mostrando uma avaliação mais balanceada entre verdadeiros e falsos para o Tree do que para o CN2.

Model	AUC	CA	F1	Prec	Recall	MCC
CN2 Rule Induction	0.947	0.827	0.827	0.828	0.827	0.784
Tree	0.927	0.856	0.856	0.856	0.856	0.819

Figura 18 - test and score dos modelos

7 Conclusão

Neste projeto, aplicamos todas as etapas do método de CRISP-DM desde o business understanding até à etapa de avaliação.

Ao analisar o dataset sobre a utilização de dispositivos móveis descobrimos as variáveis são relevantes para classificar diferentes categorias de utilização móvel através da técnica de Chi2: número de aplicações instaladas, tempo de utilização, uso de bateria, uso de dados e profissão.

Exploramos a eficiência de técnicas de classificação em prever comportamentos com diferentes modelos de previsão nomeadamente o Tree e o CN2, concluindo que o Tree obteve um desempenho ligeiramente superior na grande parte das métricas.

Através da análise exploratória, conseguimos concluir que a profissão de Direito apresenta uma maior frequência em categorias de elevada utilização, ou seja, os trabalhadores de direito têm uma maior chance de usar dispositivos moveis por um período prolongado.

Com este projeto aprendemos a aplicar o conteúdo lecionado na unidade curricular e conseguimos conhecimentos em pesquisa, em diferentes ciência de dados como a utilização do software do Orange.

Agradecemos aos professores Fernando Batista, Ana de Almeida e Vitor Bastos-Fernandes, que coordenaram esta unidade curricular e nos distribuíram as ferramentas e o conhecimento necessário para a realização deste projeto. Foi uma oportunidade enriquecedora para aplicar os conceitos adquiridos ao longo da unidade curricular "Dados na Ciência, Gestão e Sociedade".

8 Referências

Como funciona o algoritmo Árvore de Decisão. (s.d.). Obtido de Didática Tech: https://didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/

khorasani, V. (s.d.). *Mobile Device Usage and User Behavior Dataset*. Obtido de https://www.kaggle.com/datasets/valakhorasani/mobile-device-usage-and-user-behavior-dataset.

PETER CLARK, T. N. (25 de 10 de 1988). The CN2 Induction Algorithm . *Kluwer Academic Publishers*, p. 262. Obtido de https://www.inf.ufrgs.br/~engel/data/media/file/Aprendizagem/CN2.pdf