

2022/2023

DCGE - Projeto: Tema D



**Trabalho realizado no âmbito da UC Dados na ciência,
gestão e sociedade, da Licenciatura em Ciência de Dados**

Docente: Ana Maria de Almeida

Grupo 18, CD02:

Catarina Lameira, n.º 111390

Gonçalo Girão, n.º 111515

Margarida Salgueiro, n.º 111566

João Magarça, n.º 111640

Francisco Lourenço, n.º 111183

Índice:

Introdução.....	3
Descrição do conjunto de dados.....	4
Descrição do Tratamento/Preparação e Exploração dos Dados.....	5
Análise dos modelos de Machine Learning.....	10
Conclusão.....	12
Webgrafia.....	13

Introdução

Este projeto tem como objetivo pôr em prática os conteúdos e matérias lecionados na unidade curricular de Dados na Ciência Gestão e Sociedade.

A base de dados que nos foi atribuída é relativa ao desempenho dos alunos de duas escolas secundárias. Foram-nos fornecidos dados que incluem as notas, as características demográficas, sociais e escolares dos alunos, recolhidos através de relatórios escolares e questionários.

Iremos basear o nosso trabalho na metodologia de CRISP-DM (*Cross-Industry Standard Process for Data Mining*), que contém 6 etapas, sendo a primeira o Data Understanding, onde o objetivo é identificar o problema a ser resolvido. Como não nos foi fornecido nenhum problema, houve necessidade de definir questões para a especificação do trabalho e ao explorarmos as variáveis que nos foram fornecidas escolheram-se as seguintes:

1) Será que a família tem influência na decisão do aluno quanto ao ensino superior?

2) Será que a gestão do tempo dos alunos influencia o sucesso dos mesmos?

Normalmente, desde cedo, cada aluno cria os seus próprios objetivos em relação ao seu nível de escolaridade. Esses objetivos, na nossa opinião, estão muitas vezes ligados com os exemplos que cada um observa em casa. É muito mais provável que um aluno em que tanto o pai como a mãe tenham seguido o ensino superior também tenha essa ambição, do que um aluno em que os pais apenas tenham o ensino secundário, ou menos. Foi no sentido de estudar esta relação e comprovar, ou não, esta nossa opinião que surgiu a questão 1).

A questão 2) foi desenvolvida a partir do conhecimento, por parte dos elementos do grupo, de estudos/artigos nesse sentido, como por exemplo, o artigo de Melim e Veiga (2007), intitulado *Organização dos tempos de estudo em jovens alunos*. Consideramos ser um tema pertinente e que suscita opiniões divergentes, o que provoca um interesse acrescido nesta temática.

As restantes 5 etapas serão desenvolvidas ao longo do trabalho.

Para a realização deste trabalho usamos o Software Orange, uma ferramenta open-source, que nos foi bastante útil para a visualização dos dados, Data Mining e Análise de Dados.

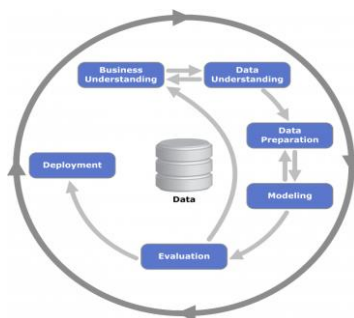


Figura 1- Metodologia CRISP-DM

Descrição do Conjunto de Dados

Analisando a base de dados fornecida, contamos com os seguintes atributos:

- *School* - escola do estudante;
- *Sex* - sexo do estudante;
- *Age* - idade do estudante;
- *Address* - morada do estudante;
- *Famsize* - tamanho da família ;
- *Pstatus* - estado de coabitação dos pais;
- *Medu* - nível de ensino da mãe;
- *Fedu* - nível de ensino do pai;
- *Mjob* - trabalho da mãe;
- *Fjob* - trabalho do pai;
- *Reason* - razão pela qual escolheu a escola;
- *Guardian* - encarregado de educação;
- *Traveltime* - tempo de viagem de casa à escola;
- *Studytime* - tempo de estudo semanal;
- *Failures* - número de reprovações em anos anteriores;
- *Schoolsup* - apoio educacional extra;
- *Famsup* - apoio educacional familiar;
- *Paid* - explicações;
- *Activities* - atividades extra-curriculares;
- *Nursery* - frequentou a creche;
- *Higher* - quer ingressar no ensino superior;
- *Internet* - acesso à internet em casa;
- *Romantic* - está num relacionamento amoroso;
- *Famrel* - qualidade das relações familiares;
- *Freetime* - tempo livre depois da escola;
- *Goout* - sair com amigos;
- *Dalc* - consumo de álcool os dias de escola;
- *Walc* - consumo de álcool durante o fim de semana;
- *Health* - estado de saúde atual;
- *Absences* - número de faltas;
- *G1* - nota do primeiro período;
- *G2* - nota do segundo período;
- *G3* - nota final.

Descrição do Tratamento/Preparação e Exploração dos Dados (Data Understanding and Data Cleaning)

No widget “*Edit Domain*” ligado ao widget “*File*”, começamos por converter variáveis contínuas em atributos categóricos. Para isso, no “*Type*” convertemos as variáveis que se encontravam em “*Number*” em “*Categorical*”, uma vez que as variáveis que nos foram atribuídas com o “*Type Number*” são valores Naturais e que não percorrem uma sequência contínua de números (ex: não há nenhuma nota de 16.4 valores, nem nenhum estudante com 17.2 anos).

Por consequência desta alteração iremos usar **técnicas de classificação** através da previsão de respostas discretas.

Posteriormente agrupamos os valores de cada variável em intervalos que permitissem uma melhor análise (ex: agrupamos os valores das variáveis *G1*, *G2*, *G3* intervalos de 0, de 1 a 9, de 10 a 16 e de 16 a 20). No widget “*Distribution*”, observámos os gráficos de forma a perceber como os grupos de valores variam dentro da mesma variável.

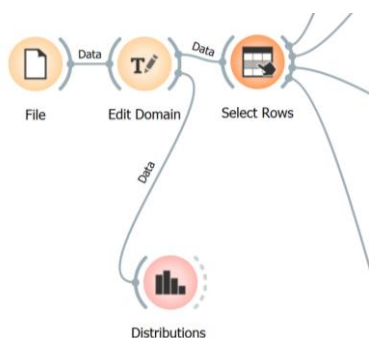


Figura 2- Modelo Orange

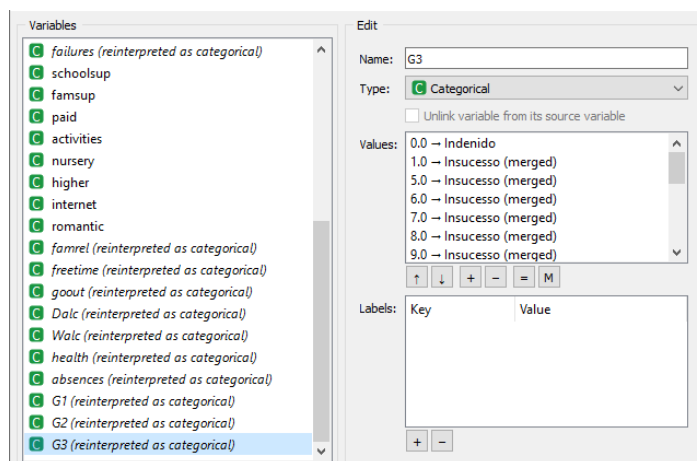


Figura 3- Widget “*Edit Domain*” viewer

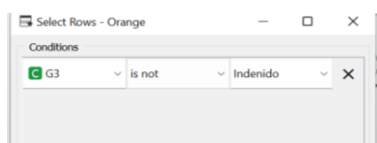


Figura 4- Widget “*Select Rows*”

Através do widget “*Select Rows*” retiramos todas as linhas que na variável *G3* pertencem ao intervalo “indefinido”, por os considerarmos dados ambíguos, que podem nem significar insucesso, sucesso ou excelência.

De seguida começámos por pensar nos nossos objetivos e nas perguntas às quais queremos responder.

Para isso agrupamos os dados no “*Select Columns*”, consoante o tema de cada pergunta e escolhemos o nosso target (aquilo que queremos analisar/prever). E fizemos isso para cada pergunta/objetivo.

Começamos por resolver o seguinte problema: **Será que a família tem influência na decisão do aluno quanto ao ensino superior?**

Para a análise da questão montamos o seguinte modelo e selecionamos a variável *Higher* como o nosso target:

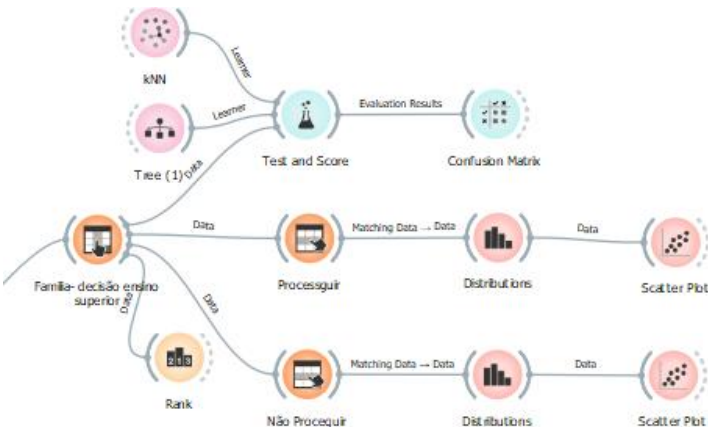


Figura 5- Modelo Orange

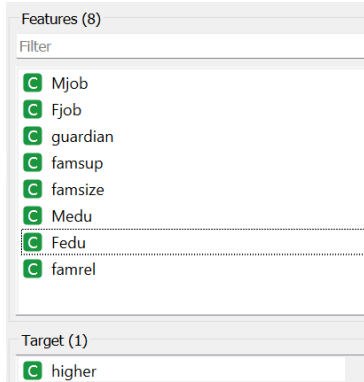


Figura 6- Widget “*Select Columns*”

No “*Select Columns*” selecionamos apenas as variáveis que estão relacionadas com fatores familiares e que nos interessavam para resolver a questão anterior. De seguida, no “*Rank*”, verificamos que a variável *Medu*, *Mjob* e *Fedu* são as que mais se relacionam com o nosso target:









		#	Info. gain	Gain ratio	Gini
1	 Medu	5	0.039	0.019	0.008
2	 Mjob	5	0.034	0.016	0.007
3	 Fedu	5	0.034	0.016	0.007
4	 guardian	3	0.019	0.017	0.007
5	 Fjob	5	0.008	0.005	0.002
6	 famsup	2	0.004	0.004	0.001
7	 famrel	2	0.001	0.001	0.000
8	 famsize	2	0.000	0.000	0.000

Figura 7- Widget “Rank Viewer”

Adicionamos, também, um “*Scatter Plot*” com as variáveis *Medu* e *Fedu* em relação ao target escolhido (*higher*):

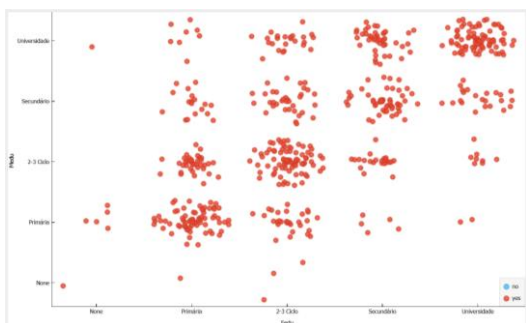


Figura 8 - Widget “Scatter Plot Viewer” - Yes

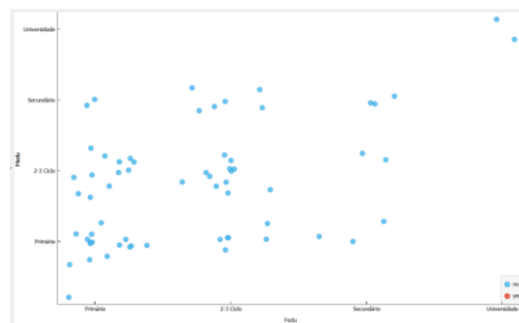


Figura 9 - Widget “*Scatter Plot Viewer*” - No

Podemos, então, concluir que os alunos em que o nível de escolaridade dos pais é o ensino superior, tem a ambição de prosseguir os estudos (apenas dois contradizem esta afirmação). Por conseguinte, os alunos que não querem seguir para o ensino superior, são os mesmos em que os pais também não têm esse nível de ensino.

Da mesma maneira, no Widget do “*Distribution*”, também podemos tirar essa conclusão:

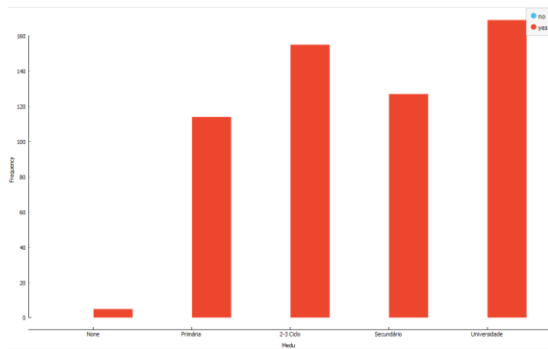


Figura 10- Distribution Viewer - Yes

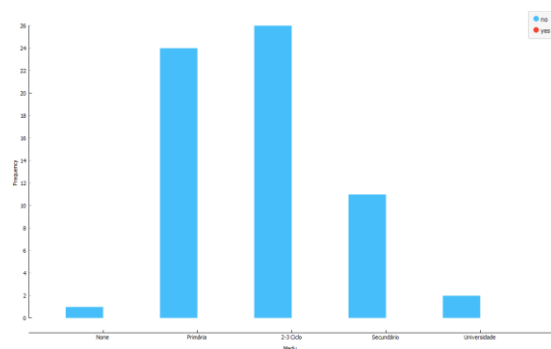


Figura 11 - Distribution Viewer - No

Vamos então proceder à análise do segundo problema:

-Será que a gestão do tempo dos alunos influencia o sucesso dos mesmos?

Para a análise desta questão montamos o seguinte workflow e seleccionamos a variável G3 como o nosso target:

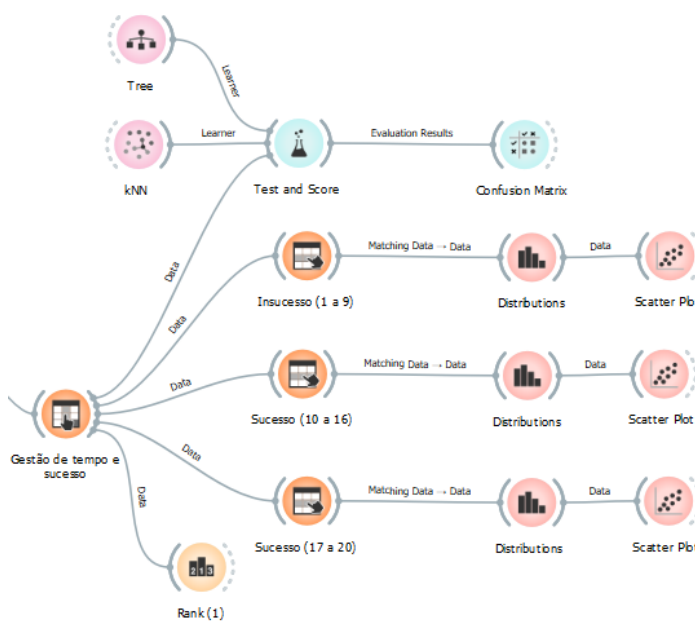


Figura 12- Modelo Orange

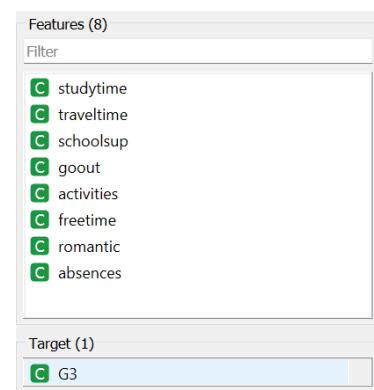


Figura 13- Widget “*Select Columns*”

No “*Select Columns*” seleccionamos apenas as variáveis que estão relacionadas com fatores que influenciam a gestão do tempo.

Para uma melhor compreensão e observação dos dados selecionamos separadamente cada um dos valores do nosso target (G3).

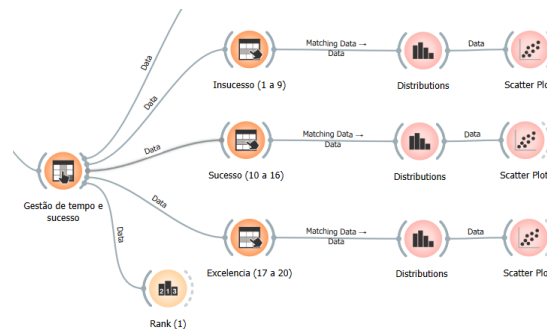


Figura 14- Modelo Orange

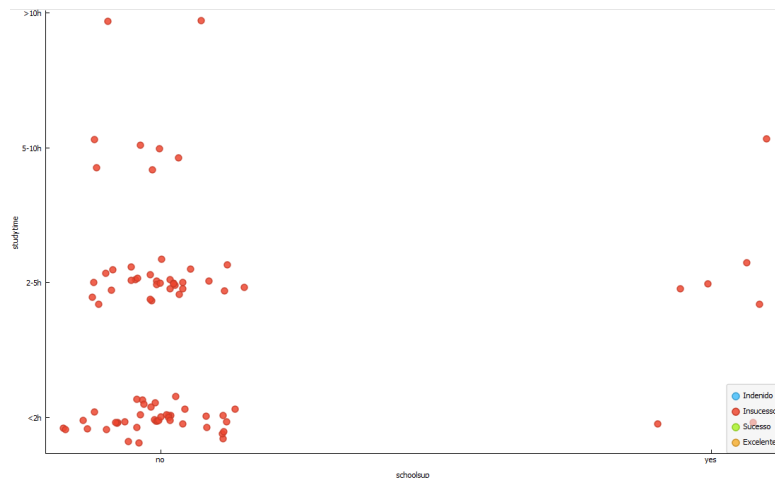
Após visualizarmos vários gráficos nos widgets “*Distributions*” e “*Scatter Plot*” verificamos que a gestão de tempo não tem uma relevância significativa no sucesso ou insucesso dos alunos. Para isso vamos analisar um exemplo que justifica esta afirmação.

Começamos por usar o widget “*Rank*” e verificamos que as variáveis “*Schoolsup*” e “*Studytime*” são aquelas que mais se relacionam com o nosso target.

		#	Gain ratio	Gini
1	schoolsup	2	0.024	0.003
2	studytime	4	0.017	0.005
3	absences	3	0.014	0.004
4	freetime	3	0.007	0.003
5	goout	3	0.006	0.002
6	traveltime	4	0.005	0.002
7	activities	2	0.003	0.001
8	romantic	2	0.003	0.001

Figura 15 - Widget “*Rank Viewer*”

Adicionamos, então, um “*Scatter Plot*”, colocamos estas variáveis nos eixos e procedemos à seguinte análise:



Os alunos que se inserem na categoria “insucesso”, na sua grande maioria, não estudam mais de 5 horas por semana e nem têm apoio escolar.

Figura 16 - Widget “*Scatter Plot Viewer*” - Insucesso

Os alunos que se inserem na categoria “sucesso”, apresentam valores nas variáveis em estudo bastante diferentes.

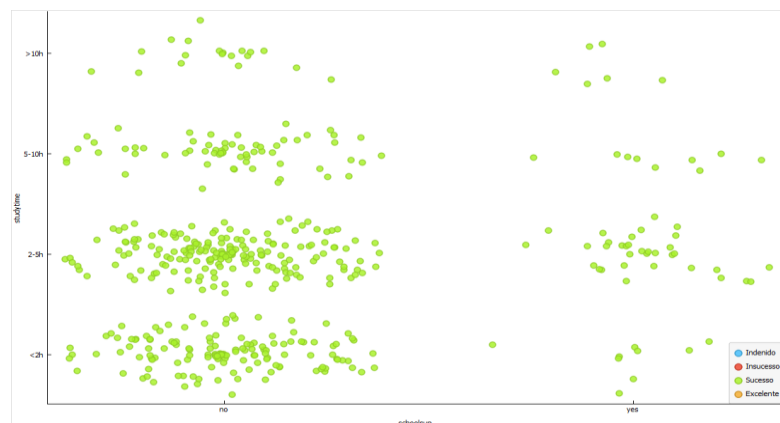
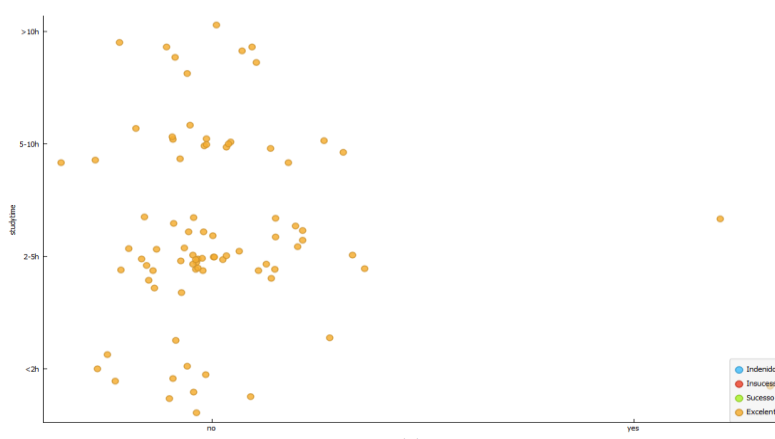


Figura 17 - Widget “*Scatter Plot Viewer*” - Sucesso



Os alunos que se inserem na categoria “excelente”, em comparação à categoria “insucesso” estudam mais horas por semana e apenas um deles tem apoio.

Figura 17 - Widget “*Scatter Plot Viewer*” - Excelente

Com estes gráficos e com a ajuda dos gráficos de barras visualizados no Widget “*Distribution*” observamos que a maioria dos alunos estuda menos de 5h por semana, pelo que concluímos que o sucesso e o insucesso não dependem das horas de estudo. Da mesma forma que a maioria dos alunos também não frequenta o apoio pelo que este também é um indicador do sucesso/insucesso.

Análise dos modelos de Machine Learning

Neste passo, o nosso objetivo é verificar qual é o conjunto de variáveis que melhor satisfazem os nossos problemas. Para isto iremos avaliar 2 modelos de classificação utilizando as ferramentas do Orange.

Para começar, inserimos a ferramenta “*Test and Score*” e fizemos a ligação com os dados pré-processados. Na saída do “*Test and Score*” inserimos os modelos de Machine Learning que serão avaliados:

Tree: É um algoritmo simples que divide os dados em nós por pureza de classe. Um nó interior na árvore representa um atributo que foi testado com os restantes e foi o vencedor como o atributo mais preditivo a este nível. Cada ramo da árvore representa um valor distinto do atributo. Com a árvore construída, define-se um trajeto que termina num nó final que determina o resultado da previsão

KNN: tenta classificar cada amostra de um conjunto de dados e avalia a sua distância em relação aos “parâmetros” mais próximos. Se os parâmetros mais próximos pertencerem maioritariamente a uma classe, a amostra em questão será classificada nessa categoria.

Desta forma, iremos avaliar qual dos modelos é que tem um melhor desempenho para atender os problema que temos desenvolvido ao longo do trabalho.

Para a avaliação dos modelos utilizamos a “*Cross Validation*” (avaliação cruzada) e definimos que 80% dos nossos dados são de teste.

Relativamente à primeira questão tivemos uma precisão de 0.911, e concluímos que o modelo kNN faz uma boa previsão do *Higher* relativamente às variáveis inseridas.

No widget “*Confusion Matrix*” verificamos que este modelo classificou 63 “no” como “yes”, por outro lado acertou em todos os “yes”. Daí podemos concluir que o facto de haver muitas amostras como “yes” o modelo passou a assumir os “no” dados como “yes”.

Model	AUC	CA	F1	Precision	Recall	Predicted			Σ
						no	yes		
						no	yes		
kNN	0.556	0.901	0.855	0.911	0.901	1	63		64
Tree	0.492	0.830	0.830	0.830	0.830	0	570		570
Σ						1	633		634

Figura 10- “Test and Score” viewer

Figura 11- “Confusion Matrix” viewer kNN metod

Referente à segunda questão tivemos uma precisão um pouco mais baixa, 0.608, e daí podemos concluir que este modelo não garante uma boa previsão do nosso target (G3). Na “*Confusion Matrix*” sabemos que o modelo kNN classificou 82 “insucessos” como “sucessos” e classificou 77 “excelentes” como “sucessos”. O que tal como na questão passada, podemos concluir que como o facto de haver muitas amostras como “sucessos” o modelo passou a assumir os “excelentes” dados como “sucesso”.

Model	AUC	CA	F1	Precision	Recall
kNN	0.526	0.705	0.623	0.608	0.705
Tree	0.512	0.585	0.579	0.578	0.585

Figura 12- “Test and Score” viewer

		Predicted				
		Indenido	Insucesso	Sucesso	Excelente	Σ
Actual	Indenido	0	0	0	0	0
	Insucesso	0	3	82	0	85
	Sucesso	0	22	442	3	467
	Excelente	0	3	77	2	82
Σ		0	28	601	5	634

Figura 13- “Confusion Matrix” viewer kNN metod

Conclusão

Neste projeto carregamos, visualizamos, tratamos e exploramos os dados visualmente utilizando “*scatter-plots*” e “*distributions*” e avaliamos e comparamos o desempenho de cada um dos modelos.

Todo este procedimento foi feito utilizando o Software Orange que nos facilitou muito no processo de análise de dados.

Atendendo aos resultados obtidos, podemos agora responder às questões acima colocadas.

Concluimos que a família têm uma forte influência na decisão dos alunos relativamente ao ensino superior, por outro lado quando abordamos a questão da gestão do tempo no sucesso e insucesso dos alunos não notamos uma influência significativa.

Na análise dos modelos de Machine Learning tivemos alguns problemas, uma vez que nos dados que nos foram fornecidos existe um grande número de amostras de variáveis para um valor em específico. Desta forma, os modelos enganam-se e assumem o valor presente em maior quantidade no estudo.

A realização deste projeto permitiu aos alunos o desenvolvimento de competências comunicativas e uma evolução no processo de recolha, veracidade e análise de dados.

Webgrafia

- *Orange: Documentation - Widgets*
<https://orangedatamining.com/widget-catalog/>
- Melim, A. & Veiga, F. (2007). Organização dos tempos de estudo em jovens alunos.

Recuperado de

<https://repositorio.ul.pt/bitstream/10451/4878/1/Organiza%C3%A7%C3%A3o%20dos%20tempos%20de%20estudo%20em%20jovens%20adultos.pdf> em 2022, outubro 30