FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Ano Letivo de 2015/2016

Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Unidade Curricular de Extração de Conhecimento de Dados I

RELATÓRIO DO TRABALHO PRÁTICO I

"Problema de Classificação"

Docentes: Professor Doutor João Gama
Professor Doutor Pavel Bradzil
Discentes: Rui Pedro Machado, 201300292
Hélder Filipe Russa, 201508409

1. INTRODUÇÃO

No âmbito da disciplina de Extração de Conhecimento de Dados I, do Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão da Faculdade de Economia da Universidade do Porto 2015/2016, foi desenvolvido este trabalho, denominado "Problemas de Classificação".

1.1 Descrição do trabalho

O objetivo principal é de explorar, selecionar e trabalhar um conjunto de dados previamente disponibilizados, relacionados com a performance financeira de múltiplas organizações, por forma a treinar o melhor algoritmo de classificação possível, e assim permitir prever a falência de uma nova organização.

A estrutura do problema além de como previamente explicado se tratar de um problema de classificação, baseia-se em modelação supervisionada, onde um conjunto de dados, denominados de treino, nos permite treinar um modelo através de um atributo-alvo, denominado na literatura, supervisor externo, que conhece para cada exemplo dado, o valor de saída, sendo que neste nosso problema se trata da falência ou não de uma empresa, dados os valores observados nos indicadores financeiros. (Gama, Carvalho, Facelli, Lorena, & Oliveira, 2012)

Para responder a este desafio serão aplicadas diferentes técnicas estudadas ao longo da disciplina de Extração de Conhecimento de Dados e comparado o desempenho das mesmas, não só ao nível dos algoritmos utilizados mas também de métodos de eliminação de atributos irrelevantes e de análise de dados.

A estrutura deste relatório segue a seguinte organização:

1. Análise dos dados de treino (Capitulo 2)
2. Introdução às tecnologias utilizados (Capitulo 3);
3. Seleção e preparação dos dados a utilizar (Capitulo 4);
4. Experiencias de treino de modelos classificativos (Capitulo 5);
5. Comparação entre os diferentes modelos de classificação utilizados (Capitulo 6);
6. Avaliação dos resultados obtidos (Capitulo 7);
7. Notas finais (Capitulo 8).

2. ANÁLISE PRELIMINAR DE DADOS

2.1 Análise das Variáveis Disponíveis

Analisando em maior detalhe o conjunto de dados disponibilizados para o estudo e treino do modelo de classificação, observou-se que este é constituído por um total de trinta rácios financeiros, denominados atributos de entrada, que juntamente com um atributo alvo, explicam a falência ou não, de um total de 118 empresas distintas.

À exceção do atributo alvo, todos os restantes denotam um tipo de dados qualitativo contínuo, dado que representam quantidades numa escala contínua e ao mesmo tempo racionais, isto porque todos representam um rácio entre diferentes variáveis financeiras.

Este tipo de analise preliminar sobre o conjunto de dados disponibilizado, com sentido critico e de ponderação da importância de atributos para um determinado problema é uma tarefa essencial para garantir a qualidade dos modelos gerados, evitando adicionar ruído aquando do treino dos mesmo. Permite assim auxiliar na tarefa de eliminação manual de atributos irrelevantes, caso seja identificado um indicador que não faça sentido para o problema em causa. (Gama et al., 2012)

Neste problema consideramos que os trinta indicadores deveriam ser tidos em consideração, dado estarem todos relacionados com o desempenho financeiro de uma organização e preferimos recorrer a métodos científicos para eliminar atributos irrelevantes ao invés de eliminação manual.

Apresenta-se de seguida em tom de resumo, a análise preliminar baseada em métodos estatísticos dos indicadores financeiros disponíveis bem como a classificação individual dos mesmos em termos de tipo de dados e características dos mesmos.

2.2 Análise Univariada

2.2.1 Análise do Atributo Alvo (de saída) – *Bankrupcy*

Utilizando a ferramenta SPSS foi possível gerar um gráfico de frequências para perceber o balanceamento das diferentes observações possíveis para este atributo. De seguida podemos observar o resultado obtido, sob a forma de gráfico e de tabela de frequências:

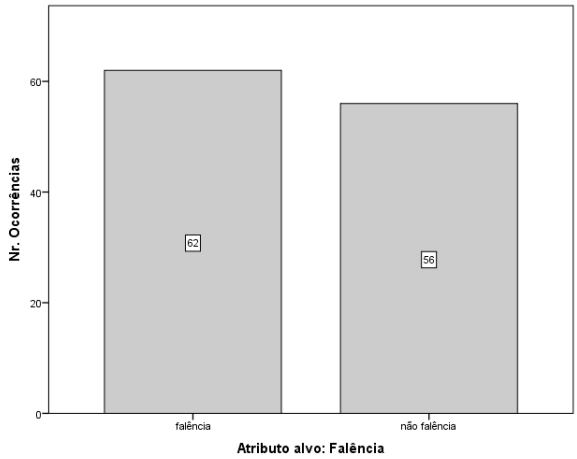

Figura 1 - Gráfico de frequências

	Frequência	Percentagem	Percentagem Cumulativa
Falência (-1)	62	52,5	52,5
Não falência (1)	56	47,5	100,0
Total	118	100,0	

Tabela 1 - Tabela de frequências relativas

Analisando o mesmo, é possível constatar que esta variável apresenta um balanceamento nas observações, dado que entre um total de 118 exemplos, 62 correspondem a uma não falência (”-1”) e 56 correspondem a uma falência (“1”).

Caso este não fosse o cenário e existisse uma clara diferença entre o número de tipos observados, deveria ser utilizada alguma técnica para balancear artificialmente as observações, tal como a técnica de redefinição do tamanho do conjunto de dados, que corresponde a adicionar casos à classe minoritária ou remover casos da maioritária, com o risco de comprometer o modelo de classificação a treinar por *overfitting* ou *underfitting*. (Gama et al., 2012)

2.2.2 Análise dos Atributos Numéricos – Indicadores Financeiros

Utilizando a mesma ferramenta estatística, foi conduzida uma análise descritiva às variáveis numéricas presentes na amostra, como forma de compreender se é necessário realizar um qualquer tipo de limpeza de dados por diferentes motivos, tais como serem inconsistentes, incompletos ou redundantes.

Para a mesma, foram utilizadas as medidas de centralidade, média; mediana e moda, assim como medidas de dispersão tais como o desvio padrão; o coeficiente de dispersão; coeficiente de assimetria e de achatamento. No Anexo I, apresenta-se um quadro resumo dos valores observados para as variáveis da amostra, sendo que no parágrafo seguinte é realizada uma interpretação dos valores mais destacáveis.

Tal como é possível verificar no Anexo referido, a variável X8 e X16 têm exatamente os mesmos valores estatísticos observados, correspondendo ao rácio "*sales/receivables"*, tendo assim uma igual influência no variável alvo, desta forma, optamos por excluir uma delas do treino do nosso modelo de classificação, neste caso a X8.

Outro tipo de análise possível recai sobre a dispersão dos valores observados para uma variável. Tal como é possível observar na tabela, a variável X21 tem valores de desvio padrão muito altos o que indica elevada dispersão, esta que é comprovada pelos respetivos valores dos coeficientes de dispersão. De seguida ilustra-se, com a construção de gráficos *boxplot*, a diferença entre distribuições aproximadamente simétricas e pouco dispersas (X4) e distribuições assimétricas e muito dispersas (X21). Desta forma podemos identificar variáveis que serão pouco explicativas dos valores do atributo alvo (Elevada dispersão) versus variáveis mais explicativas (Baixa Dispersão).

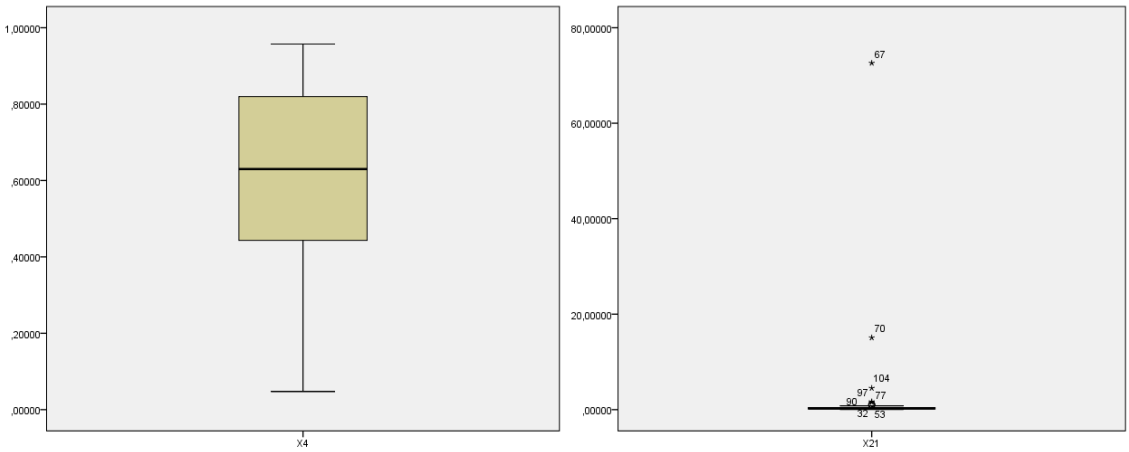

Figura 2 Profilling por Boxplot

Aquando da análise estatística de variáveis existem alguns pontos a ter em atenção, a dispersão dos valores observados para cada variável, que podem tornar a média uma medida não representativa da amostra isto porque é uma medida facilmente influenciada por *outliers*, sendo assim necessário, neste casos, utilizar outro tipo de medidas como a mediana para ter a certeza do comportamento das mesmas. (Gama et al., 2012)

2.3 Análise Multivariada

Para a análise multivariada, foi realizado um teste de correlação, baseado no teste de *Pearson*. Os resultados desta análise pode ser visualizados no Anexo II - Tabela Correlações Multivariadas. Nesta tabela é possível verificar a correlação entre variáveis que assim explica o comportamento entre duas variáveis, nomeadamente o que acontece quando os valores de uma variam positivamente, nos valores de outra e vice-versa. Em Estatística pode assumir-se que um valor superior a 0,75 explicam uma correlação forte no sentido positivo e um valor inferior a -0,75, uma correlação forte no sentido negativo.

Em *Datamining*, este tipo de análise permite identificar atributos redundantes, quando os valores da correlação são precisamente superiores a 0,75 ou inferiores a -0,75, demonstrando alta correlação entre ambos e assim contribuindo de forma similar para a explicação do atributo alvo.

Podemos através desta tabela retirar algumas conclusões, nomeadamente comprovar que dado que a varável X8 tem um conjunto de valores iguais à variável X16, o valor da correlação é 1, dado que sempre que um valor de uma varia positivamente o valor da outra acompanha esta variação.

Assim podemos verificar na tabela que alguns atributos têm forte correlação com outros, por exemplo a variável X6 tem alta correlação com as variáveis X19, X21, X22 e X30, desta forma é irrelevante e pode ser excluída do conjunto de atributos a utilizar na construção dos modelos classificativos.

Seguimos esta abordagem não para excluir atributos irrelevantes, mas para compreender a correlação entre as diferentes variáveis disponíveis e ter uma melhor compreensão sobre as mesmas. Recorremos, no entanto ao WEKA, para através de algoritmos de seleção de atributos escolher os melhores a utilizar na construção do modelo classificativo.

No Anexo III, é possível visualizar o código em R, utilizado para gerar a matriz de correlações.

3. PRÉ-PROCESSAMENTO

3.1 Limpeza de Dados

Após análise dos valores disponíveis em cada um dos indicadores da amostra, foi possível concluir que estes já se encontravam previamente tratados e limpos, como por exemplo sem valores nulos ou casos aberrantes nem dados redundantes. Outro tipo de tarefa que não foi necessário realizar foi qualquer tipo que quantificação de variáveis qualitativas, nomeadamente conversões simbólico-numérico nem o contrario.

3.2 Análise de Relevância de Atributos

Tal como é possível constatar na análise ao enunciado deste problema, o objetivo é treinar um modelo de classificação que consiga prever a falência ou não de uma empresa com base em

30 rácios financeiros, no entanto nem todos estes atributos podem ser úteis e importantes para a previsão do valor do atributo alvo.

Desta forma é importante avaliar a contribuição de cada um dos rácios financeiros para a explicação deste atributo alvo, que neste caso se caracteriza pela falência das organizações. Assim conseguiremos não só eliminar ruido, mas também reduzir tempos de processamento no treino do modelo que no final se espera refletir na melhoria da performance do mesmo.

Existem diferentes técnicas a aplicar para este tipo de análise, nomeadamente, técnicas supervisionadas de ordenação e técnicas de seleção de subconjuntos. Dado que as técnicas de seleção de subconjuntos constituem um processo mais custoso a nível computacional quando comparada com as técnicas de ordenação, muito embora possam trazer resultados mais apurados, decidimos recorrer às primeiras como forma de encontrar uma resposta rápida e leve, para a escolha dos atributos a selecionar para treinar o nosso modelo.

Para este tipo de técnica, que visa simplesmente listar os diferentes atributos, ordenando-os de acordo com a sua importância na explicação do atributo alvo, existem diferentes critérios aplicáveis, nomeadamente:

- ***Information Gain****,* que calcula de forma independente o valor da explicação de cada uma das diferentes variáveis explicativas face ao atributo alvo;
    - No entanto para valores numéricos com elevada dispersão não é a melhor escolha dado que para estes casos, este critério pode sobrestimar o valor de uma variável e não permitir treinar modelos que capturem o padrão de observações.
- ***Info Gain Ratio*** que tem exatamente o mesmo objetivo do *Information Gain,* no entanto resolve o problema de sobrestimar atributos, tendo em consideração, também, o valor da informação (Dada pela sua variabilidade).
- ***Chi-squared*** que testa a independência entre dois eventos, neste caso avalia as distribuições seguidas entre o atributo alvo e o atributo explicativo e avalia a sua interdependência.

No Anexo IV – Tabela de seleção de atributos, gerada através do WEKA, podemos visualizar uma tabela resumo que ilustra os resultados obtidos, após aplicações dos três critérios, na plataforma WEKA.

Podemos ver que existe uma ordenação dos atributos consistente entre os diferentes critérios usados, desta forma, decidimos usar como *input* original do nosso modelo, um valor de referência de 30% do total das variáveis, correspondente a 9 atributos explicativos (X9; X10; X11; X13; X14; X15; X24; X29), marcados a verde na tabela. Desta forma concluímos que o rácio financeiro X9, correspondente ao indicador ROA ("*return on assets*") é a mais explicativa relativamente à falência ou não de uma empresa.

4. CONSTRUÇÃO DO MODELO

4.1 Modelos de Classificação

A escolha de um modelo deve sempre contemplar a utilidade prática para o âmbito da análise, o que neste caso representa (por exemplo), como base em atributos financeiros, prever se uma dada empresa pode falir ou não.

Sendo um dos algoritmos de classificação mais utilizador em todo dada a sua fácil interpretação e simplicidade de compreensão as árvores de decisão foram a primeira escolha para a elaboração de uma resposta a este problema.

Este algoritmo baseia-se na construção de uma árvore onde cada nó contém uma avaliação para um dado atributo e cada folha está associada a uma classe, é por esta razão que é tipicamente denominado um algoritmo de procura.

Como segunda experiência decidimos utilizar o classificador baseado em probabilidades *Naive Bayes* que parte do principio que os atributos de um exemplo são independentes entre si, dado o atributo alvo. Tal como foi dito este algoritmo calcula as probabilidades de observação de um valor X sabendo que já observamos um valor Y de uma outra variável. Este algoritmo é bastante eficiente e fácil de implementação tal como as árvores de decisão.

De seguida são demonstrados os resultados finais obtidos em cada uma das experiências realizadas utilizando os algoritmos previamente identificados. De realçar que a lista de atributos

utilizados baseia-se na lista selecionada através do WEKA, cujo processo foi explicado no ponto anterior.

4.2 Experiências Realizadas

Para a realização das seguintes experiencias foi utilizado o *software Knime* dada a sua simplicidade de utilização baseada em *workflows* montados visualmente.

4.2.1 Experiencia 1 – Árvores de decisão

A árvore de decisão inicial foi construída com todos os parâmetros do Knime como *default*, incluindo o critério, neste caso, *info gain ratio*. Apresenta-se na matriz de confusão da Tabela 4, os resultados obtidos, na melhor execução do nosso modelo classificativo, com um particionamento dos dados de treino, em 80% para treino e 20% para avaliação do desempenho:

			Classificação	
			TP	**FN**
			-1	**1**
Classe	**-1**	*Falência*	12	1
	1	*Não Falência*	6	5
			FP	**TN**

Tabela 2 - Resultados Árvore de Decisão

Para a análise do desempenho do modelo, são assim contruídas as seguintes métricas, baseadas nos resultados da matriz de confusão acima demonstrada, resumidas na Tabela 5:

- *Precision* - Número de instâncias corretamente classificadas numa classe, a dividir pelas instâncias previstas pelo classificador para essa classe *(TP/(TP+FP));*
- *Recall* – Número de instâncias corretamente classificadas numa dada classe, a dividir pelo número de exemplos reais dessa classe *(TP/(TP+FN)) ;*
- *TP Rate* – Taxa de instâncias corretamente classificadas de uma determinada classe *(TP/(TP+FN));*
- *FP Rate* – Taxa de instâncias incorretamente classificadas de uma determinada classe *(FP/(FP+TN));*

			Precision	*Recall*	*TP Rate*	*FP Rate*
Classe	-1	*Falência*	0.67	0.92	0.92	0.08
	1	*Não Falência*	0.83	0.45	0.45	0.55
		weighted mean	0.75	0.6865385		

Tabela 3 - Métricas de Performance do Modelo

Através da análise da tabela de performance, podemos constatar que o nosso modelo, é capaz de classificar assertivamente os casos onde o atributo alvo tem o valor -1, referente a falência, mas nos casos contrários, de não falência, apenas consegue prever corretamente 45% os casos, sendo este o nosso gargalo de desempenho.

A taxa de acerto global do modelo é de 70,87% (Total de casos corretos / Total de casos de testes). Na figura seguinte, é possível visualizar o resultado final, da construção do processo de classificação, no Knime.

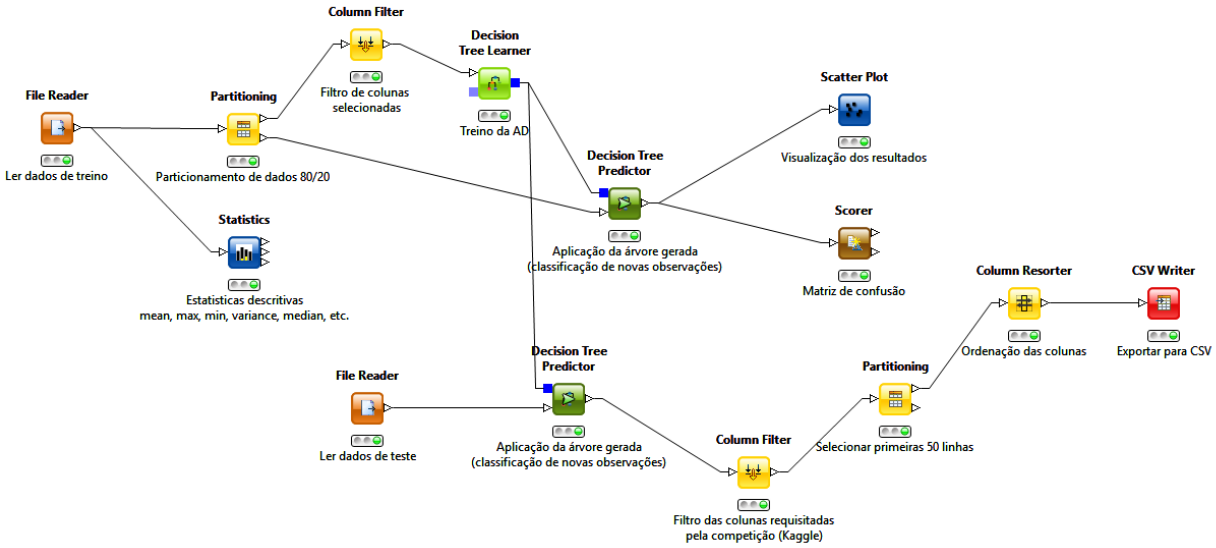

Figura 3 - Treino da Árvore de Decisão no Knime

4.2.2 Experiencia 2 – *Naive Bayes*

O modelo baseado em Naive Bayes inicial foi construído com todos os parâmetros do Knime como *default*. Apresenta-se na matriz de confusão da Tabela 6, os resultados obtidos, na melhor execução do nosso modelo classificativo, com um particionamento dos dados de treino, em 80% para treino e 20% para avaliação do desempenho:

			Classificação	
			TP	**FN**
			-1	**1**
Classe	**-1**	*Falência*	10	3
	1	*Não Falência*	3	8
			FP	**TN**

Tabela 4 - Resultados *Naive Bayes*

Para a análise do desempenho do modelo, são assim utilizadas as mesmas métricas, do classificador anterior, e resumidas na Tabela 7:

			Precision	*Recall*	*TP Rate*	*FP Rate*
Classe	-1	*Falência*	0.77	0.77	0.77	0.23
	1	*Não Falência*	0.73	0.45	0.73	0.27
		weighted mean	0.75	0.6096154		

Tabela 5 - Métricas de Performance do Modelo Naive Bayes

Através da análise da tabela de performance, podemos constatar que os valores obtidos melhoraram substancialmente com o algoritmo do Naive Bayes, quando comparados com a utilização da árvore de decisão.

A taxa de acerto global do modelo é de 75% (Total de casos corretos / Total de casos de testes)

Na figura seguinte, é possível visualizar o resultado final, da construção do processo de classificação, no Knime.

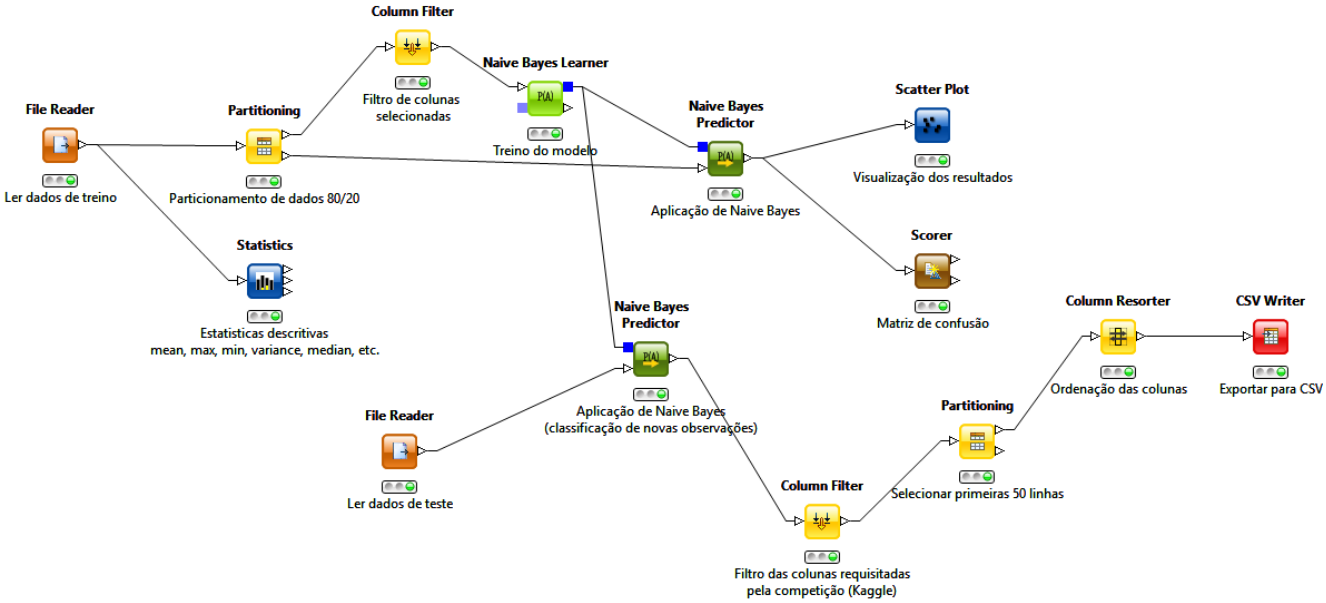

Figura 4 - Treino do Naive Bayes no Knime

5. AVALIAÇÃO DOS MODELOS DE CLASSIFICAÇÃO UTILIZADOS

5.1 Análise Curva ROC

Para escolher o melhor classificador, as curvas ROC analisam graficamente a razão entre os verdadeiros positivos (TP - Eixo Y) e os falsos positivos (FP - Eixo X). Através destas curvas podemos encontrar a área por baixo da curva, *Area Under ROC*, que é a base de comparação entre classificadores, sendo que o objetivo passa por maximizar essa área em cada um deles.

Para este teste, optamos por recorrer ao *Software* WEKA, que de uma forma simples nos permite obter os valores necessários à construção da curva ROC para os classificadores baseados em Árvores de Decisão e Naive Bayes.

Dado que os modelos foram construídos no KNIME, por forma a realizarmos a análise ROC no WEKA, foi necessário recorrer ao algoritmo J48, correspondente ao classificador de árvores de decisão bem como ao algoritmo NaiveBayes para o segundo classificador baseado em redes Bayesianas.

Para a construção da nossa curva ROC, recorremos ao método de cross-validation, com 10 folds e mantivemos a opção de seleção de atributos, já que esta tinha sido criada através do WEKA. O diagrama final pode ser visto na seguinte figura:

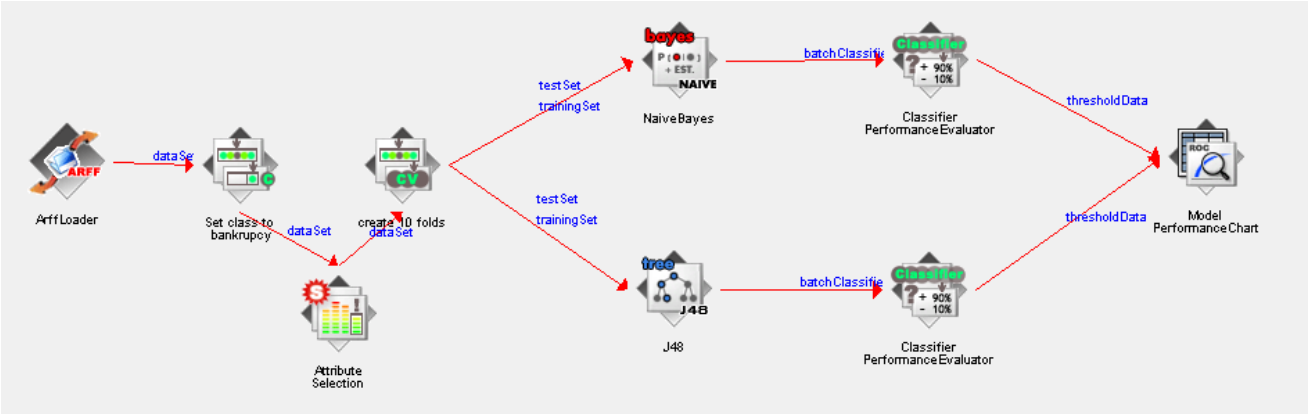

Figura 5 - Construção de curvas ROC em WEKA

Na figura 6, podemos assim visualizar os resultados obtidos pela combinação das curvas ROC de ambos os classificadores, tendo como base o valor -1 (Faliu) como referência, isto porque decidimos que este valor seria o nosso positivo.

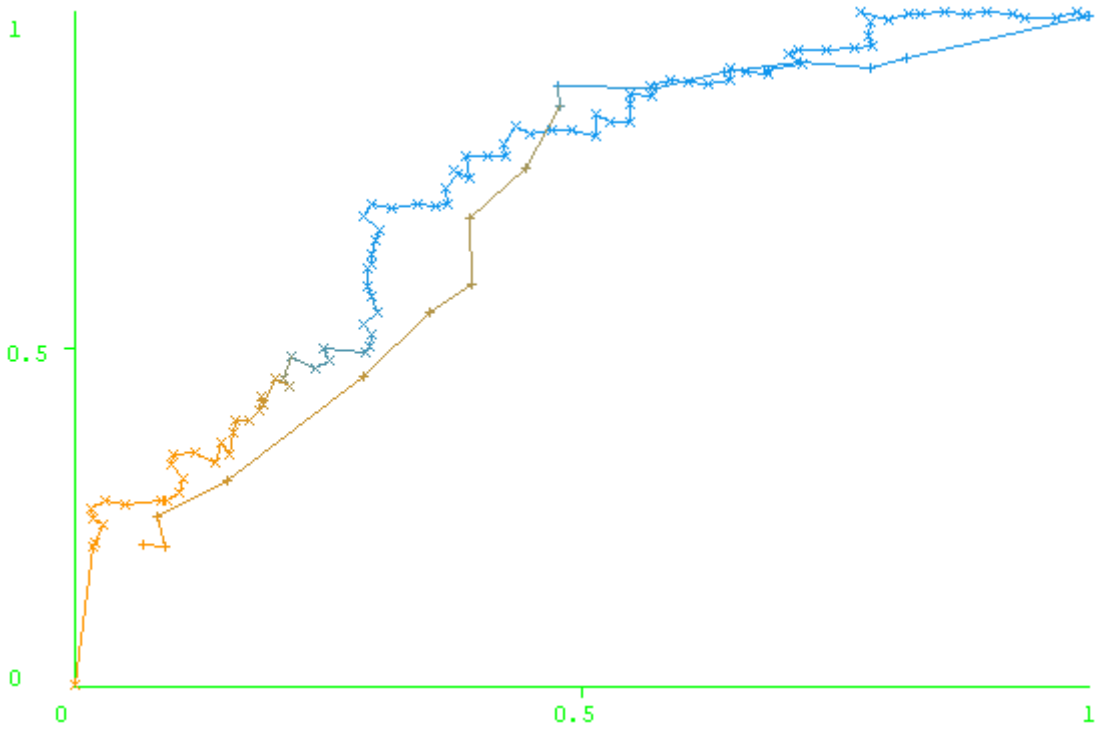

Figura 6 - Curvas ROC geradas

Legenda:

- x - NaiveBayes
- + - J48 (Árvore de Decisão)

Através da análise das curvas ROC, constatamos que a prever valores de falência (-1), o classificador NaiveBayes é sensivelmente melhor para este problema e dada a nossa amostra, visível no gráfico, por ter valores quase sempre acima da árvore de decisão, gerando uma AUC maior, tal como pretendido.

6. CONCLUSÕES

Em termos práticos, acerca dos resultados obtidos, para este problema e dada a abordagem seguida pelo grupo, o algoritmo do Naive Bayes apresentou uma performance relativamente superior (cerca de 5 pontos percentuais), conseguindo uma accuracy geral de 75%, após ter sido realizada uma análise prévia das variáveis bem como uma seleção de melhores atributos a utilizar na construção do modelo.

Podemos assim concluir e aceitar que os objetivos gerais deste trabalho foram atingidos, tendo o grupo realizado as diferentes componentes aprendidas nas aulas, nomeadamente: análise univariada; análise multivariada; limpeza e pré-processamento de dados; seleção de atributos manualmente e cientificamente; treino de modelos com diferentes classificadores e por último, análise de desempenho comparativa entre classificadores.

Acreditamos existirem mecanismos que permitem aumentar ainda mais a performance dos classificadores, no entanto achamos que a realização deste exercício contribuiu para uma melhor compreensão acerca dos passos a seguir na construção de modelos de datamining baseados em tarefas de classificação bem como para compreender como aplicar métodos estatísticos e que decisão tomar com os mesmos para melhor a qualidade dos nossos modelos.

7. BIBLIOGRAFIA

Gama, J., Carvalho, A., Facelli, K., Lorena, A., & Oliveira, M. (2012). *Extração de Conhecimento de Dados*. (E. Silabo, Ed.) (1st ed.). Lisboa: Silabo, Edições.

8. ANEXOS

8.1 Anexo I – Tabela de Análise Univariada

Variável	Amplitude	Mínimo	Máximo	Média	Mediana	Desvio Padrão	Coef. Assimetria	Coef. Achatamento	Coef. Variação
X1	1,81392	0,00018	1,81410	0,2129238	0,0612065	0,38969936	2,609	6,103	1,83022941
X2	0,42512	0,00023	0,42535	0,0620921	0,0272560	0,08723088	2,254	4,884	1,404863279
X3	6,62096	0,18504	6,80600	1,4962928	1,1231500	1,18045536	2,505	6,912	0,788920031
X4	0,90965	0,04764	0,95729	0,6070486	0,6299450	0,23307004	-0,393	-0,605	0,383939645
X5	1,69887	-1,01560	0,68327	0,0286972	0,0330560	0,31212534	-0,612	1,440	10,87651747
X6	85,55894	-85,01000	0,54894	-0,7625846	0,0096980	7,83845362	-10,795	116,980	-10,27879852
X7	29053,66710	1,33290	29055,00000	333,8258746	12,0170000	2680,51049616	10,698	115,508	8,029666662
X8	551,01560	0,19440	551,21000	14,9413924	6,5871000	53,90285259	9,018	86,832	3,607619106
X9	1,08792	-0,59345	0,49447	0,0265338	0,0183760	0,13857875	-0,669	4,979	5,222721371
X10	2,03591	-1,40780	0,62811	0,0330800	0,0260660	0,25684160	-1,630	8,406	7,764253229
X11	25,24780	-17,34000	7,90780	0,0015783	0,0540820	1,85982156	-6,374	67,969	1178,351138
X12	26,32270	-17,34000	8,98270	0,1162044	0,1223600	1,94493001	-5,421	59,353	16,73714666
X13	2,72067	-0,46197	2,25870	0,1371308	0,0308780	0,38832481	2,901	11,576	2,831784597

X14	35,82420	-33,18700	2,63720	-0,2691161	0,0872685	3,13779024	-10,056	105,939	-11,65961696
X15	7,18650	-4,54930	2,63720	0,0077629	0,0541495	0,65952472	-2,950	21,637	84,95810138
X16	551,01560	0,19440	551,21000	14,9413924	6,5871000	53,90285259	9,018	86,832	3,607619106
X17	**13,49456**	**0,00444**	**13,49900**	**2,6335418**	**1,9615500**	**2,35219886**	**2,632**	**8,336**	**0,893169355**
X18	26,49181	0,09319	26,58500	4,3964541	3,5872500	3,65351032	3,077	13,016	0,831012962
X19	1876,93782	0,66218	1877,60000	81,4705998	55,4115000	173,75827333	9,629	99,687	2,132772727
X20	13,49456	0,00444	13,49900	2,6335418	1,9615500	2,35219886	2,632	8,336	0,893169355
X21	72,56914	0,03186	72,60100	1,0936982	0,2783100	6,79216630	10,221	107,608	6,210274523
X22	9,13207	0,02433	9,15640	0,3872689	0,2331750	0,94648730	7,986	68,765	2,444005459
X23	2,94516	0,01194	2,95710	0,7477759	0,6052400	0,57866821	1,737	3,666	0,773852469
X24	0,97739	-0,40568	0,57171	0,0057590	0,0096898	0,09644413	0,264	13,684	16,74658782
X25	1,84060	0,05810	1,89870	0,6210687	0,5926350	0,32351006	0,830	2,008	0,520892514
X26	1410,70500	-26,80500	1383,90000	16,7960549	1,5494000	127,38633513	10,746	116,267	7,584300966
X27	773,49700	-17,73700	755,76000	6,7998346	0,0000000	69,59254731	10,836	117,612	10,234447
X28	637,20800	-9,06800	628,14000	9,9962207	1,4828500	58,27589857	10,386	110,824	5,829793104
X29	1,15510	-0,37934	0,77576	0,0785586	0,0601185	0,16050138	0,979	4,199	2,043077763
X30	10,69339	0,03762	10,73100	0,4302768	0,2787700	0,98402188	9,983	104,968	2,286950786

8.2 Anexo II – Tabela Correlações Multivariadas

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
X1		0,84	0,70	0,10	0,55	-0,02	-0,03	-0,01	0,37	0,31	0,18	0,21	0,64	0,08	0,14	-0,01	-0,06	-0,11	0,01	-0,06	0,02	-0,05	0,46	0,30	-0,46	-0,06	-0,05	-0,08	0,35	0,06
X2	0,84		0,44	0,25	0,37	0,05	-0,02	0,27	0,33	0,28	-0,08	-0,04	0,50	0,10	0,18	0,27	0,20	0,07	-0,09	0,20	-0,06	-0,13	0,28	0,24	-0,21	-0,07	-0,06	-0,08	0,33	-0,06
X3	0,70	0,44		0,12	0,70	0,08	-0,08	-0,10	0,34	0,29	0,24	0,26	0,58	0,03	0,13	-0,10	-0,13	-0,21	-0,05	-0,13	-0,09	-0,18	0,76	0,27	-0,64	-0,03	-0,01	-0,04	0,29	-0,02
X4	0,10	0,25	0,12		0,32	0,25	-0,14	0,09	0,16	0,11	-0,01	-0,01	0,14	0,00	0,07	0,09	0,40	-0,04	-0,22	0,40	-0,28	-0,31	0,12	0,10	0,28	0,03	0,00	0,07	0,16	-0,23
X5	0,55	0,37	0,70	0,32		0,15	-0,10	-0,38	0,54	0,46	0,39	0,40	0,54	-0,02	0,29	-0,38	-0,13	-0,23	-0,10	-0,13	-0,15	-0,24	0,59	0,32	-0,67	0,05	0,05	0,04	0,50	-0,10
X6	-0,02	0,05	0,08	0,25	0,15		0,01	0,03	0,03	0,01	0,00	0,00	0,05	-0,01	0,01	0,03	0,11	0,11	-0,97	0,11	-0,99	-0,89	-0,02	-0,52	0,02	0,01	0,00	0,02	0,05	-0,97
X7	-0,03	-0,02	-0,08	-0,14	-0,10	0,01		-0,01	0,03	0,08	0,02	0,02	-0,01	0,02	0,03	-0,01	0,01	0,16	-0,03	0,01	-0,01	-0,02	-0,05	0,02	0,03	-0,01	-0,01	-0,02	0,06	-0,04
X8	-0,01	0,27	-0,10	0,09	-0,38	0,03	-0,01		-0,18	-0,11	-0,73	-0,69	-0,04	0,01	-0,05	1,00	0,55	0,45	-0,09	0,55	-0,04	-0,06	-0,19	-0,02	0,43	-0,03	-0,02	-0,04	-0,19	-0,07
X9	0,37	0,33	0,34	0,16	0,54	0,03	0,03	-0,18		0,93	0,54	0,58	0,76	0,16	0,55	-0,18	0,15	0,11	-0,04	0,15	-0,04	-0,11	0,32	0,60	-0,46	-0,04	-0,03	-0,04	0,96	-0,04
X10	0,31	0,28	0,29	0,11	0,46	0,01	0,08	-0,11	0,93		0,48	0,52	0,70	0,15	0,50	-0,11	0,14	0,18	-0,04	0,14	-0,04	-0,15	0,29	0,70	-0,42	-0,02	-0,02	-0,02	0,86	-0,03
X11	0,18	-0,08	0,24	-0,01	0,39	0,00	0,02	-0,73	0,54	0,48		1,00	0,44	0,05	0,22	-0,73	-0,22	-0,15	0,01	-0,22	-0,01	-0,03	0,13	0,30	-0,38	0,00	0,00	0,00	0,50	0,01
X12	0,21	-0,04	0,26	-0,01	0,40	0,00	0,02	-0,69	0,58	0,52	1,00		0,50	0,06	0,24	-0,69	-0,20	-0,13	0,00	-0,20	-0,01	-0,04	0,13	0,33	-0,40	-0,01	-0,01	-0,01	0,55	0,00
X13	0,64	0,50	0,58	0,14	0,54	0,05	-0,01	-0,04	0,76	0,70	0,44	0,50		0,11	0,34	-0,04	0,09	0,05	-0,07	0,09	-0,06	-0,13	0,39	0,50	-0,47	-0,05	-0,04	-0,06	0,78	-0,05
X14	0,08	0,10	0,03	0,00	-0,02	-0,01	0,02	0,01	0,16	0,15	0,05	0,06	0,11		0,22	0,01	0,10	0,12	-0,08	0,10	-0,01	-0,05	0,09	0,14	-0,15	-0,97	-0,97	-0,96	0,17	-0,07
X15	0,14	0,18	0,13	0,07	0,29	0,01	0,03	-0,05	0,55	0,50	0,22	0,24	0,34	0,22		-0,05	0,11	0,09	-0,01	0,11	-0,01	-0,05	0,14	0,38	-0,20	-0,02	-0,01	-0,03	0,51	-0,03
X16	-0,01	0,27	-0,10	0,09	-0,38	0,03	-0,01	1,00	-0,18	-0,11	-0,73	-0,69	-0,04	0,01	-0,05		0,55	0,45	-0,09	0,55	-0,04	-0,06	-0,19	-0,02	0,43	-0,03	-0,02	-0,04	-0,19	-0,07
X17	-0,06	0,20	-0,13	0,40	-0,13	0,11	0,01	0,55	0,15	0,14	-0,22	-0,20	0,09	0,10	0,11	0,55		0,82	-0,23	1,00	-0,14	-0,22	-0,11	0,05	0,39	-0,08	-0,08	-0,07	0,17	-0,23
X18	-0,11	0,07	-0,21	-0,04	-0,23	0,11	0,16	0,45	0,11	0,18	-0,15	-0,13	0,05	0,12	0,09	0,45	0,82		-0,25	0,82	-0,14	-0,22	-0,20	0,02	0,27	-0,09	-0,08	-0,09	0,14	-0,25
X19	0,01	-0,09	-0,05	-0,22	-0,10	-0,97	-0,03	-0,09	-0,04	-0,04	0,01	0,00	-0,07	-0,08	-0,01	-0,09	-0,23	-0,25		-0,23	0,97	0,91	0,11	0,47	-0,04	0,08	0,09	0,08	-0,08	0,98
X20	-0,06	0,20	-0,13	0,40	-0,13	0,11	0,01	0,55	0,15	0,14	-0,22	-0,20	0,09	0,10	0,11	0,55	1,00	0,82	-0,23		-0,14	-0,22	-0,11	0,05	0,39	-0,08	-0,08	-0,07	0,17	-0,23
X21	0,02	-0,06	-0,09	-0,28	-0,15	-0,99	-0,01	-0,04	-0,04	-0,04	-0,01	-0,01	-0,06	-0,01	-0,01	-0,04	-0,14	-0,14	0,97	-0,14		0,94	0,00	0,43	0,00	0,01	0,02	0,00	-0,07	0,97
X22	-0,05	-0,13	-0,18	-0,31	-0,24	-0,89	-0,02	-0,06	-0,11	-0,15	-0,03	-0,04	-0,13	-0,05	-0,05	-0,06	-0,22	-0,22	0,91	-0,22	0,94		-0,10	0,21	0,10	0,03	0,04	0,03	-0,14	0,90

X23	0,46	0,28	0,76	0,12	0,59	-0,02	-0,05	-0,19	0,32	0,29	0,13	0,13	0,39	0,09	0,14	-0,19	-0,11	-0,20	0,11	-0,11	0,00	-0,10		0,32	-0,61	-0,08	-0,07	-0,09	0,29	0,05
X24	0,30	0,24	0,27	0,10	0,32	-0,52	0,02	-0,02	0,60	0,70	0,30	0,33	0,50	0,14	0,38	-0,02	0,05	0,02	0,47	0,05	0,43	0,21	0,32		-0,32	-0,03	-0,03	-0,04	0,53	0,46
X25	-0,46	-0,21	-0,64	0,28	-0,67	0,02	0,03	0,43	-0,46	-0,42	-0,38	-0,40	-0,47	-0,15	-0,20	0,43	0,39	0,27	-0,04	0,39	0,00	0,10	-0,61	-0,32		0,13	0,11	0,16	-0,42	-0,06
X26	-0,06	-0,07	-0,03	0,03	0,05	0,01	-0,01	-0,03	-0,04	-0,02	0,00	-0,01	-0,05	-0,97	-0,02	-0,03	-0,08	-0,09	0,08	-0,08	0,01	0,03	-0,08	-0,03	0,13		1,00	1,00	-0,07	0,08
X27	-0,05	-0,06	-0,01	0,00	0,05	0,00	-0,01	-0,02	-0,03	-0,02	0,00	-0,01	-0,04	-0,97	-0,01	-0,02	-0,08	-0,08	0,09	-0,08	0,02	0,04	-0,07	-0,03	0,11	1,00		0,98	-0,06	0,09
X28	-0,08	-0,08	-0,04	0,07	0,04	0,02	-0,02	-0,04	-0,04	-0,02	0,00	-0,01	-0,06	-0,96	-0,03	-0,04	-0,07	-0,09	0,08	-0,07	0,00	0,03	-0,09	-0,04	0,16	1,00	0,98		-0,07	0,06
X29	0,35	0,33	0,29	0,16	0,50	0,05	0,06	-0,19	0,96	0,86	0,50	0,55	0,78	0,17	0,51	-0,19	0,17	0,14	-0,08	0,17	-0,07	-0,14	0,29	0,53	-0,42	-0,07	-0,06	-0,07		-0,09
X30	0,06	-0,06	-0,02	-0,23	-0,10	-0,97	-0,04	-0,07	-0,04	-0,03	0,01	0,00	-0,05	-0,07	-0,03	-0,07	-0,23	-0,25	0,98	-0,23	0,97	0,90	0,05	0,46	-0,06	0,08	0,09	0,06	-0,09	

8.3 Anexo III – Código em R para gerar tabela de correlação multivariada

```
set.seed(7)
# load the library
library(caret)
library(class)
#set working direcotryWork
global.dir <- "C:\\Users\\ruima_000\\iCloudDrive\\FEP - Mestrado\\Extração de Conhecimento de Dados\\Work I"
setwd(global.dir)
#read input
input <- read.csv('training_1.csv', header=T, sep= ',' )
# calculate correlation matrix
correlationMatrix <- cor(input[2:ncol(input)])
# summarize the correlation matrix
write.csv(file = "C:\\Users\\ruima_000\\iCloudDrive\\FEP - Mestrado\\Extração de Conhecimento de Dados\\Work I\\correlation.csv" ,correlationMatrix)
```

8.4 Anexo IV – Tabela de seleção de atributos, gerada através do WEKA

Variável	Chi-Squared	Info Gain	Info Gain Ratio
X8	*Eliminada Manualmente*		
X9	39.0983	0.2542	0.2544
X24	37.055	0.2405	0.2467
X10	36.9076	0.2391	0.2392
X11	35.6786	0.2328	0.246
X13	34.4334	0.2257	0.2296
X29	34.1938	0.2735	0.2226
X15	32.7774	0.2112	0.2112
X14	28.7768	0.1906	0.2029
X12	28.2832	0.1803	0.1821
X1	21.9894	0.1745	0.2934
X21	15.4641	0.097	0.0976
X3	15.2189	0.0954	0.1032
X5	12.0966	0.0974	0.2327
X2	0	0	0
X4	0	0	0
X6	0	0	0
X7	0	0	0
X16	0	0	0
X17	0	0	0
X18	0	0	0
X19	0	0	0
X20	0	0	0
X22	0	0	0
X23	0	0	0
X25	0	0	0
X26	0	0	0
X27	0	0	0
X28	0	0	0
X30	0	0	0