Data Mining

Sistemas de Informação II

Tópicos

Decision Tree;

Clustering;

Neuronal Network;

Assiciation Rules;

Logistic Regression;

Native Bayes.

Tópico 1

Decision Tree

Decision Tree

- Primeiro efetuámos os testes com os parâmetros default (com o valor discrete para tudo o que for numérico).
- SCORE_METHOD: 4.
- SPLIT_METHOD: 3.
- Podemos observar que em modo geral os valores das métricas são mais elevados no teste 2, que foi realizado com a *relationship*, *capitalgain*, *capitalloss*, *educationum*, *occupation*, *hoursperweek*, *age* e *sex*.

Inputs	Teste 1	Teste 2
maritalstatus	discrete	
relationship		discrete
capitalgain	discrete	discrete
capitalloss	discrete	discrete
educationum		discrete
education	discrete	
occupation	discrete	discrete
hoursperweek	discrete	discrete
age	discrete	discrete
sex	discrete	discrete
Score	0.91	0.92
Accuracy	0.8315	0.8402
Precision	0.7714	0.7211
Recall	0.4929	0.6000

Decision Tree

- A razão pela qual dividimos em dois testes é porque tanto a *education* e *educationum* estão relacionados, enquanto que na educação diz, por exemplo, *Bachelors*, conseguimos observar que na parte de *educationum*, este tipo de grau académico corresponde ao seu respetivo número.
- Assim como o campo da *relationship* e *marital-status*, achamos que não fazia sentido testá-los juntos, porque eles no fundo estão um pouco relacionados também. Quando uma pessoa, por exemplo, do sexo masculino tem a *relationship* "husband" no seu marital-status seria "married" e casos assim que conseguimos observar.

Decision Tree

- Primeiro fizemos os testes com os dados todos a default, ou seja, com o valor *discrete*, como vimos anteriormente. O que acaba por não fazer sentido, os dados *discrete* consistem em valores distintos que são separados uns dos outros, por exemplo, o número de alunos numa sala, o número de pessoas numa casa. Quando falamos em dados *discretized*, são dados contínuos, ou seja, que foram divididos em intervalos ou compartimentos. Esse processo é chamado de discretização e é frequentemente usado para simplificar análise de dados contínuos. Por exemplo, se estivermos a medir a altura de um grupo de pessoas, podemos dividir em intervalos como "menos de 1m40", "de 1m40 a 1m70" . Isso vai facilitar a análise de dados para poder comparar o número de pessoas dos diferentes intervalos, em vez de tentar medidas de altura individuais. Por isso optámos por repetir os testes anteriores com os mesmos inputs alterando apenas o tipo de dados, de *discrete* para *discretized*, para tudo o que for numérico.

Decision Tree

- Mudámos os dados numéricos para *discretized*.
- Voltámos a efetuar exatamente com os mesmos inputs que nos davam por defeito, apenas alterámos os valores numéricos para *discretized*.
- Podemos observar que o score do teste 1 melhorou em relação ao teste passado e a *accuracy* e *recall* também.
- No teste 2 o score manteve-se o mesmo e a *precision* e *recall* melhoraram um pouco também.

Inputs	Teste 1	Teste 2
maritalstatus	discrete	
relationship		discrete
capitalgain	discretized	discretized
capitalloss	discretized	discretized
educationum		discretized
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.92	0.92
Accuracy	0.8420	0.8415
Precision	0.7557	0.7603
Recall	0.5338	0.5317

Decision Tree

- Tanto os inputs *capitalgain* e *capitalloss*, considerámos que são parâmetros que não vamos incluir no nosso modelo, porque não nos dão valores reais. Tanto pode ter acontecido este ano para uma pessoa, como para o próximo ano acontecer com outra. Ou seja, é algo não garantido, não é certo.

- Descartámos também os inputs *relationship* e *maritalstatus*, porque não achamos que sejam uma mais valia neste modelo, não nos acrescenta nada. São dados confusos.

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.88	0.87
Accuracy	0.7984	0.7877
Precision	0.6344	0.6526
Recall	0.4366	0.3939

Decision Tree - Testes 1 e 2

Comparando os resultados anteriores:

- *Score*: 0.92 é igual em ambos os testes;
- *Accuracy*: muito similar, ligeiramente maior no teste 1;
- *Precision*: muito similar, ligeiramente maior no teste 2;
- *Recall*: muito similar, ligeiramente maior no teste 1.

Em comparação são ambos muito parecidos!

O objetivo é ver se a e*ducation* é ou não importante para o nosso modelo.

Decision Tree - Conclusões

Teste 1 – Com *education*:

- Em 264 pessoas que têm doutoramento, 201 ganham mais de 50K;
- Em 1126 pessoas que têm mestrado, 635 ganham mais de 50K (pouco mais de metade), sendo que a maioria é do sexo masculino (527) e que por sua vez têm idade superior a 29 anos (515);
- Em contra partida para 3539 pessoas com grau bacharel, 1483 ganham mais de 50K (menos de metade), ou seja, há mais pessoas a ganhar menos 50K do que mais de 50K.

Faz sentido na realidade de hoje em dia?

Decision Tree - Conclusões

Teste 2 – Sem *education*:

- Sem a educação é dado mais foco à idade e à ocupação.
- À medida que a idade aumenta a probabilidade de ganhar mais de 50K aumenta e verificamos também que esse aumento vai estar dependente da ocupação que cada pessoa tem.
- **Exemplo:**
- Para idades entre 49 e 60 anos, há 77 pessoas que têm a ocupação de *tech support*, em que mais de metade (48) ganha mais de 50K, sendo que a maior parte é homem e a maioria destes ganham mais de 50K, o que nas mulheres é o inverso, isto é, há menos mulheres com esta ocupação e a maioria ganha menos de 50K (apenas 3).

Faz sentido na realidade de hoje em dia?

Decision Tree

- Decidimos alterar o parâmetro relativo do "Split method", para tentar tirar conclusões na parte dos valores das métricas.
- Nos algoritmos de árvore de decisão, o "Split method" é usado para fragmentar os dados num determinado nó da árvore em dois ou mais subconjuntos com base em determinadas condições. Esse processo é repetido em cada subconjunto recursivamente até que um determinado critério de paragem seja atendido, como todos os elementos do subconjunto pertencem à mesma classe ou o subconjunto é muito pequeno para ser dividido posteriormente.
- Em geral, o método *split* é uma parte importante do algoritmo da árvore de decisão, pois determina como os dados são fragmentados e, em última análise, como a árvore é construída.

Decision Tree

- SLIPT_METHOD = 1 (***Binary***)

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.88	0.87
Accuracy	0.8032	0.7932
Precision	0.6815	0.6719
Recall	0.3834	0.3504

- SLIPT_METHOD = 2 (***Complete***)

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.87	0.86
Accuracy	0.7939	0.7675
Precision	0.6192	0.5810
Recall	0.4334	0.2767

Decision Tree – Comparar métricas

- Podemos observar que nos testes com a educação (teste 1) o *score* em todos os testes matêm-se. Em relação á *accuracy* e precisão obtivemos melhores resultados quando o *split method* é igual a 1, em contra partida o *recall* é maior quando o *split method* é default (usa ambas, *binary* e *complete*).
- Em relação ao teste sem a educação (teste 2), o *score* apenas diminuiu um pouco quando o *split method* é igual a 2 (*complete*). Tanto a *accuracy* como a precisão, também obtivemos melhores resultados quando o *split method* é igual a 1, e em contra partida observamos que o *recall* é melhor quando o *split method* é default (usa ambas, *binary* e *complete*).

Decision Tree

- Decidimos alterar o parâmetro relativo do "Score method", para tentar tirar conclusões na parte dos valores das métricas.
- O *Score Method* é usado para avaliar a qualidade da árvore. É normalmente usado para avaliar o desempenho da árvore num conjunto de dados de teste, após a árvore ter sido treinada num conjunto de dados de treinamento.
- Diferentes algoritmos de árvore de decisão podem ter diferentes maneiras de calcular o *score*. No caso de árvores de classificação, o *score* geralmente é calculado como a percentagem de amostras de dados de teste classificadas corretamente pela árvore. Para árvores de regressão, o *score* pode ser calculado como o erro quadrático médio entre os valores previstos e os valores verdadeiros para os dados de teste. Em geral, o score é uma importante métrica de avaliação para árvores de decisão, pois fornece uma medida de quão bem a árvore é capaz de generalizar para dados não vistos.

Decision Tree

- SCORE_METHOD= 1 (***Entropy***)

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.88	0.87
Accuracy	0.7995	0.7955
Precision	0.6385	0.6325
Recall	0.4526	0.4500

- SCORE_METHOD= 3 (***Bayesian With K2 Prior***)

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.88	0.87
Accuracy	0.7986	0.7944
Precision	0.6350	0.6398
Recall	0.4369	0.4215

Decision Tree – Comparar Métricas

- Podemos observar que no teste com a educação (teste 1) o score em todos os testes matêm-se (0,88). Em relação à *accuracy*. precisão e *recall* obtivemos melhores resultados quando o *score method* é igual a 1.
- No que toca ao teste feito sem a educação (teste2), o score também se mantém em todos os testes (0,87). Na *accuracy* e *recall* obtivemos melhores resultados quando o *score method* é igualmente igual a 1, mas em contra partida, a precisão é ligeiramente melhor quando o *score method* é igual a 3.

Tópico 2

Clustering

Clustering

- Número de clusters default (=10).

Matriz classificação – com *education*

Contagens para Clustering em Salary:

Previsto	<=50K (Real)	>50K (Real)
<=50K	6699	2089
>50K	111	149

Matriz classificação – sem *education*

Contagens para Clustering1 em Salary:

Previsto	<=50K (Real)	>50K (Real)
<=50K	6768	2280
>50K	0	0

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.84	0.83
Accuracy	0.7569	0.7480
Precision	0.5731	ERRO
Recall	0.0666	0

Não vamos considerar os valores da *precision* e *recall*, do teste 2, na nossa análise.

Clustering

Número de clusters = 5

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.85	0.83
Accuracy	0.7738	0.7508
Precision	0.6321	0.5093
Recall	0.2042	0.2982

Número de clusters = 15

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.84	0.84
Accuracy	0.7536	0.7612
Precision	0.5046	0.5630
Recall	0.2203	0.2350

Clustering – Com educação

- Comparando as métricas dos testes anteriores feitos obtivemos melhores resultados usando **5 clusters** e, de entre os clusters apresentados escolhemos os que consideramos mais relevantes:

Clustering – Com educação

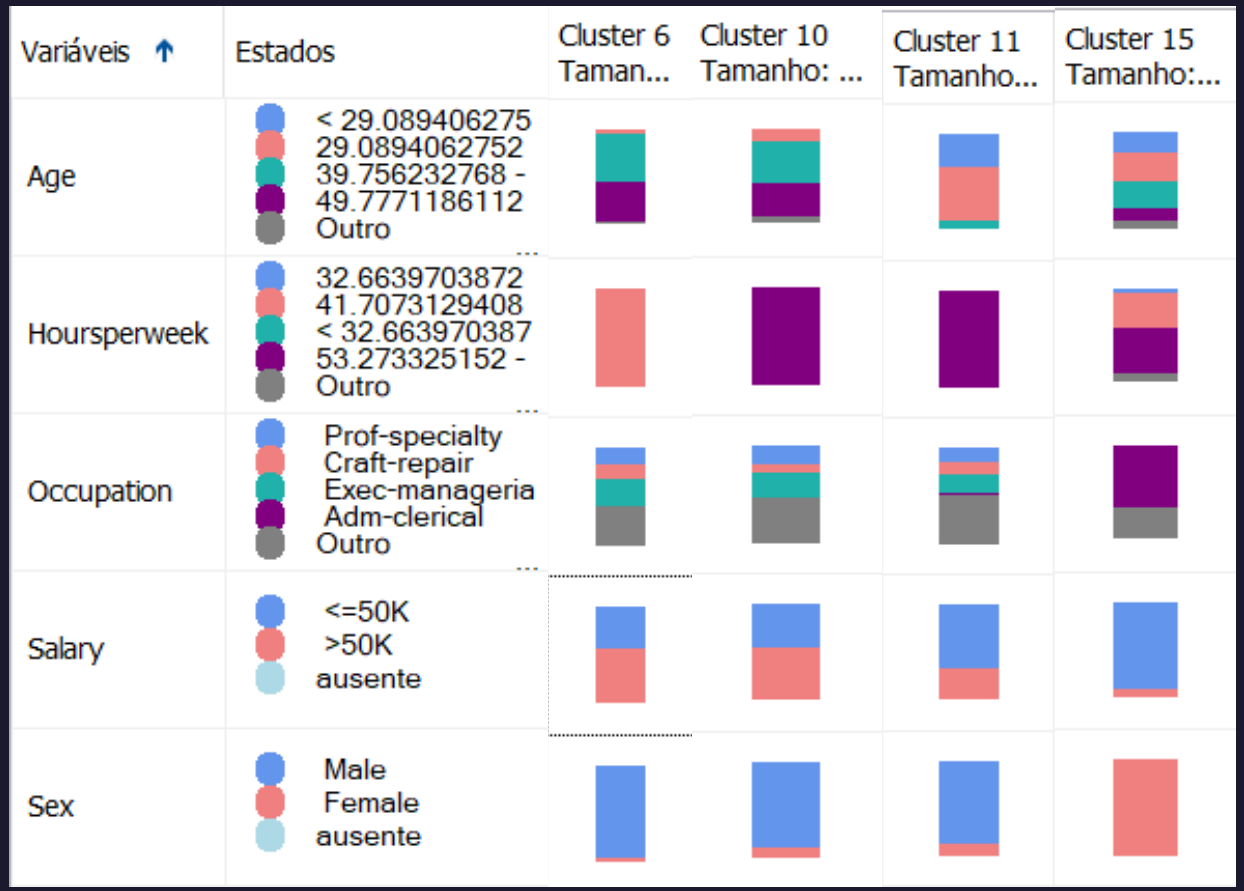
- **Cluster 1:** este cluster indica-nos que a maioria das pessoas ganham menos de 50K (a percentagem de ganhar acima de 50K é quase nula – 0.05), pertencem ao grau de escolaridade "High School" que equivale ao nosso nível de secundário e todas têm menos de 29 anos (100%);

- **Cluster 3:** de todos os clusters escolhemos também este, porque é o segundo que tem maior probabilidade de ganhar mais de 50K. Este cluster apresenta-nos um salário mais ou menos equilibrado, mais para o menos do que para o mais (69% para menos de 50K e 31% para mais de 50K). As horas por semana estão também equilibradas mas onde está mais balanceado é na educação, onde conseguimos ver que as percentagens para os diferentes graus de educação não variam muito, isto é, os valores estão próximos uns dos outros e faz sentido, pois a maioria ganha menos de 50K e tem idade igual a 29 anos, dado este que nos chamou à atenção por ser único;

- **Cluster 5:** este é o cluster com maior probabilidade de ganhos superiores a 50K, aqui a maioria trabalha 41 horas por semana e têm ocupação "Prof-Specialty" e "Exec-managerial", a maioria tem grau Bacharel, pois há mais pessoas com este grau de escolaridade e mais uma vez observamos que esta maioria têm idades entre os 39 e os 49 anos e são do sexo masculino.

Clustering – Sem educação

- Comparando as métricas dos testes anteriores feitos obtivemos melhores resultados usando **15 clusters** e, de entre os clusters apresentados escolhemos os que consideramos mais relevantes:



Clustering – Sem educação

- **Cluster 6:** este cluster é muito semelhante ao cluster anterior (5). Tem maior probabilidade de ganhos superiores a 50K, onde aqui todos trabalham 41 horas por semana e a maioria são do sexo masculino (95%), as ocupações que mais se destacam são as mesmas (“Prof-Specialty” e “Exec-managerial”) e têm idades entre os 39 e 49 anos;
- **Cluster 10:** de todos os clusters escolhemos também este, porque é o segundo que tem maior probabilidade de ganhar mais de 50K. Este cluster apresentam-nos um salário mais ou menos equilibrado, a maioria ganha mais que 50K (53%), em que as pessoas fazem entre 53 a 65 horas por semana, sendo a maioria do sexo masculino. Neste caso as idades estão mais compreendidas entre 39 e os 60 anos;

Clustering – Sem educação

- **Cluster 11:** escolhemos este cluster pois os dados não são bem balanceados. A maioria ganha menos de 50K (60%) com idades compreendidas entre 49 e 60 anos, trabalham entre 32 a 41 horas e são do sexo masculino;
- **Cluster 15:** depois escolhemos este cluster, devido à baixa probabilidade de ganhar mais de 50K (89%). Podemos observar que a totalidade dessas pessoas são do sexo feminino, com idades muito bem distribuídas, onde a maioria trabalha entre as 53 e 65 horas por semana. A ocupação destacada é a "Adm-clerical".

Tópico 3

Neuronal Network

Neuronal Network

- Hidden_node_ratio default (=4.0).
- Na rede neuronal, efectuámos os primeiros testes em default e fomos alterando o parâmetro que corresponde aos nós ocultos (*hidden node*) para perceber se as métricas melhoram ou não. Uma grande proporção de nós ocultos para nós de entrada e saída pode permitir que a rede aprenda padrões mais complexos nos dados, mas também pode aumentar o risco de *overfitting*, onde a rede funciona bem nos dados de treinamento, mas mal em dados novos e não vistos. Por outro lado, uma pequena proporção de nós ocultos pode não ser capaz de capturar a complexidade dos dados, levando a um *underfitting* e baixo desempenho.

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.89	0.88
Accuracy	0.8029	0.7918
Precision	0.6393	0.6359
Recall	0.4665	0.4066

Neuronal Network

- Hidden_node_ratio (=5.0)

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.89	0.88
Accuracy	0.8053	0.7906
Precision	0.6462	0.6268
Recall	0.4700	0.4184

- Hidden_node_ratio (=3.0)

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.89	0.88
Accuracy	0.8064	0.7929
Precision	0.6412	0.6290
Recall	0.4942	0.4351

Neuronal Network – Comparação Métricas

- Relativamente ao teste com educação (teste 1) observamos que o *score* mantêm-se igual (0.89) em todos os testes efetuados. A *accuracy* e o *recall* são ligeiramente melhores quando os nós ocultos são menores (*hidden node* = 3) e a precisão é melhor quando há um maior número de nós ocultos (quando o *hidden node* = 5).
- No que toca aos testes efetuados sem a educação (teste 2), o *score* também se mantém igual (0.88) em todos os teste efetuados. A *accuracy* e o *recall* são também ligeiramente melhores quando os nós ocultos são menores (= 3), a precisão é melhor em default (*hidden node* = 4).

Neuronal Network – O que favorece? Com educação

Conseguimos observar que o que favorece mais o ganho superior a 50K é:

- *Education*: "Prof-school", "Doctorate", "Masters" e "Bachelors";
- *Occupation*: "Exec-manegerial";
- *Age*: entre os 49 a 60 anos;
- *Hoursperweek*: horas entre 39 a 65.

Conseguimos observar que o que **não** favorece o ganho superior a 50K é:

- *Education*: "7th-9th", "9th", "10th";
- *Occupation*: "Priv-house-serv";
- *Age*: inferiores a 29 anos;
- *Hoursperweek*: inferiores a 32 horas;
- *Sex*: "Female".

Neuronal Network – O que favorece? Sem educação

Conseguimos observar que o que favorece mais o ganho superior a 50K é:

- *Occupation*: "Armed-Forces", "Prof-specialty" e "Exec-manegerial";
- *Age*: entre os 40 a 60 anos;
- *Hoursperweek*: horas entre 53 e 65 e horas maiores que 65 horas;
- *Sex*: "Male".

Conseguimos observar que o que **não** favorece o ganho superior a 50K é:

- *Occupation*: "Other-services";
- *Age*: inferiores a 29 anos;
- *Hoursperweek*: inferiores a 32 horas.

Tópico 4

Association Rules

Association Rules

- Tudo em default.

Matriz classificação – com *education*

Contagens para Association_Rules em Salary:

Previsto	<=50K (Real)	>50K (Real)
<=50K	6718	1955
>50K	92	283

Matriz classificação – sem *education*

Contagens para Association_Rules1 em Salary:

Previsto	<=50K (Real)	>50K (Real)
<=50K	6768	2280
>50K	0	0

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.85	0.82
Accuracy	0.7738	0.7480
Precision	0.7547	ERRO
Recall	0.1264	0

Não vamos considerar os valores da *precision* e *recall*, do teste 2, na nossa análise.

Association Rules – Com educação

As 3 regras que têm probabilidade maior (maior importância), para ganhar mais de 50K são:

- *Education* = *Doctorate*, Occupation = *Exec-managerial* -> *Salary* => 50K;
- *Education* = *Prof-school*, *Hoursperweek* = 41 - 53 -> *Salary* => 50K;
- *Education* = *Prof-school*, Age = 39 - 49 -> *Salary* => 50K.

Association Rules – Com educação

As 3 regras que têm probabilidade maior, para ganhar **menos** de 50K são:

- *Age* < 29 -> *Salary* < 50K;
- *Age* < 29 *Education* = *HS-grad* -> *Salary* < 50K;
- *Age* < 29, *Hoursperweek* = 32 – 41 -> *Salary* < 50K.

Association Rules – Sem educação

As 3 regras que têm probabilidade maior, para ganhar mais de 50K são:

- *Ocuppation* = *Exec-managerial*, *Sex*= Male -> *Salary* => 50K;
- *Occupation* = *Exec-managerial*, *Age* = 49 – 60 -> *Salary* => 50K;
- *Hoursperweek* = 53 – 65, *Occupation* = *Exec-magerial* –> *Salary* => 50.

Association Rules – Sem educação

As 3 regras que têm probabilidade maior, para ganhar **menos** de 50K são:

- *Age* = 49 – 60, *Sex* = Female -> *Salary* < 50K;
- *Age* = 60, *Hoursperweek* < 32 -> *Salary* < 50K;
- *Age* >= 60, *Occupation* = *Adm-clerical* -> *Salary* < 50K.

Tópico 5

Logistic Regression

Logistic Regression

- Tudo em default.
- Podemos observar que, entre os dois testes efetuados, o teste com a educação é melhor, tanto o *score*, a *accuracy*, como a *precisão* e o *recall* são melhores em relação aos testes efetuados sem a educação.

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.89	0.88
Accuracy	0.8070	0.7929
Precision	0.6353	0.6234
Recall	0.4678	0.4508

Logistic Regression– O que favorece? Com educação

Conseguimos observar que o que favorece mais o ganho superior a 50K é:

- *Education*: “Prof-school”, “Doctorate”, “Masters”;
- *Occupation*: “Exec-manegerial”;
- *Age*: Idades entre os 49 a 60 anos;
- *Hoursperweek*: horas entre 53 a 65.

Conseguimos observar que o que **não** favorece o ganho superior a 50K é:

- *Education*: “Preschool”, “5th-6th”, “1st-4th”;
- *Occupation*: “Armed-Forces”;
- *Age*: Idades inferiores a 29 anos;
- *Hoursperweek*: inferiores a 32 horas;
- *Sex*: “Female”.

Logistic Regression– O que favorece? Sem educação

Conseguimos observar que o que favorece mais o ganho superior a 50K é:

- *Occupation*: “Prof-specialty” e “Exec-manegerial”;
- *Age*: Idades entre os 49 a 60 anos;
- *Hoursperweek*: horas entre 41 a 53 horas;
- *Sex*: “Male”.

Conseguimos observar que o que **não** favorece o ganho superior a 50K é:

- *Occupation*: “Armed-Forces”, “Priv-house-serv” e “Farming-fishing”;
- *Age*: Idades inferiores a 29 anos;
- *Hoursperweek*: inferiores a 32 horas;
- *Sex*: “Female”.

Tópico 6

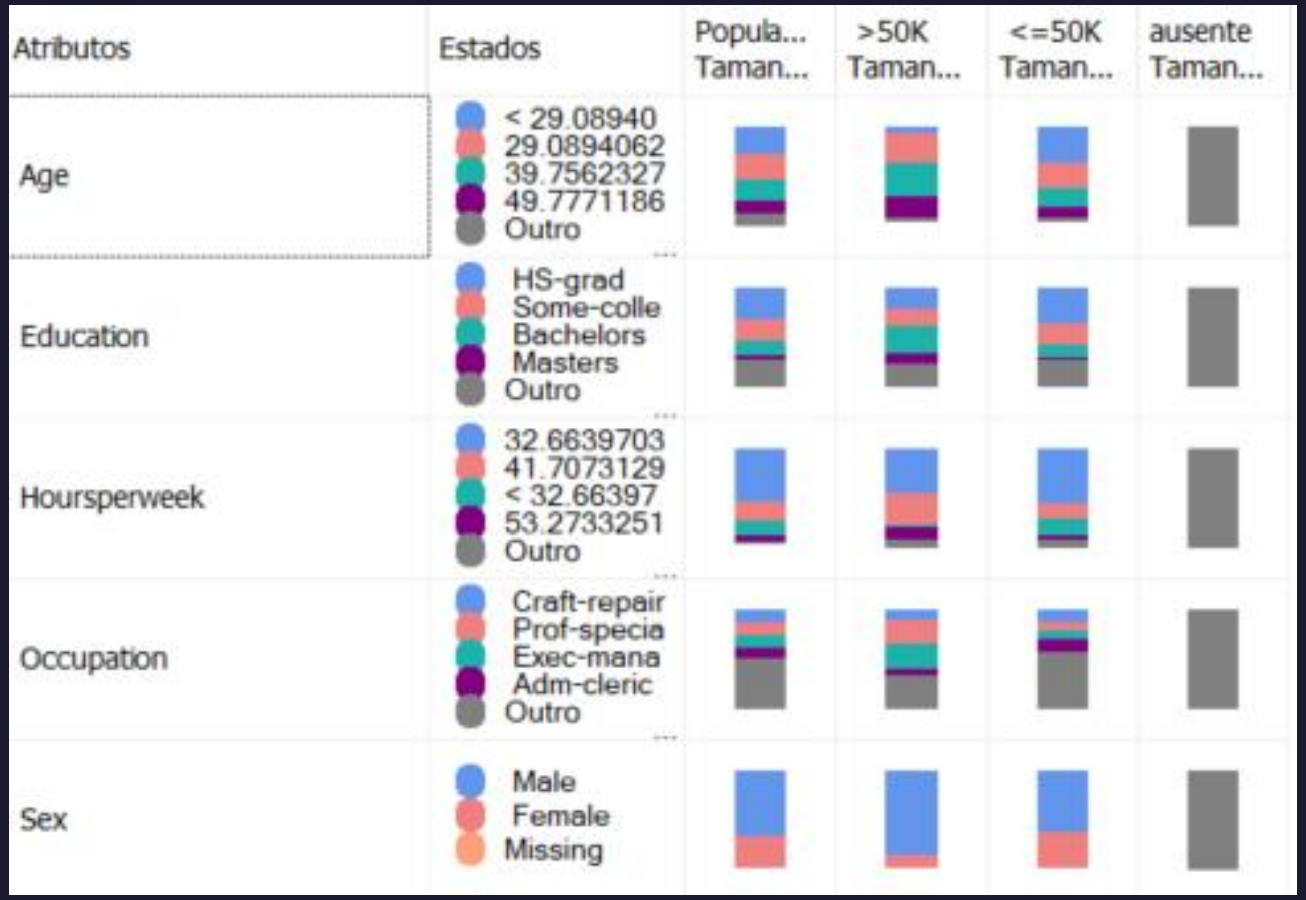
Native Bayes

Native Bayes

- Tudo em default.
- Podemos observar que, entre os dois testes efetuados, o teste com a educação é melhor, tanto o *score*, a *accuracy*, como a precisão e o *recall* são melhores em relação aos testes efetuados sem a educação.

Inputs	Teste 1	Teste 2
education	discrete	
occupation	discrete	discrete
hoursperweek	discretized	discretized
age	discretized	discretized
sex	discrete	discrete
Score	0.88	0.87
Accuracy	0.7980	0.7876
Precision	0.5975	0.5892
Recall	0.5626	0.5184

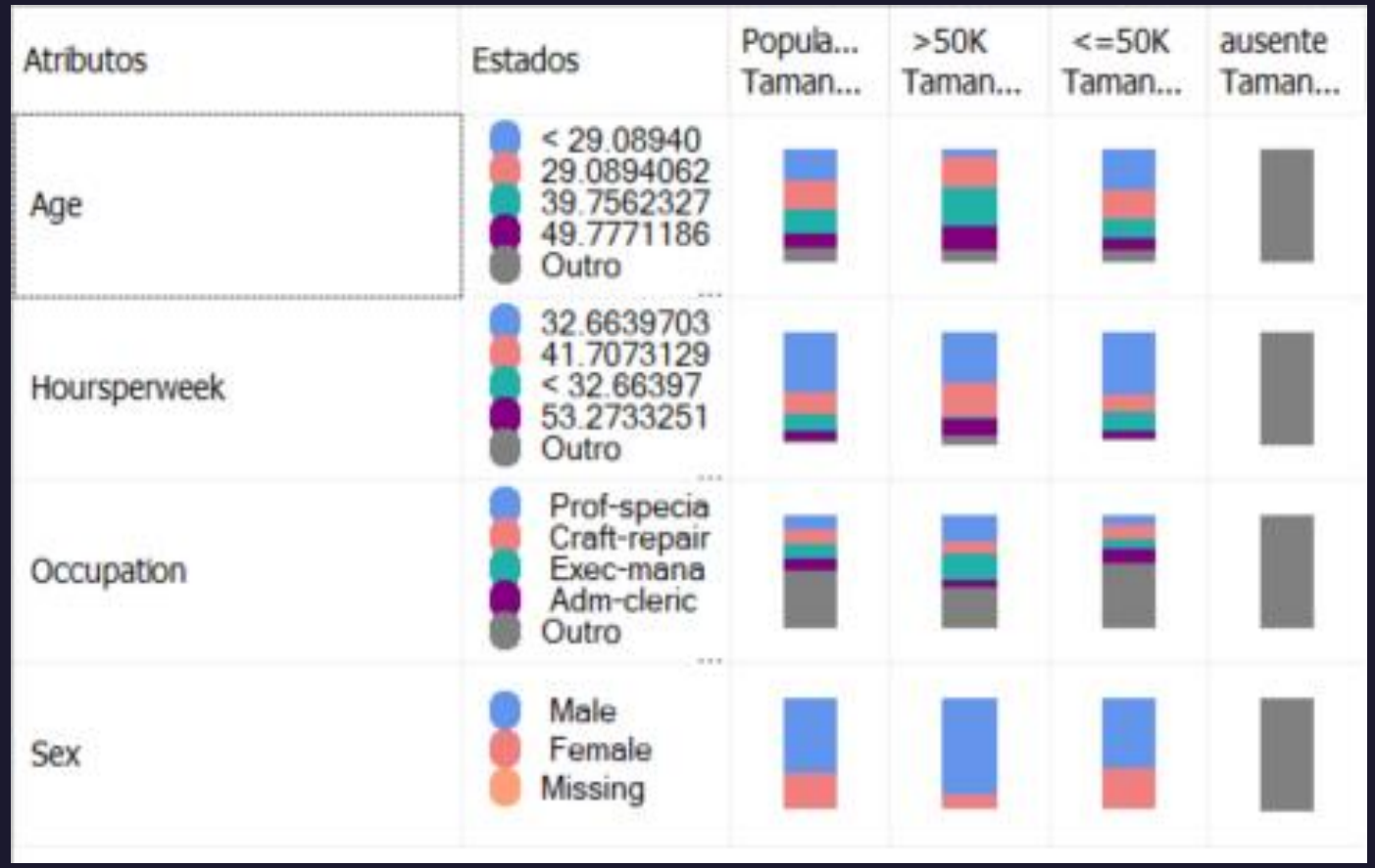
Native Bayes – Com educação



Native Bayes – Com educação

- A grande maioria das pessoas que ganham mais de 50K são do sexo masculino, temos vindo a observar ao longo dos diferentes classificadores, têm grau Bacharel, as principais ocupações são "Prof-Specialty" e "Exec-managerial" com idades compreendidas entre 39 e 49 anos e trabalham entre 32 a 41 horas por semana.
- Por outro lado para as pessoas que ganham menos de 50K, reparámos que a percentagem do sexo feminino aumenta, trabalham mais entre 32 a 41 horas, a percentagem de grau de escolaridade de secundário é maior e a maioria tem menos de 29 anos.

Native Bayes – Sem educação



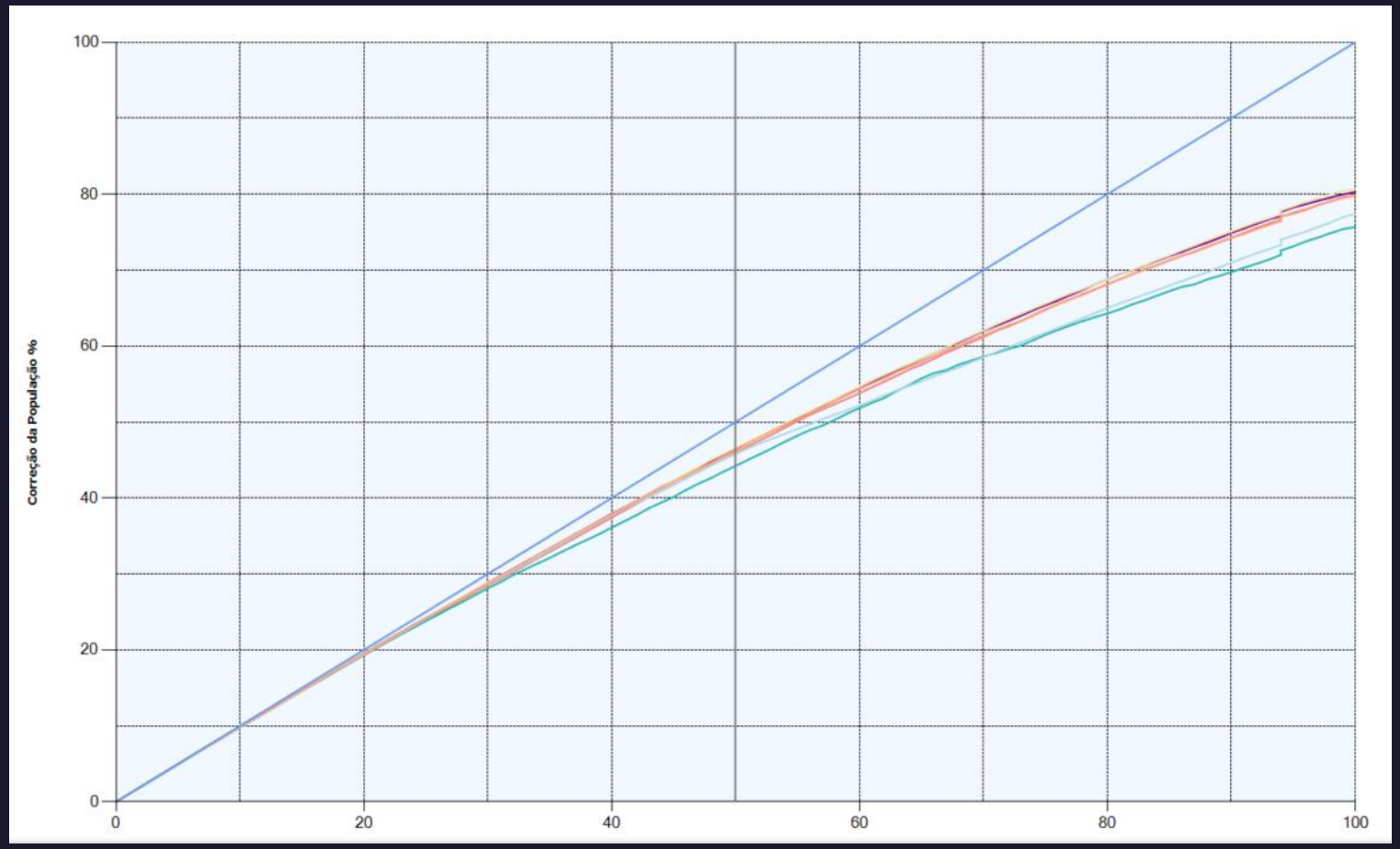
Native Bayes – Sem educação

- Aqui observamos mais uma vez que a maioria das pessoas que ganham mais de 50K são de sexo masculino, têm idades compreendidas entre 39 a 49 anos, com ocupação de "Prof-Specialty" e "Exec-managerial" e trabalham entre 32 a 41 horas por semana.
- Por outro lado, voltamos a observar, que quando as pessoas ganham menos de 50K a percentagem aumenta para o sexo feminino, trabalham entre 32 a 41 horas por semana, onde a maioria tem menos de 29 anos.

Gráfico de Comparação de Precisão **Com educação**

Percentual de população: 49,50%

Série, Modelo	Pontuaç...	Correçã...	Probabili...
Salaries 6	0,88	45,94%	84,08%
Clustering	0,84	44,23%	76,66%
Neural Netwo...	0,89	46,50%	84,06%
Logic_Regress...	0,89	46,63%	83,85%
Association_R...	0,85	45,80%	79,49%
Native Bayes	0,88	46,28%	89,90%
Modelo Ideal		50,00%	



Conclusões – Perguntas

- A educação é um parâmetro a considerar neste estudo?
- Qual é o melhor modelo?
- Os modelos indicam todos o mesmo?

A educação é um parâmetro a considerar neste estudo?

Sim pois a educação influencia bastante o salário ao final do mês, no entanto, depende se a pessoa vai ou não trabalhar na área correspondente ao grau de escolaridade. Foi por este motivo que fizemos ambos os testes, com e sem educação.

Qual é o melhor modelo?

Os dois classificadores que nos mostram melhores resultados são a Logistic Regression e a Rede Neuronal, onde para valores de correção da população a logistic apresenta valores melhores enquanto que para valores de probabilidade de previsão, a Rede Neuronal dá-nos melhores valores (os valores são muito próximos uns dos outros, as diferenças entre estes dois é muito pouca). Neste momento não conseguimos decidir qual das duas é a melhor, por isso fomos comparar as métricas de ambos e observámos que a Logistic Regression apresenta-nos melhores valores logo escolhemos este classificador como o melhor modelo.

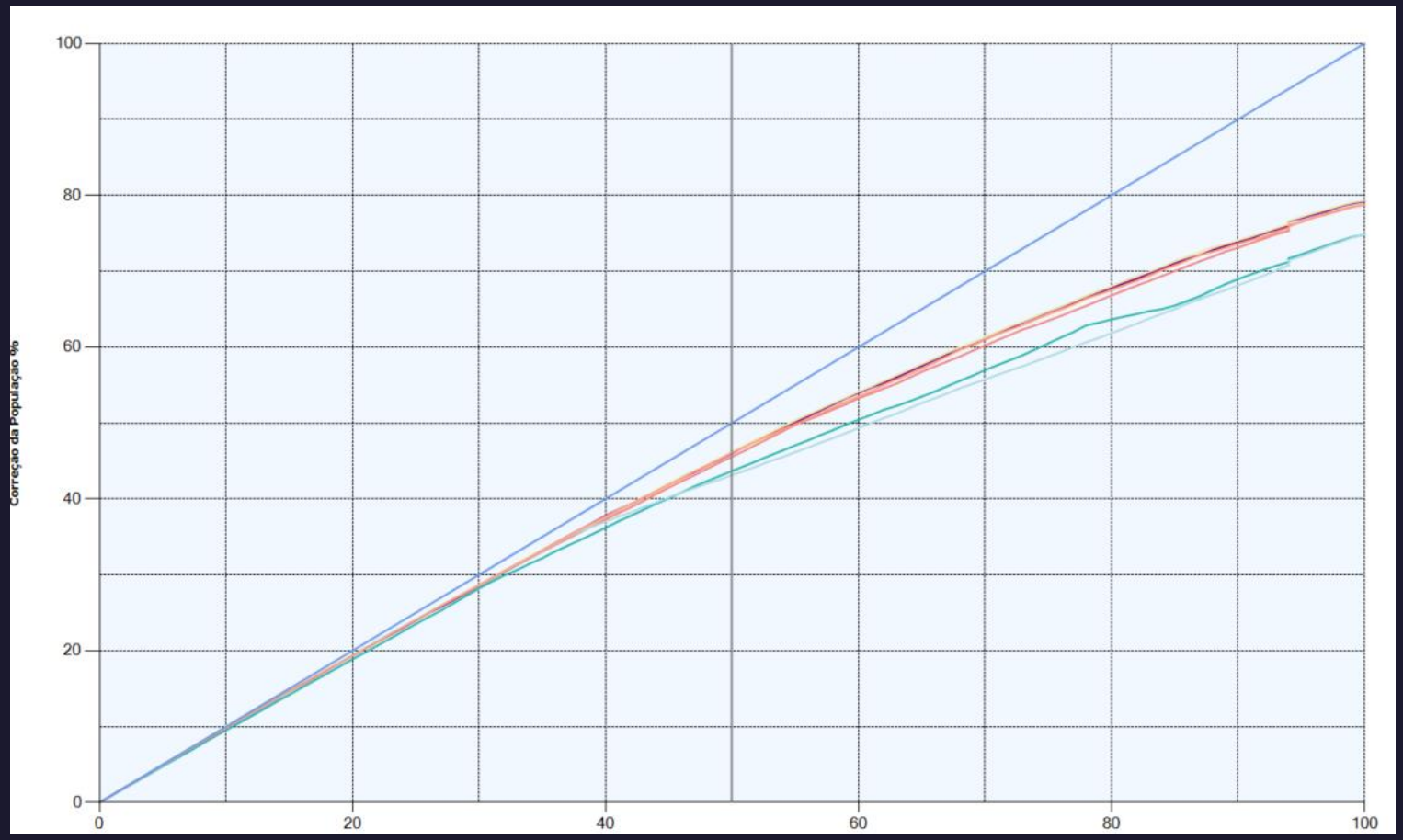
Os modelos indicam todos o mesmo?

De uma maneira geral, como visto anteriormente, todos os modelos indicam-nos quase sempre a mesma coisa. E ao comparar estes dados percebemos que correspondem à nossa realidade.

Gráfico de Comparação de Precisão **Sem educação**

Percentual de população: 49,50%

Série, Modelo	Pontuaç...	Correçã...	Probabili...
Salaries 8	0,87	45,56%	82,02%
Clustering1	0,83	43,68%	68,27%
Neuronal Net...	0,88	45,99%	82,27%
Logist_Regres...	0,88	46,16%	81,31%
Association_R...	0,82	43,15%	75,24%
Native Bayes1	0,87	46,07%	83,79%
Modelo Ideal		50,00%	



Conclusão Final – Sem educação

Qual é o melhor modelo?

Os dois classificadores que nos mostram melhores resultados são a Logistic Regression e a Rede Neuronal, onde para valores de correção da população a logistic apresenta valores melhores, enquanto que para valores de probabilidade de previsão a Rede Neuronal dá-nos melhores valores (os valores são muito próximos uns dos outros, as diferenças entre estes dois é muito pouca). Neste momento não conseguimos decidir qual das duas é a melhor, por isso fomos comparar as métricas de ambos e observamos que a Rede Neuronal apenas tem a precision como melhor valor, e como a Logistic Regression apresenta-nos melhores valores para as outras métricas (accuracy e o recall), continuamos a escolher este classificador como o melhor modelo.

Obrigado

Patrícia Silva - 2016014544

Emanuel Saraiva - 2019130219