U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

ELEIÇÕES

Agentes e Inteligência Artificial Distribuída
Mestrado Integrado em Engenharia Informática e Computação

Grupo 24
Bárbara Sofia Silva (201505628)
Julieta Frade (201506530)
Ventura Pereira (201404690)

2018/2019

APRESENTAÇÃO

Primeira Parte

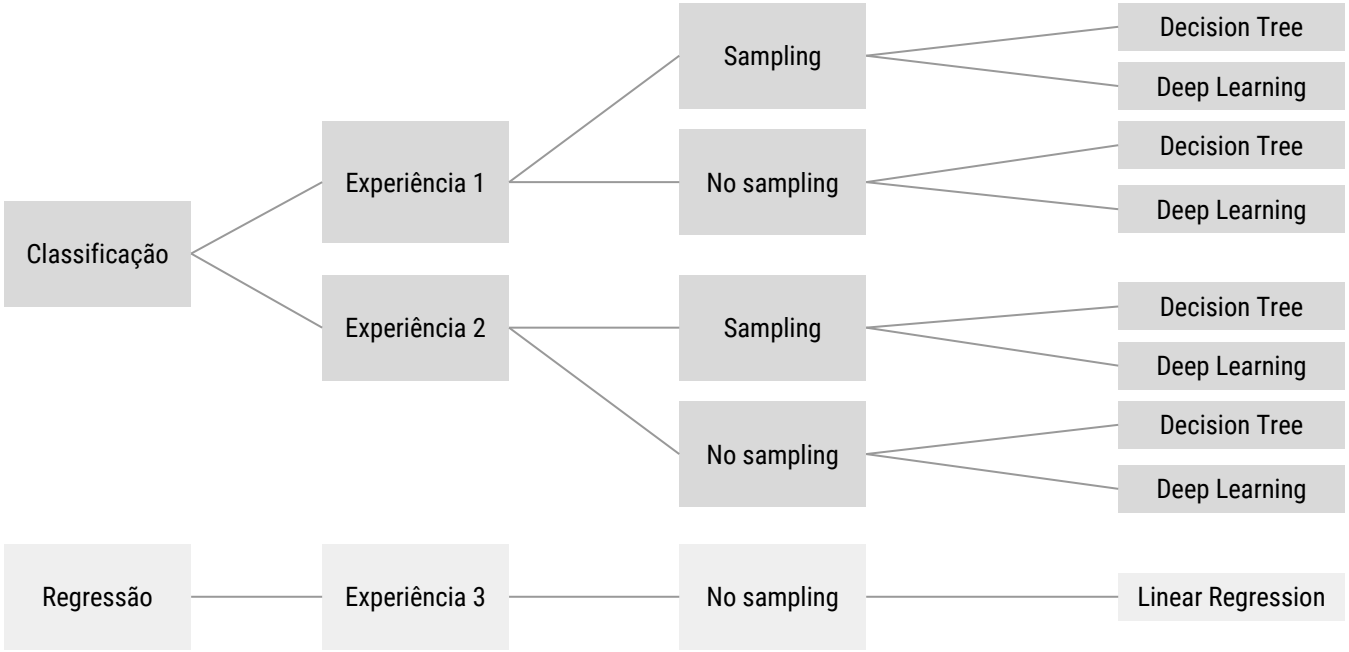
DESCRIÇÃO DO PROBLEMA DE ANÁLISE DE DADOS

Tanto os **candidatos** como os **eleitores** possuem **crenças políticas**. No caso dos candidatos este conjunto representa a sua **estratégia** que poderá mudar ao longo do processo eleitoral, contudo se a alterarem em demasia, vão sofrer perdas no que toca ao fator de **credibilidade**.
Cada candidato tem uma **equipa de campanha** composta por **chefes de estado**. Cada elemento desta equipa é responsável por fazer **sondagens** à população do estado que lhe foi atribuído e seguidamente **aconselhar** o seu candidato em relação à próxima mudança de estratégia. O candidato decide se altera ou não a crença de acordo com a sua **teimosia**.
Face ao fator de credibilidade de cada candidato, o eleitor exige um valor mínimo de credibilidade por parte destes, caso nenhum tenha atingido este nível a pessoa abster-se-á. Se nenhum candidato tiver em pelo menos metade das crenças o nível mínimo requerido pelo eleitor, este também poderá exercer o seu direito de **abstinência**.

***PROBLEMA 1 (CLASSIFICAÇÃO)**: Como é que o facto de um candidato ter chefe de estado ou não num estado, a sua credibilidade e teimosia, a população desse mesmo estado, se mudou de crença de acordo com a opinião pública e se a sua crença principal é a mesma que a do estado afeta a vitória de um candidato nesse estado?*

***PROBLEMA 2 (CLASSIFICAÇÃO)**: Como é que a credibilidade, teimosia, número de chefes de estado e percentagem de população que comunica com os seus chefes de estado afeta a vitória global de um candidato?*

EXPERIÊNCIAS REALIZADAS



EXPERIÊNCIAS REALIZADAS

EXPERIÊNCIA 1

Objetivo:

Responder ao Problema 1 (**Classificação**).

Variáveis independentes:

hasChief - *boleano*; se o candidato tinha ou não chefe de estado naquele estado.

percentageStatePopulation - *inteiro*; percentagem de população que pertence ao estado.

credibility - *inteiro*; percentagem de credibilidade do candidato.

stubbornness - *inteiro*; percentagem de teimosia do candidato.

changedBelief - *true, false ou unknown;* se o candidato mudou a crença de acordo com a opinião do estado. Toma o valor de unknown se o candidato não tomou conhecimento desta.

sameBelief - *boleano*; se a crença principal do candidato é a mesma do que a do estado

Variável dependente:

wonState - *boleano*; se o candidato ganhou ou não naquele estado

EXPERIÊNCIA 2

Objetivo:

Responder ao Problema 2 (**Classificação**).

Variáveis independentes:

nChiefs - *inteiro*; quantidade total de chefes de estado de cada candidato.

percentageCandidatePopulation - *inteiro*; percentagem de população que comunica com os chefes de estado do candidato.

credibility - *inteiro*; percentagem de credibilidade do candidato.

stubbornness - *inteiro*; percentagem de teimosia do candidato.

Variável dependente:

won - *boleano*, se o candidato ganhou ou não.

ESTATÍSTICAS SOBRE OS DADOS RECOLHIDOS

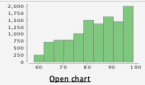
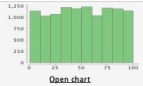
De forma a obter estatísticas adequadas face aos dados recolhidos, foi decidido gerar um **ficheiro CSV**. Neste ficheiro de dados, **cada linha corresponde a um candidato num estado**, ou seja, nessa linha conseguimos perceber a posição desse candidato num determinado estado. Cada coluna corresponde a uma certa informação sobre o mesmo.

INFORMAÇÕES SOBRE UM CANDIDATO
candidateId, stateId, hasChief, percentageStatePopulation, wonState, percentageStateVotes, credibility, stubbornness, won, percentageVotes, changedBelief, sameBelief, nChiefs, percentageCandidatePopulation

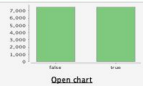
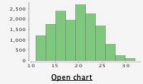
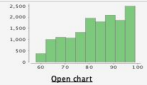
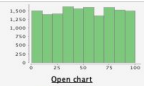
Para cada experiência foram tidos em conta **diferentes dados**, neste caso colunas. Assim como utilização dos dados **com sampling** ou **sem sampling**. Na experiência 1 foi usado um sampling de **5750 true, 5750 false**. Na experiência 2 **6500 true, 6500 false**.
Nos slides seguintes averiguamos detalhadamente as estatísticas face aos dados de cada experiência.

ESTATÍSTICAS DOS DADOS DA EXPERIÊNCIA 1

COM SAMPLING

Label											
wonState	Polynominal	0	Least: true (5750)	Most: false (5750)	Values: false (5750), true (5750)						
hasChief	Binominal	0	 Open chart	Least: false (5747)	Most: true (5753)	Values: true (5753), false (5747) Details...					
percentageStatePopulation	Integer	0	 Open chart	Min: 11	Max: 32	Average: 19.580	Deviation: 4.179				
credibility	Integer	0	 Open chart	Min: 58	Max: 99	Average: 83.393	Deviation: 10.435				
stubbornness	Integer	0	 Open chart	Min: 1	Max: 100	Average: 51.163	Deviation: 28.689				
changedBelief	Polynominal	0	 Open chart	Least: false (2768)	Most: unknown (5747)	Values: unknown (5747), true (2985), false (2768) Details...					
sameBelief	Binominal	0	 Open chart	Least: true (2480)	Most: false (9020)	Values: false (9020), true (2480) Details...					

SEM SAMPLING

Label											
wonState	Polynominal	0	Least: true (5769)	Most: false (9231)	Values: false (9231), true (5769)						
hasChief	Binominal	0	 Open chart	Least: true (7496)	Most: false (7504)	Values: false (7504), true (7496) Details...					
percentageStatePopulation	Integer	0	 Open chart	Min: 11	Max: 32	Average: 19.552	Deviation: 4.177				
credibility	Integer	0	 Open chart	Min: 58	Max: 99	Average: 83.023	Deviation: 10.549				
stubbornness	Integer	0	 Open chart	Min: 1	Max: 100	Average: 50.894	Deviation: 28.723				
changedBelief	Polynominal	0	 Open chart	Least: false (3736)	Most: unknown (7504)	Values: unknown (7504), true (3760), false (3736) Details...					
sameBelief	Binominal	0	 Open chart	Least: true (3164)	Most: false (11836)	Values: false (11836), true (3164) Details...					

ESTATÍSTICAS DOS DADOS DA EXPERIÊNCIA 2

COM SAMPLING

Label			Least	Most	Values
won	Polynominal	0	true (6500)	false (6500)	false (6500), true (6500)

credibility	Integer	0	Min 58	Max 99	Average 83.262	Deviation 10.500	Open chart
stubbornness	Integer	0	Min 1	Max 100	Average 50.932	Deviation 28.767	Open chart
nChiefs	Integer	0	Min 0	Max 5	Average 2.504	Deviation 1.103	Open chart
percentageCandidatePopulation	Integer	0	Min 0	Max 100	Average 49.603	Deviation 22.581	Open chart

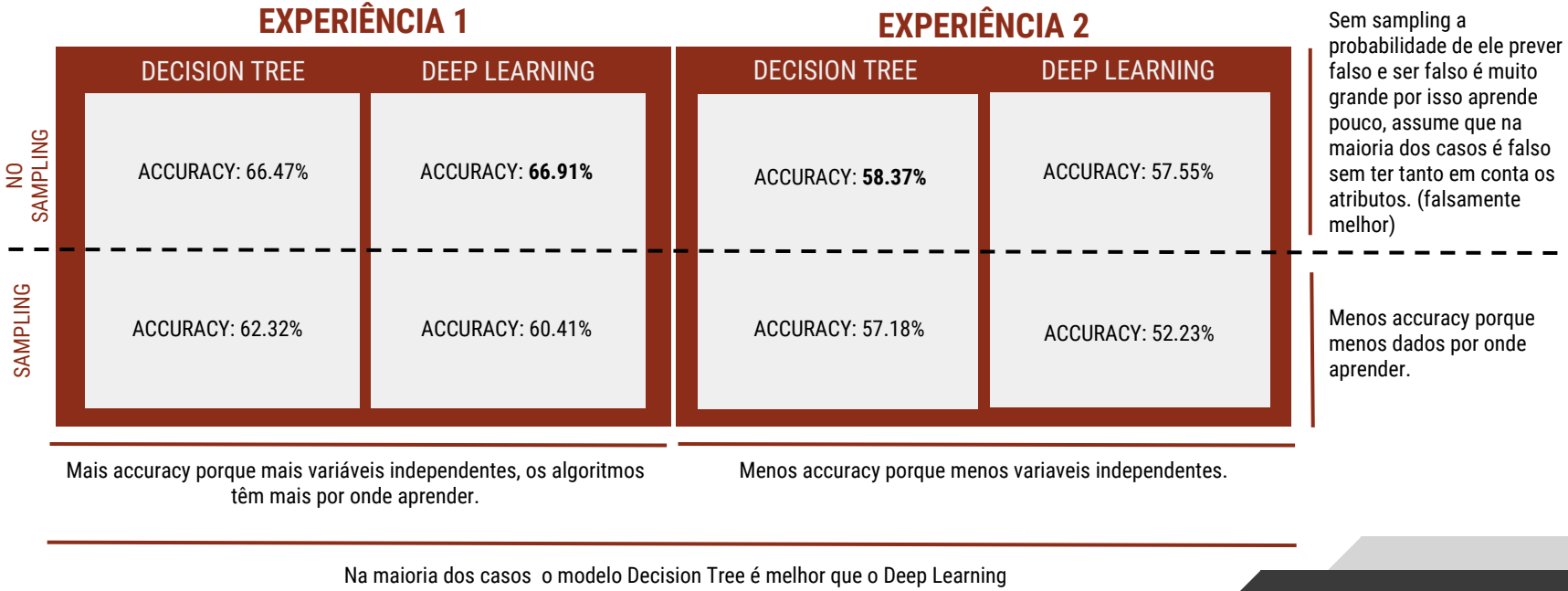
SEM SAMPLING

Label			Least	Most	Values
won	Polynominal	0	true (6595)	false (8405)	false (8405), true (6595)

credibility	Integer	0	Min 58	Max 99	Average 83.023	Deviation 10.549	Open chart
stubbornness	Integer	0	Min 1	Max 100	Average 50.894	Deviation 28.723	Open chart
nChiefs	Integer	0	Min 0	Max 5	Average 2.500	Deviation 1.105	Open chart
percentageCandidatePopulation	Integer	0	Min 0	Max 100	Average 49.536	Deviation 22.604	Open chart

ANÁLISE DE DADOS COM RAPIDMINER

Obs: *accuracy=100%-classification error*



CONCLUSÕES

No final destas experiências, o grupo pode chegar a várias considerações relativas aos resultados das mesmas. Primeiramente, de forma a tornar os resultados mais fiáveis, no âmbito de diminuir o ***erro*** e aumentar a fonte de ***informação*** , deveríamos ter gerado mais dados para análise. Contudo, restringe-se este pensamento ao facto de baixar o erro no treino, não significa que o modelo seja melhor com os novos casos adicionados. Para além disso, o grupo não quis tornar o modelo ***overfitted*** nem **sobreajustado**, visando uma generalização mais correta.
Em segundo lugar, no seguimento do nosso tema "***Eleições***", considerámos que dever-se-ia ter envolvido, mais veemente, variáveis que envolvessem as ***crenças dos eleitores e dos voters*** , uma vez que este parâmetro era fulcral no algoritmo e respetivo resultado do nosso trabalho. Contudo, foram encontradas dificuldades em enquadrar determinados tipos de variáveis, nomeadamente *arrays* e intervalos de valores, na análise.
Por último, mais problemas poderiam ter sido analisados, como seria algum cingente na ***abstinência*** de um voter.
O grupo conclui, portanto, e em tom e aprendizagem para um trabalho futuro que envolva processos de data mining, que se deve ter especial atenção ao conjunto de variáveis escolhidas para análise, assim como a quantidade de dados e possíveis problemas de análise, nunca tentando manter um modelo óptimo, sem **sobreajustamento**.

INFORMAÇÃO ADICIONAL

Segunda Parte

PROCESSOS RAPIDMINER

Como ilustrado no **esquema do slide 4**, no total foram feitos **9 processos RapidMiner** para analisar os dados gerados pelo programa.
Em alguns destes processos decidimos utilizar o operador **Sample**. Este operador cria uma amostra dos dados selecionando exemplos aleatoriamente, e o objetivo da sua utilização foi equilibrar a amostra.
Já o operador **Split Validation** foi utilizado em todos os processos, este distribui aleatoriamente os dados por conjunto de treino e conjunto de teste. Justifica-se separar a amostra pois uma avaliação com dados de treino de modelação:

- Estimador pouco fiável do comportamento do modelo em novos exemplos;
- Assume que os casos do futuro serão iguais aos de treino.

PROPORÇÃO UTILIZADA

- **70%** dos casos para treino
- **30%** dos casos para teste

PROCESSOS RAPIDMINER

O que diferencia cada experiência é nomeadamente a utilização de um dos seguintes operadores:

DECISION TREE
Gera um modelo de árvore de decisão, que pode ser usado para classificação e regressão.

DEEP LEARNING
Executa o algoritmo Deep Learning utilizando H2O 3.8.2.6.

LINEAR REGRESSION
Calcula um modelo de regressão linear a partir da amostra fornecida.

RESULTADOS DE OUTRAS EXPERIÊNCIAS

PROBLEMA 3 (REGRESSÃO): Como é que a credibilidade, teimosia, número de chefes de estado e percentagem de população que comunica com os seus chefes de estado afeta a percentagem de votos de cada candidato?

EXPERIÊNCIA 3

Objetivo:

Responder ao Problema 3 (**Regressão**).

Variáveis independentes:

nChiefs - *inteiro*; quantidade total de chefes de estado de cada candidato.

percentageCandidatePopulation - *inteiro*; percentagem de população que comunica com os chefes de estado do candidato.

credibility - *inteiro*; percentagem de credibilidade do candidato.

stubbornness - *inteiro*; percentagem de teimosia do candidato.

Variável dependente:

percentageVotes - *inteiro*, percentagem de votos de cada candidato

Como para este modelo só podemos usar variáveis independentes numéricas, fomos obrigados a escolher poucos das que tínhamos. Isto resultou numa fraca correlação e num alto root relative squared error.

> CORRELATION: 0.124

> ROOT MEAN SQUARED ERROR: 7.910 +- 0.0

> ROOT RELATIVE SQUARED ERROR: 0.993 [está no intervalo entre 0 e 1, portanto pode-se considerar que o modelo é útil em relação à previsão trivial (médio dos valores) visto que tem menor erro que esta.]

OUTRAS OBSERVAÇÕES

De forma a dar **setup aos processos RapidMiner**, deverá ser **atualizado o path do ficheiro CSV** a ser importado pelo operador **Read CSV**, visto que este path varia de computador para computador.
O ficheiro de dados **rapidData.csv** deverá ser importado e ajustado a cada experiência, como detalhado no slide 5 e 14.