Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Bacharelado em Tecnologia da Informação

Disciplina: Aprendizado de Máquina
Discentes:
Gabriel Cristian Melo da Silva
Gabriel Estácio de Souza Passos
Rodrigo Faustino de Sousa

Docente: SAMUEL DA SILVA OLIVEIRA

Aprendizado de Máquina 2020.2- Checkpoint 1: Pré-proces
Descrição da sua base de dados samento dos dados

1. Faça uma breve descrição do seu problema a ser resolvido e da sua base de dados.

O problema que vamos abordar busca determinar, a partir de um conjunto de estatísticas para cada jogador da NBA (liga americana de basquete), quem foi o MVP (most valuable player) da temporada, partindo da temporada de 1996/97 até a de 2019/20. O MVP é um prêmio dado ao jogador com melhor desempenho na competição.

A base, disponibilizada em formato .csv, possui inicialmente 11145 instâncias, divididas entre 22 atributos.

2. Como os dados foram adquiridos?

Os dados foram adquiridos pela plataforma Kaggle, sendo disponibilizados por Justinas Cirtautas, e pode ser obtida aqui.

3. Responda:

a. Quantos atributos categóricos e numéricos a base possui?
Inicialmente, são 8 atributos categóricos e 14 numéricos.

b. Para os atributos categóricos, quantas classes existem para cada atributo?
player_name: 2235

team_abbreviation: 36
college: 316
country: 76
draft_year: 45
draft_round: 8
draft_number: 75
season: 24

c. Para os atributos numéricos, são todos escalares? Se sim, quais as escalas? Nenhum dos atributos foram transformados para novas escalas. Os atributos ‘Unnamed: 0’, ‘age’, ‘player_height’, ‘player_weight’, ‘gp’, ‘pts’, ‘reb’, ‘ast’, ‘net_rating’ se encontram nos seus valores originais. Já os atributos oreb_pct, dreb_pct, usg_pct, ts_pct, ast_pct são valores percentuais, portanto, não há necessidade de transformação.

d. A sua base possui valores faltosos? Não.

Primeiras manipulações com os dados

4. A sua base de dados vai precisar de algum tipo de processamento? Se sim, descreva o que foi feito para limpar e organizar a base

A base já estava praticamente pronta: não possuía valores faltantes, nem valores duplicados e os outliers detectados faziam sentido dentro do contexto do problema. Logo, a única limpeza necessária foi remover as colunas que não geram impacto no resultado e transformar as variáveis categóricas em numéricas.

Sobre transformações, foi necessário converter o atributo 'season' de categórico pra numérico e o atributo 'player_name' (categórico) para 'player_id' (numérico), e aplicar uma normalização nas entradas do dataset.

Além disso, foi necessário criar uma coluna target para o problema que queremos abordar. Para isso, criamos um atributo binário chamado **'was_mvp'**, onde 0 indica que o jogador não foi o MVP da temporada e o 1 indica o caso contrário.

Redução de instâncias

5. Aplique um método para redução de instâncias na sua base de dados. Pode ser um visto em sala de aula ou um escolhido pelo aluno. Caso seja um método escolhido pelo aluno, é importante fazer uma descrição do mesmo. Defina esta base como: BaseReduzida1

O método escolhido foi o de amostragem aleatória simples, onde um determinado número de instâncias é retirado da base de forma aleatória e todos os elementos têm a mesma probabilidade de serem selecionados.

6. Responda:

a. Qual foi a redução de instâncias como resultado do método de redução?

Foram selecionadas 70% das instâncias da base original, o que corresponde a 7.801, gerando uma redução de 3.344 instâncias.

Seleção de atributos

7. Aplique um método para seleção de atributos na sua base de dados. Novamente, pode ser um visto em sala de aula ou um escolhido pelo aluno. Caso seja um método escolhido pelo aluno, é importante fazer uma descrição do mesmo. Defina esta base como: BaseReduzida2

8. Responda:

a. Qual foi a redução no número de atributos como resultado do método de redução? O método usado para seleção dos atributos foi o de Florestas Aleatórias, que foi descrito no material extra fornecido no GitHub da disciplina. No lugar de usar somente uma árvore de decisão, são criadas múltiplas árvores de decisão usando amostras do conjunto original, com fim de reduzir a variância e o *overfit* que são inerentes ao método de árvores de decisão. No caso em si, foi usado com parâmetros bem semelhantes ao exemplo do GitHub, com max_features = 6 e threshold = 'median'. Valores maiores de 6 não resultaram em novos campos sendo escolhidos. Executando o método múltiplas vezes causou pouca mudança, onde as vezes o atributo ts_pct apareceu no lugar do reb ou net_rating.

O resultado foram as seguintes colunas escolhidas pelo método: pts, reb, ast, net_rating, usg_pct, ast_pct. Dessa forma, contando a coluna target, reduzimos de 13 para 6 atributos.

Extração de Atributos

9. Aplique um método para extração de atributos na sua base de dados. Novamente, pode ser um visto em sala de aula ou um escolhido pelo aluno. Caso seja um método escolhido pelo aluno, é importante fazer uma descrição do mesmo. Defina esta base como: BaseReduzida3

10. Responda a seguinte pergunta:

a. Qual foi a redução no número de atributos como resultado do método de redução?

Para a Extração de atributos foi utilizado o algoritmo PCA (Principal Component Analysis). Redução de 13 para 5.

Preencha a seguinte tabela:

	Número de Instâncias	**Número de atributos**
Base Original	11145	22
Base Reduzida 1	7801	13
Base Reduzida 2	7801	7
Base Reduzida 3	7801	5