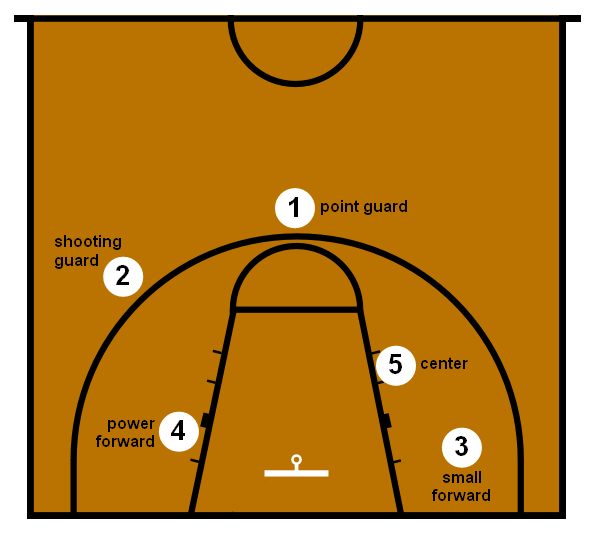
**Predição da posição de jogadores de basquete**

**Autores: Felipe Machado e Samuel Porto**

**Introdução**

* **Objetivo do projeto**

A “National Basketball Association” (NBA) é a principal liga de basquetebol profissional da América do Norte. A NBA é uma das 4 “grandes ligas” de esporte profissional do continente. Os seus jogadores são os mais bem pagos esportistas do mundo, por salário médio anual. No basquete, os jogadores são separados de acordo com três posições: center, forward, guard.



Diante disso, o intuito do projeto é realizar uma análise de dados correspondentes à jogadores da NBA (National Basketball Association) e a partir dos mesmos, conseguir determinar a posição de cada jogador no seu time. O projeto envolve uma análise estatística dos dados, e posteriormente, uma série de algoritmos que possibilitam que o usuário forneça os dados do jogador, e o código retorne a posição do mesmo.

Esse algoritmo pode ser aplicado em outros esportes, tais como vôlei, futsal e futebol americano.

* **Organização do documento (Jupyter)**

1. Introdução: importação da base de dados e conversão da mesma, para as unidades desejadas.
2. Análise Centers
3. Análise Guards
4. Análise Forwards
5. Junção dos DataFrames
6. Análise da altura dos jogadores
7. Análise do peso dos jogadores
8. Relacionando as variáveis
9. Separando em treinamento e teste
10. Prevendo as posições

**Metodologia e Materiais**

1. **Introdução**

Nas próximas linhas, a base de dados está sendo preparada para a análise. Primeiramente, é importante ressaltar que nem sempre a base de dados são ideais para todos. No nosso projeto, a base de dados foi obtida do site oficial da NBA, e foram utilizados dados da Temporada de 2017-2018. Dessa base, foram obtidas três outras sub-bases, correspondentes à cada posição. Por serem dados norte-americanos, as informações físicas dos jogadores estavam em outro sistema de unidades. O peso encontrava-se em libras, enquanto que a altura estava em pés. Com isso, foi necessário converter essas informações para as unidades mais utilizadas no Brasil, que são: quilograma para o peso dos jogadores, e metro/centímetros (nesse caso convertemos para centímetros) para altura dos jogadores.

Diante disso, o excel entendia a altura como data, e para reverter a situação, usufruímos de ferramentas do excel. Com isso, a coluna "HEIGHT" corresponde a altura oficial do jogador, em centímetros. A coluna "HEIGHT 1" corresponde ao digito anterior a vírgula na base de dados; esse primeiro digito estava em pés, enquanto que os dígitos após a vírgula, representados pela coluna "HEIGHT 0,1" estavam em inches. Diante disso ao converter os dígitos antes da virgula, de pés para centímetros e os dígitos posteriores à virgula de inches para centímetros, somamos e obtemos a altura dos jogadores em centímetros.

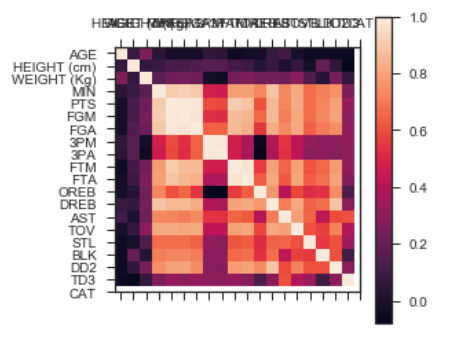
Vale ressaltar, que na análise de cada posição, foi adicionado uma coluna indicando a posição de cada jogador. Isso foi feito, visto que os dataframes vieram separados por posição, e para distinguirmos as posições

1. **Análise Centers**

Ao analisarmos os dados referentes aos “Centers”, observamos alguns comportamentos interessantes. Primeiramente, ao obtermos as correlações entre a idade desses jogadores, e as demais variáveis, percebemos correlações muito baixas. Já em relação à altura e peso, foi observado maiores correlações, que em relação à idade, porém, não permitiram obtermos grandes conclusões. Podemos constatar que os centers possuem médias bastante elevadas. A média da altura desses jogadores é 210 cm, quanto que a média do peso é 112kg.

Normalmente esses são os maiores jogadores em quadra, e são excelentes em garantir pontos de 2 perto da cesta, pegar rebotes, impedir e bloquear bandejas dos adversários. Diante disso, as duas médias citadas anteriormente, comprovam isso, visto que são altos e pesados. Por jogarem debaixo da cesta, no garrafão, não costumam arremessar da linha de três, algo comprovado pela média de cesta de 3 por jogo ser bastante baixa, aproximadamente 0,3. Em compensação, são responsáveis pela maioria dos rebotes defensivos; isso pode ser mostrado pela média de rebotes defensivos dos centers ser aproximadamente 4. Além dos rebotes defensivos, no quesito “Rebotes Ofensivos” os “Centers” possuem uma média muito superior as demais posições. Enquanto os “Centers” tem uma média de 1,46 para rebotes ofensivos, os “Guards” possuem a média de 0,43 e os “Forwards” de 0,84.

Diante disso, para visualizarmos melhor as correlações existentes entre cada variável, foi utilizada uma matriz de correlação. Nela, podemos perceber que para os Centers, as únicas correlações significativamente elevadas são as relacionadas ao tempo de jogo e pontos feitos, ou arremessos feitos e arremessos tentados, algo que não muda entre as posições. Com a análise dessas variáveis relacionadas somente à uma posição, acaba tornando-se insuficiente para conclusões mais aprofundadas. Assim, é necessário analisar os dados das outras duas posições separadamente, e posteriormente analisa-las conjuntamente.

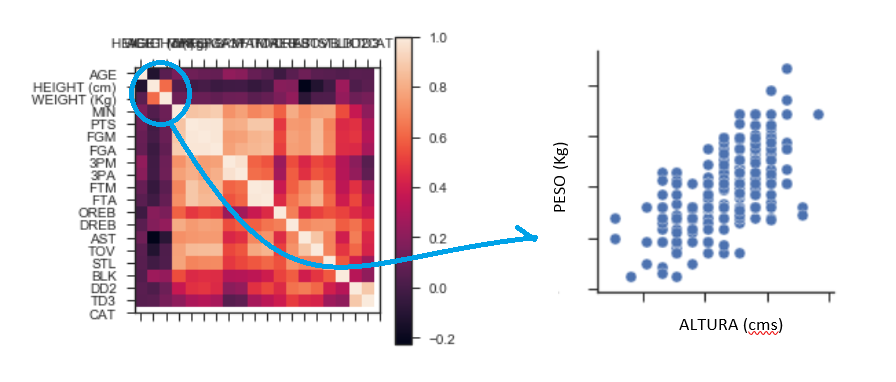


1. **Análise Guards**

Os jogadores “Guards” são considerados os com melhor controle de bola, passe e arremesso. Normalmente são ágeis e conseguem efetuar arremessos de dois e três pontos. São esses jogadores que organizam as jogadas, e são considerados em sua maioria os “play-makers”. Normalmente são os menores jogadores em quadra.

Diante dessas constatações, algumas observações são imprescindíveis para prever a posição de futuros jogadores. Ao observamos a média dos dados dos “Guards”, sendo esta 193,7 centímetros, percebemos que a média de altura é significativamente menor que a dos “Centers” (210 centímetros). Além da altura, o peso também teve uma média bem menor, sendo esta 90,46 kg. Isso demonstra o porquê são considerados jogadores mais ágeis e menores. Ademais, outro dado importante é o da média de assistências, que é bem comparado as demais posições. Enquanto a média de assistência dos “Guards” é de 2,24, a mesma média para “Centers” é de 1,18 e dos “Forwards” 1,36.

Em relação aos “Guards”, ao olharmos para as correlações, podemos ver que a Altura e o Peso possuem uma correlação bem mais significativa que no caso dos “Centers”. O Coeficiente de Pearson dessas duas variáveis corresponde à aproximadamente 0,62, uma correlação significativa. Não há grandes constatações a serem feitas analisando somente uma única posição. Diante disso, analisaremos os “Forwards”, e logo em seguida, uma análise das diferenças entre os dados obtidos das três posições.



Na imagem acima, fica claro a moderada correlação existente para os “Guards”, que antes, não estava presente no grupo dos “Centers”.

1. **Análise Forwards**

Os jogadores considerados “Forwards” são os mais versáteis de todos em quadra. Esses jogadores são considerados fortes e ágeis, e são capazes de realizar arremessos certeiros e também lidar com bandejas e pontos na área do garrafão.

Diante das afirmações acima, os dados obtidos dos “Forwards” nos mostram isso. Primeiramente, esses possuem uma média de altura de 204,1 centímetros, ou seja, estão no meio entre os “Centers” com média de altura de 210 centímetros e os “Guards” cuja média de altura é de 193,7 centímetros. Além da altura, o peso também comprova a versatilidade desses jogadores: enquanto os “Forwards” apresentaram uma média de peso de 102,9 kg, os “Centers” mostraram-se mais pesados, atingindo 112kg, enquanto os “Guards” apresentaram-se mais leves, com uma média de 90,46kg.

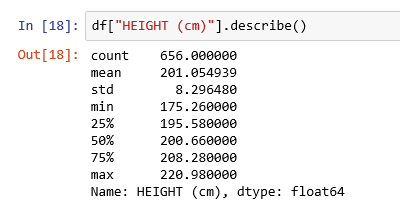
Algo interessante dos “Forwards” é que esses jogadores possuem uma média de tentativas de arremessos de três pontos um pouco menor que os “Guards” e muito maior que os “Centers”. Em compensação, esses mesmos jogadores possuem uma média de rebotes (tanto ofensivos, quanto defensivos) significativamente maior que a média dos “Guards”, porém, muito menor que a média dos “Centers”. Isso reforça a ideia de que os “Forwards” são jogadores “híbridos”, ou seja, os mais versáteis em quadra.

Além disso, as correlações entre as variáveis não são tão significativas para o projeto, quando analisadas de posições diferentes, separadamente. As correlações seguem o mesmo padrão da das outras posições, com algumas mudanças ínfimas, incapazes de provocar mudanças significativas no algoritmo.

**OBS: Após a análise separada em posições, dos DataFrames, a próxima etapa é juntar as três posições em um único DataFrame e fazer uma análise completa, procurando evidências que possam auxiliar na predição da posição.**

1. **Análise da Altura**

Nessa etapa analisamos o comportamento da altura dos jogadores, para tal juntamos os data frames de cada posição com o comando *pd.concat()* do pandas. A variável *df* recebeu essa junção de data frames e com as informações de todas as posições numa só variável seguimos a análise inicialmente com um *df.describe()*.

Uma imagem contendo texto

Descrição gerada automaticamente

A altura dos jogadores da NBA vai de 1,75m até o maior jogador, que tem uma altura de 2,20m. Essa diferença é observada por conta das posições, os “Guards” (menores em campo) são os jogadores que entram menos em contato com os adversários, e são os responsáveis por fazer a bola chegar no “Center”. Por sua vez, esses são os maiores em quadra e estão mais próximos da cesta, participando mais de disputas, bloqueios e roubos de bola na região do garrafão. Já os “Forwards”, desempenham tanto o papel dos centers quanto o dos “Guards”, dependendo da situação do jogo; por conta disso, possuem uma altura que fica na média, como observado no histograma acima.

Além disso, podemos constatar que o peso dos jogadores segue uma distribuição normal. Isso pode ser observado por meio da função “probplot”. Além disso, através de um histograma, essa evidência fica ainda mais clara.

Uma imagem contendo texto, mapa

Descrição gerada automaticamente

1. **Análise do Peso**

Analisando o peso percebemos que ele também é uma variável que se correlaciona com a posição:

Como constatado na análise da altura, o os menores pesos são correspondentes aos jogadores da posição “Guard”, seguido pelos “Forwards”, e por fim os mais pesados e mais altos são os “Centers”.

Uma imagem contendo captura de tela

Descrição gerada com alta confiança

Um fato interessante, é que a média do peso de todos os jogadores, é muito próxima da média do peso dos “Forwards”. Isso se dá, visto que os “Forwards” são versáteis, e considerados híbridos entre os “Guards” e os “Centers”.

Além disso, o peso dos jogadores também segue uma distribuição normal, e pode ser confirmado pelo gráfico da função “probplot”, que indica o quão próximo de uma distribuição Gaussiana, é a distribuição dos dados selecionados. Quanto mais sobreposto da reta vermelha, mais próximo será de uma Gaussiana.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

1. **Relacionando as Variáveis**

Ao utilizar a função “sns.pairplot” foram plotados todos os gráficos relacionando duas variáveis. Diante disso, muitos gráficos possuíam uma forte correlação entre as variáveis, porém não ajudariam a encontrar a posição de um jogador. São essas: tempo jogado, pontos marcados, número de bloqueios, número de assistências. Porém, apesar dessa boa correlação, decidimos não as utilizar, visto que quando plotadas com as categorias separadas, não houve nenhum padrão de distinção entre essas categorias, o que torna muito difícil de relacionar dados futuros dessas variáveis, com as três categorias. Com isso, ao aprofundar-nos ainda mais no estudo dos gráficos percebemos uma clara distinção entre as posições nos gráficos ligados à altura. Todavia, apesar da separação, a maioria não tinha correlações satisfatórias que pudessem auxiliar na determinação da posição. No entanto, um gráfico em especial chamou a atenção. Ao analisarmos o gráfico da Altura pelo Peso, percebemos uma significativa distinção entre as posições, além de um Coeficiente de Pearson bastante elevado (0,792). Com essas informações, torna-se possível relacionar a altura dos jogadores, com o peso, e por fim com a posição. Assim sendo, determinamos como variáveis principais para a determinação da posição de um jogador a altura e o peso.

Uma imagem contendo mapa, texto

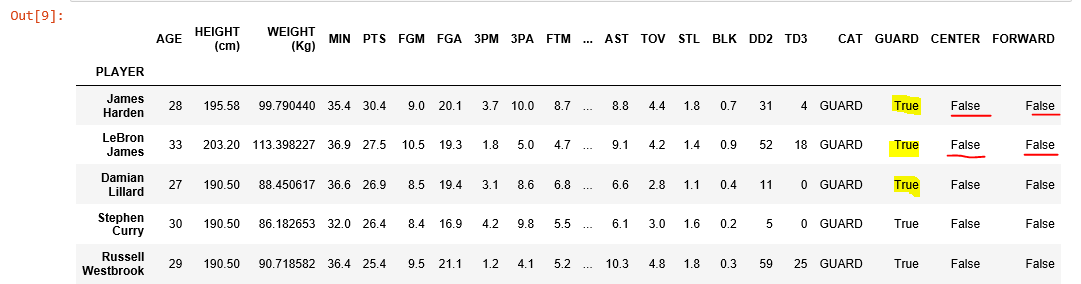
Descrição gerada automaticamente

Para fazermos a classificação e conseguirmos definir a posição de um dado jogador com base nas suas características recorreremos à uma Regressão Logística. Esse tipo de “regressão” tem como principal objetivo prever uma variável categoria (ou qualitativa), a partir de uma variável dependente binária, nesse caso, são elas: peso e altura. A Regressão Logística nos fornecerá sempre uma previsão entre 0 e 1, com isso conseguimos interpretar os resultados como uma probabilidade válida. Além disso, regressão logística não será influenciada por outliers, dados que não agregam muita informação para o modelo.

1. **Separando Treinamento e Teste**

Primeiramente usamos o modelo de Regressão Logística para prevermos se o jogador tem probabilidade de pertencer a alguma posição em específico, ou seja, com base no peso e altura de certo jogador previmos se ele seria ou não guard, analogamente fizemos o mesmo para as outras posições.

Contudo para tal primeiro acionamos outras três colunas ao DataFrame concatenado com todas as posições, a coluna GUARD recebe o valor 1 quando o jogador e 0 quando ele é de outra posição, o mesmo para as colunas CENTER e FORWARD**.**



Após tirarmos as conclusões acima, separamos uma amostra de teste e outra de treinamento. Isso serve para que a máquina aprenda treinando com uma amostra e aplique o aprendizado prevendo as possíveis posições para as outras amostras de teste.

Uma imagem contendo captura de tela

Descrição gerada com muito alta confiança

1. **Prevendo a Posição**

Com a separação, é possível aplicar o algoritmo em uma análise de testes. E com isso, obter uma certa acurácia em relação à uma base de dados já classificada**.**

* Posição Guard

Aqui vemos que o modelo categorizou com uma precisão de 76% os atletas na posição guard com base em seus pesos e alturas.

**Uma imagem contendo captura de tela

Descrição gerada com muito alta confiança**

* **Posição Center**

**Uma imagem contendo captura de tela

Descrição gerada com alta confiança**

Aqui vemos que o modelo categorizou com uma precisão de 84% os atletas na posição center com base em seus pesos e alturas.

* Posição Forward

**Uma imagem contendo captura de tela

Descrição gerada com alta confiança**

Aqui vemos que o modelo categorizou com uma precisão de 84% os atletas na posição center com base em seus pesos e alturas.

1. **Conclusão**

A partir da análise dos resultados, podemos concluir que o modelo é capaz de determinar a posição da maioria dos jogadores. O modelo final, é capaz de analisar a altura e o peso de um determinado jogador, e indicar qual a sua posição, em um time de basquete, com um acurácia de 62%. Diferentemente do modelo anterior, é possível prever a posição diretamente, sem ter uma posição analisada pré-definida. No modelo intermediário, era analisado se o jogador correspondia ou não à uma certa posição definida, e a resposta era dada como: sim para a posição, ou não, se o jogador não correspondesse àquela posição. Já no modelo final, não há posições pré-definidas, o modelo responde a posição de cada jogador de acordo com os dados obtidos, e não somente se o jogador corresponde ou não àquela posição especificada previamente.

Uma imagem contendo captura de tela

Descrição gerada automaticamenteAlgoritmo do Modelo Final: