Erick Silvestre Lima de Brito;
Germano Andrade Brandão;
João Alcindo Ribeiro de Azevedo;
Patrick Saul Costa do Amaral;
Sávio Vinícius Costa do Amaral;

Relatório - Trabalho A2

Rio de Janeiro
2020

Erick Silvestre Lima de Brito;
Germano Andrade Brandão;
João Alcindo Ribeiro de Azevedo;
Patrick Saul Costa do Amaral;
Sávio Vinícius Costa do Amaral;

Relatório - Trabalho A2

Relatório referente ao trabalho final da disciplina Linguagens de Programação - 2020.2 da Graduação em Matemática Aplicada e da Graduação em Ciência de Dados na Escola de Matemática Aplicada - Fundação Getúlio Vargas.

Professor:	Dr. Rafael de Pinho André
Monitora:	Bianca Gotaski de Melo
Monitor:	Igor Cortes Junqueira
Monitor:	Rener de Souza Oliveira

Rio de Janeiro
2020

Sumário

1 Introdução

1.1 Escolha das Bases

Decidimos escolher as tabelas real_state.real_state_values e fifa.fifa_players disporníveis no banco de dados. a tabela real_state refere-se a valores residenciais nos subúrbios de Bostos(Massachusetts,EUA). Já a tabela fifa_players é uma base detalhada com os dados dos jogodadores no jogo FIFA 19(2018), desenvolvido pela empresa Eletronic Arts Sports.

1.2 Definição de papéis

Os papéis foram decididos da seguinte maneira:

Papel 1: Cientista de Dados / Especialista de Negócio

- Sávio Vinícius
- Patrick Saul[1]

Papel 2: Engenheiro de Dados / Engenheiro de Software

- João Alcindo

Papel 3: Especialista em Visualização de Dados

- Germano Andrade

Papel 4: Especialista de Garantia da Qualidade:

- Erick Brito
- Patrick Saul

1.3 Repositório

Todo o trabalho de limpeza e manipulação feito (Códigos, documentação, galeria de imagens, etc) pode ser encontrado no Repositório do Trabalho no GitHub.

[1]Devido à quantidade de papéis divergir da quantidade de integrantes do grupo, este aluno "flutuou" entre o Papel 1 e o Papel 4.

2 Perguntas de Negócio

- Sobre a Base do Fifa 19

    1. Dado um empresário com qualquer orçamento, qual o melhor elenco que ele pode ter, visando ganhar vários campeonatos? E qual seria o custo para ter esse elenco?
    2. Dado um time com qualquer orçamento, qual o melhor elenco que ele pode ter visando ter o melhor retorno com as vendas dos jogadores no longo prazo? E qual seria o custo para ter esse elenco?
    3. Dado o conjunto dos 50 melhores jogadores avaliados, qual a porcentagem deles que preferem chutar com o pé esquerdo?

- Sobre a Base do Real State Values

    1. Se uma pessoa prioriza acima de tudo a segurança, qual seria a distância ponderada para os cinco centros de empregos de Boston?
    2. Se uma pessoa acredita que uma taxa de aluno-professor não possa ser maior que 15 alunos por professor, qual seria o menor índice de criminalidade que ele conseguiria?
    3. Se uma pessoa quer que sua área seja limitada por um rio, qual seria a maior e a menor quantidade média das casas ocupadas?

3 Diagrama de Soluções

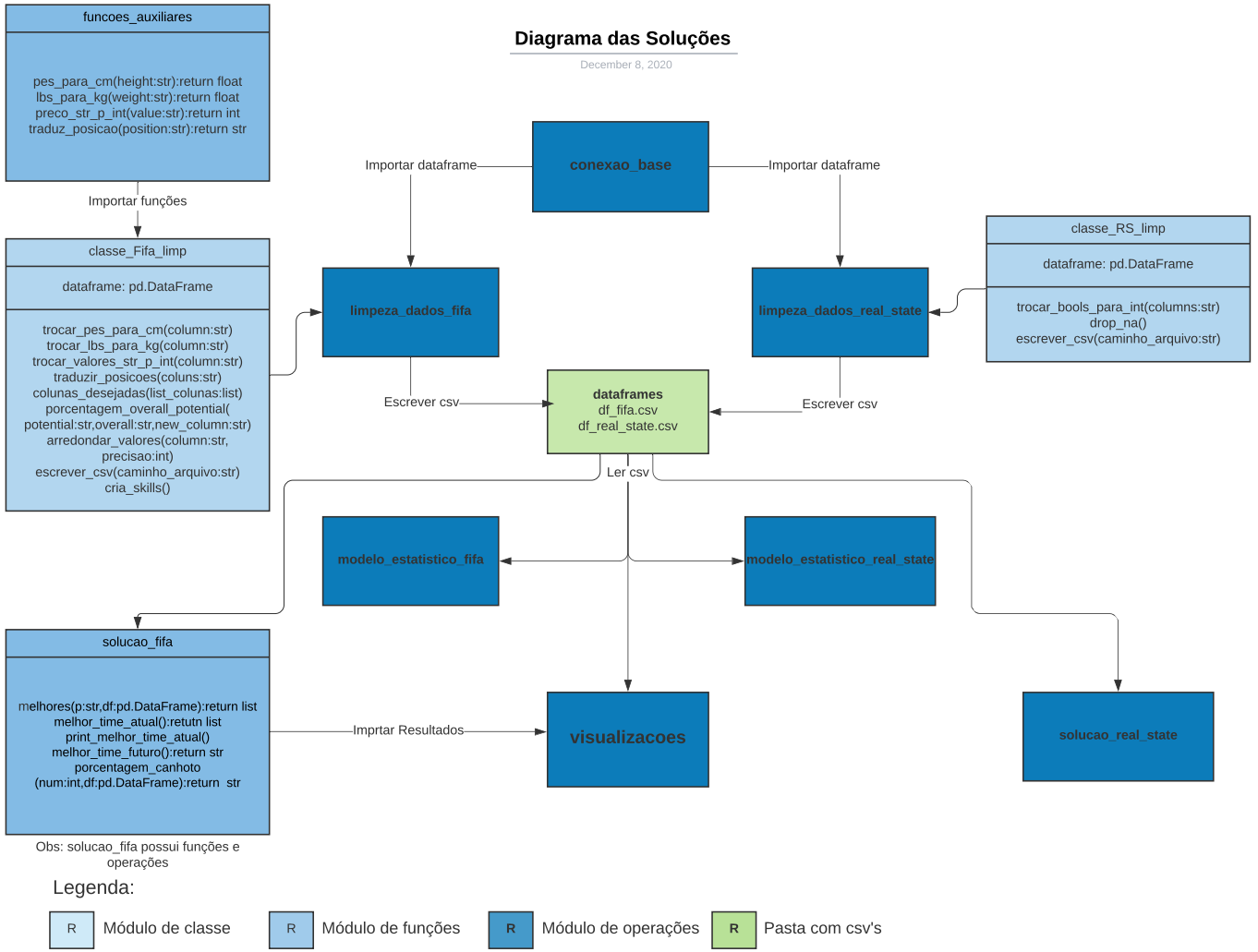


Figura 1: Diagrama de soluções

4 Modelos Estatísticos

4.1 Fifa 19

Começamos entendendo os dados da base de dados por meio dos códigos *df.head()* , *df.columns()* e *df.describe()*. Depois de fazer as análises descritivas, exploratória do modelo, começamos a limpar os dados (trocar várias unidades de medidas, trocar os type de string para int, entre outras coisas) para assim começar a fazer os modelos estatísticos dessa base de dado.

Para a construção do modelo estatístico, olhamos que o valor de contrato de um jogador parece está relacionado com o valor da quebra de seu contrato(multa rescisória) , então fizemos o modelo de regressão linear para vê se eles tinham alguma correlação e chegamos em um impressionante R^2 de $0,961$, como mostra a imagem abaixo.

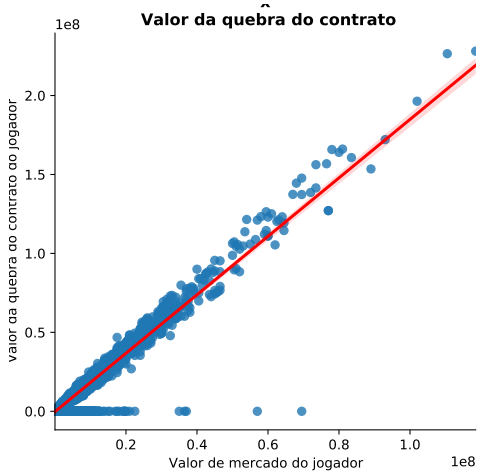
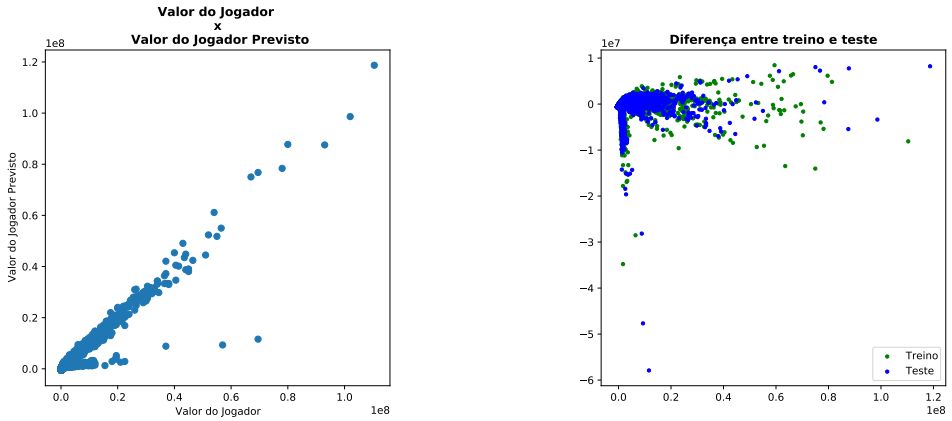


Figura 2: Valor de quebra de contrato vs Valor de mercado do jogador

Por fim, dividimos o conjunto de dados em treino e teste para fazer uma previsão dos valores dos jogadores com os valores estimados e também chegamos a um resultado bom com um R^2 de $0,961$ como mostra a figura abaixo:



(a) Valor Real vs Valor Previsto (b) Diferença entre treino e teste

Figura 3: FIFA 19

4.2 Real State Values

Começamos fazendo uma análise descritiva e exploratória dos dados, e assim começamos a pensar que o índice de indústrias não varejistas($INDUS$) tinha uma correspondencia com o índice da concetração do nitrato de oxigênio(NOX). Com

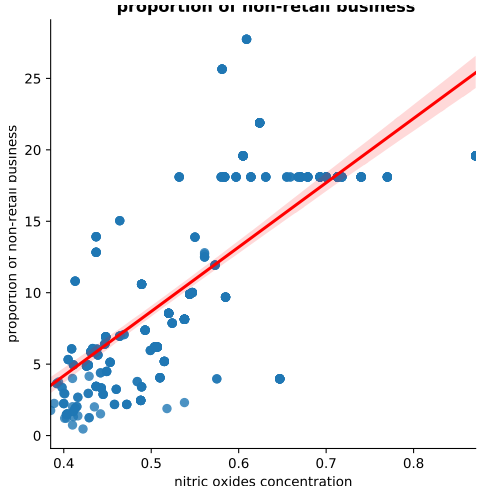
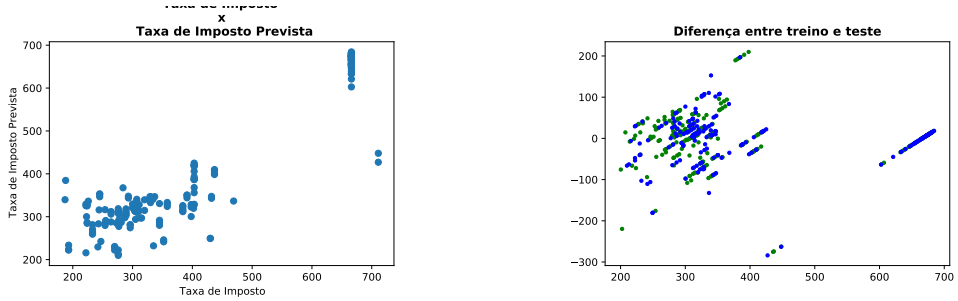


Figura 4: NOX vs INDUS

base nisso começamos a limpar os dados(trocar type de algumas colunas) e depois disso começamos a fazer o método estatístico usando regressão linear para vê se o NOX tem alguma correlação com o $INDUX$ e chegamos a um R^2 de $0,582$ como mostra a figura abaixo.



(a) Taxa Real vs Taxa Prevista (b) Diferença entre treino e teste

Figura 5: Real State Values

Por fim dividimos o conjunto de dados em treinos e testes para fazermos uma análise das taxas e as taxas prevista e encontramos um R^2 de $0,886$ como mostra a figura acima.

5 Visualizações

5.1 FIFA

A partir das muitas colunas do dataframe, visando a diminuir a quantidade de dados para entender melhor os jogadores/times, criamos colunas com os atributos técnicos dos jogadores de linha e golerio, conforme as tabelas a seguir.

Os seis principais atributos técnicos dos jogadores de linha.

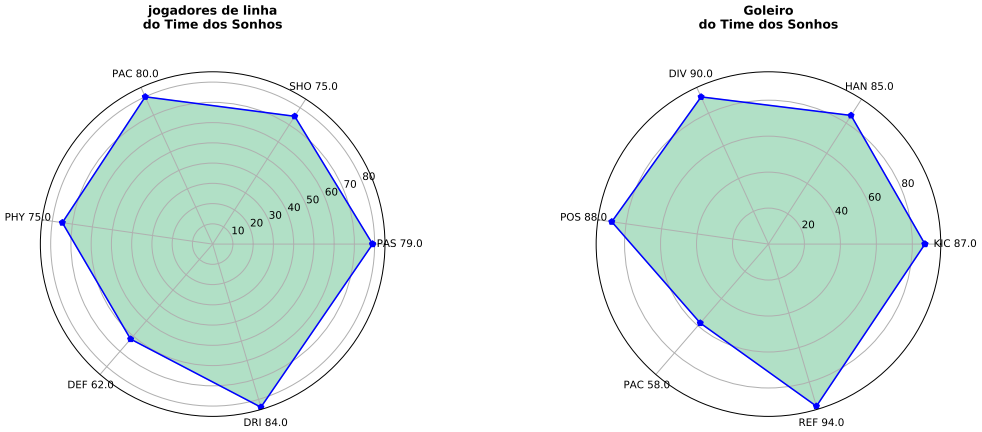
Sigla	Significado	Tradução
PAS	Passing	Passe
SHO	Shooting	Remate
PAC	Pace	Velocidade/ Ritmo
PHY	Physical	Físico/Resistência
DEF	Defense	Defesa
DRI	Dribbling	Drible/Finta

Os seis principais atributos técnicos dos guarda-redes / goleiros.

Sigla	Significado	Tradução
KIC	Kicking	Pontapé/Chutão
HAN	Handling	Manuseio
DIV	Diving	Mergulho
POS	Position	Posicionamento
SPD[2]	Speed	Velocidade
REF	Reflexes	Reflexos

[2]Equivalente ao PAC dos jogadores de linha

Feito isso, de acordo com a solução encontrada para as perguntas da base do FIFA(2), para um time baseado no *Overall* atual dos jogadores, tivemos o seguinte

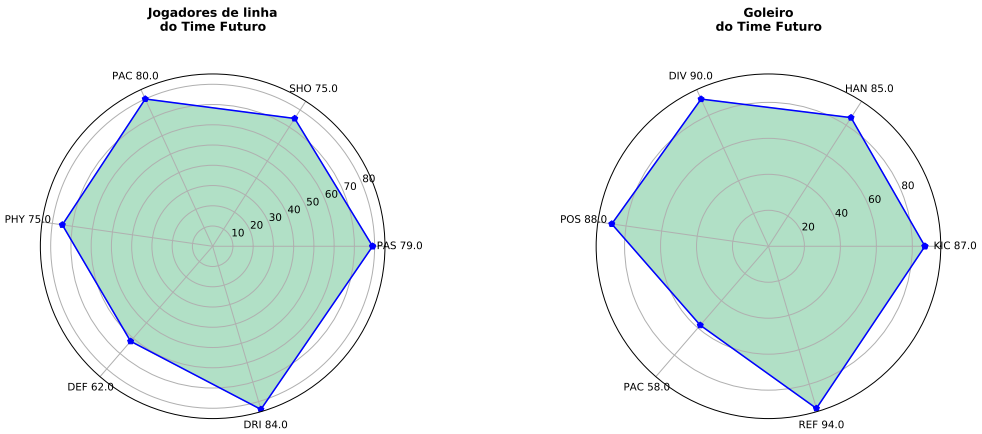


(a) Jogadores de linha (b) Goleiro

Figura 6: Time dos Sonhos

Para o gráfico (a) foi feito a média dos *skills* dos jogadores de linha. Já para o gráfico (b), foi feito o gráfico do melhor goleiro.

Já pensando em longo prazo, o time baseado no *Potential* dos jogadores, teríamos as seguintes pontuações dos *Skills*:



(a) Jogadores de linha (b) Goleiro

Figura 7: Time Futuro

Um ponto interessante a destacar é a diferença de idade entre esses times. Isso, porque é normal que o time visando no longo prazo seja composto em sua maioria por jogadores mais jovens, e isso é o que podemos ver no seguinte gráfico: Para

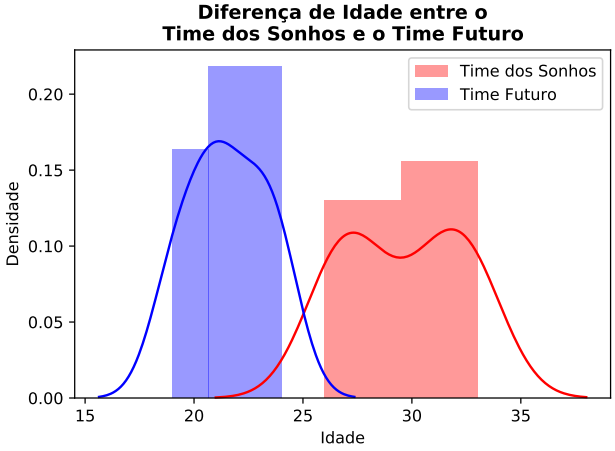


Figura 8: Diferença de idades

o gráfico (a) foi feito a média dos *skills* dos jogadores de linha, assim como nos gráficos anteriores. Já para o gráfico (b), foi feito o gráfico do melhor goleiro.

Podemos observar claramente que o "Time para o futuro" tem uma média de idade muito inferior ao "Time para o presente".

Outra análise interessante entre os dois times é no quesito de preço (valor total que o time custa).

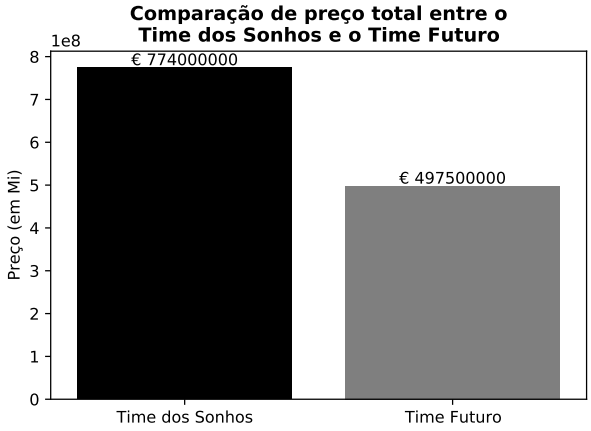


Figura 9: Comparação de preços

Podemos observar que o investimento em um time para o futuro compensa muito mais, visto que a diferença entre o preço dos times é gigante.

Agora, em relação ao **Pé Dominante** dos jogadores (conforme a 3ª pergunta sobre essa base), podemos ver a diferença em percentual entre *Canhotos* e *Destros*, dado o conjunto dos 50 jogadores com maior *Overall*.

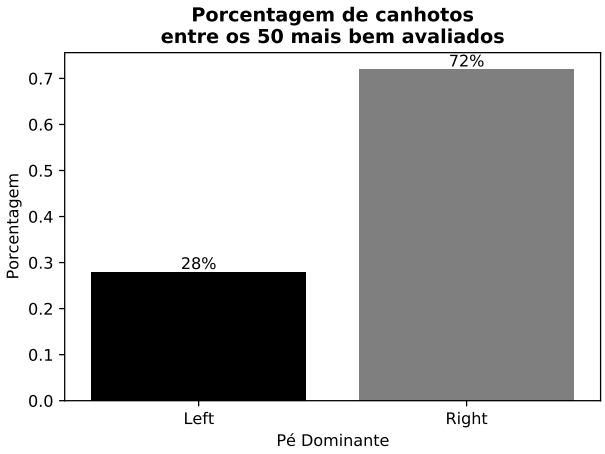


Figura 10: Porcentagem de canhotos

5.2 Visualizações Real State

Para o conjunto de dados do Real State Values (2), uma relação interessante foi entre a Distância entre os grandes centros de empregos de Boston e a Taxa de Crimes.

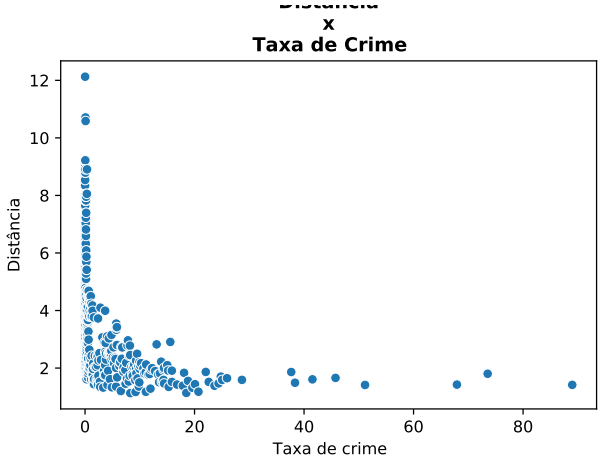


Figura 11: Distância vs Taxa de Crime

Podemos observar que ao passo que a distância diminui, as taxas de crimes são maiores.