PREDIÇÃO DA POSIÇÃO DO RANKING DE UMA CIDADE NA LISTA DAS TOP-K MELHORES CIDADES PARA SE VIVER

Cibele Castelo Nogueira
Pontifícia Universidade Católica de Minas Gerais (PUC Minas- Contagem) - Brasil

RESUMO
Este artigo tem como objetivo realizar a predição de ranking de cidades na lista das *top-k* melhores cidades para se viver, utilizando dados da competição Movehub City Rankings, que se encontram no site da *Kaggle*. Apresenta-se a utilização da biblioteca *scikit-learn* para implementação de um algoritmo de *Machine Learning* para resolução do problema.

PALAVRAS-CHAVE: Movehub, Predição, *Machine Learning*, *Scikit-learn*.

1.INTRODUÇÃO

A predição do ranking das melhores cidades para se viver com base em dados conhecidos é um problema alcançável, pois é constatável que melhores cidades para se viver possuem em sua localidade pessoas com boa qualidade de vida, com bons salários e com poder de compra, e em contraste cidades menos favorecidas possuem altas taxas de crime, pobreza e poluição. Assim, cada fator citado, entre outros influenciam em como uma cidade pode ser avaliada.

Nesta competição já há uma variável, 'Movehub Rating', que é a combinação de todas as pontuações para uma classificação geral para uma cidade. Dessa forma, para realizar a predição será necessário considerar todos os fatores existentes dentro do conjunto de dados, pois eles contribuem para a obtenção do resultado do *ranking*. O *Scikit-Learn*, uma biblioteca de *Machine Learning*, será utilizada para implementação de um algoritmo de regressão multivariada para a modelagem dos dados. Esse algoritmo irá criar o modelo que será capaz de prever os valores dos *rankings* das cidades.

Este artigo tem como objetivo resolver a competição passando por todas as etapas do ciclo de ciência de dados, sendo assim, será visto nas seções posteriores: a coleta de dados, o processamento de dados, a análise exploratória, a modelagem dos dados, a avaliação do modelo e por fim, uma conclusão final que avaliará se o modelo construído atende ou não a necessidade do problema de negócio.

2.1 COLETA DE DADOS

Inicialmente, antes de qualquer análise, é necessário obter os dados que serão utilizados para o projeto. Os conjuntos de dados foram baixados do site da *Kaggle*, mais especificadamente da competição Movehub City Rankings. Para resolução do problema, será utilizada a variável dependente 'Movehub Rating', que é uma combinação de todas as pontuações para uma classificação geral para uma cidade ou país. As outras variáveis, são independentes, e podem ter influência direta ou indireta no valor do *ranking*.

2.3 PROCESSAMENTO DE DADOS

O Processamento de dados tem como objetivo realizar a limpeza e correção dos dados. Esta etapa visa a qualidade dos dados e garante que a base de dados contenha dados precisos e confiáveis. Sua importância é muito grande, pois senão for feita, as próximas etapas podem ser prejudicadas pois conterão dados sujos, duplicados e valores faltantes que irão impactar diretamente na avaliação final do modelo e consequentemente na consistência da informação.
Para isso, utilizou-se a biblioteca Pandas que fornece suporte para estruturação e análise de dados. Ela possui familiaridade com a linguagem de consulta estruturada SQL e permite fazer operações DML e DLL dentro de sua própria sintaxe.

Realizou-se análise, extração e transformação dos três conjuntos de dados que se encontravam no formato '.csv'. As seguintes tarefas foram executadas: junção de todos os *datasets*; atualização dos nomes de cidades com erros, inserção dos nomes de países faltantes para cada cidade, retirada de duplicação de cidades, sendo que cidades que possuíam mais países como não há como saber qual cidade é de qual país, permaneceu-se mais registros, como no caso Cambridge e Valência. E também, alteração do nome das colunas, pois não foi permitido a manipulação dos dados em variáveis que possuíam espaços como: 'Avg Rent', para isso foi feita a Substituição do espaço entre os títulos das colunas por '_', ficando todas as colunas que tinha espaço assim: 'Avg_Rent'.

Assim sendo, a manipulação de dados permitiu obter informações dos conjuntos de dados, possibilitando verificar, selecionar, atualizar, inserir e remover campos que haviam dados sujos, valores duplicados e faltantes.

Abaixo, a visualização do gráfico com a visão de registros duplicados e o gráfico após a remoção dos registros duplicados.

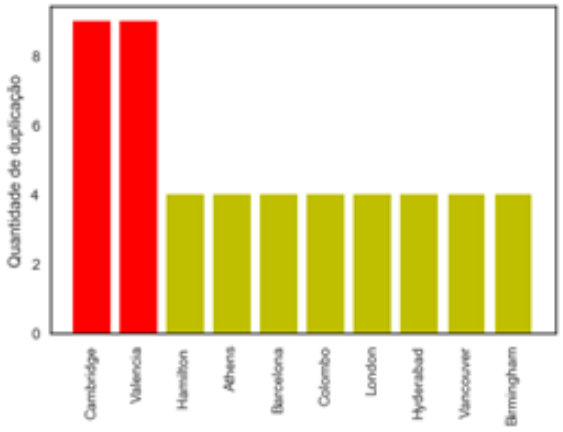

Figura 1: Registro de cidades duplicadas

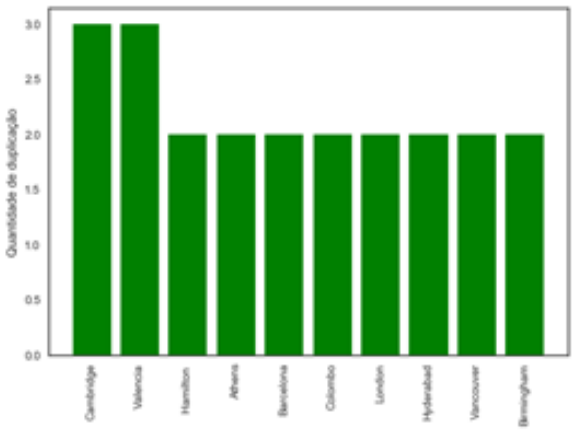

Figura 2: Após remoção de registro de cidades duplicadas

2.4 ANÁLISE EXPLORATÓRIA

A análise exploratória é a etapa que permite você conhecer seus dados. Para análise de regressão multivariada, faz-se necessário a exploração da associação entre as variáveis, no caso, de cada variável independente numérica com a variável dependente numérica ('Movehub Rating'). Essa análise permite entender como a variável independente se comporta em relação a variável

dependente. Para obter numericamente essa associação entre as variáveis, foi realizado o cálculo do coeficiente de correlação linear Pearson, utilizando para isso a função pearsonr da biblioteca Scipy.

No gráfico abaixo, pode-se visualizar o resultado obtido da relação de cada variável independente (x) com a variável dependente (y). As cores das barras são diferentes, pois abordam diferentes grupos que contêm variáveis de acordo com seu valor de correlação linear. Assim, cada variável independente pertence a um grupo:

O 1º grupo possui as barras na cor verde e são aqueles que apresentam valor Pearson próximo a 1. Quanto mais próximo de 1, maior correlação positiva, ou seja, essas variáveis tendem a aumentar o valor do 'Movehub Rating'.

O 2º grupo possui barras na cor amarela e são aqueles que apresentam valor Pearson próximo a 0. Quanto mais próximo de 0, menor a correlação linear.

O 3º grupo possui barras na cor vermelha e são aqueles que possuem valor Pearson próximo a -1. Quanto mais próximo de -1, maior correlação negativa, ou seja, essas variáveis tendem a diminuir o valor do 'Movehub Rating'.

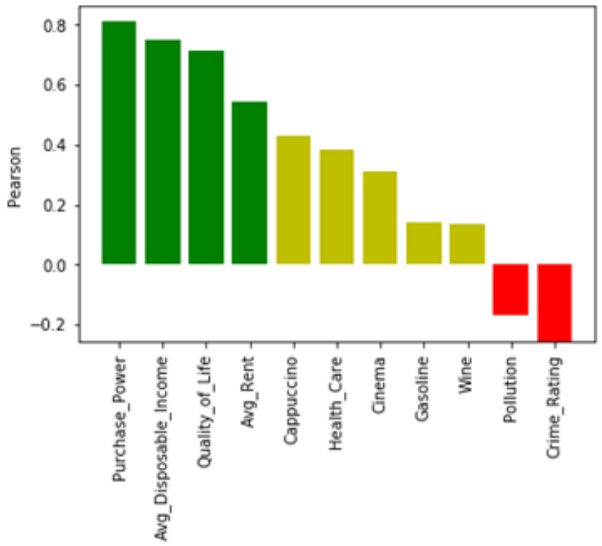

Figura 3: Coeficientes de correlação linear para cada variável independente

2.4.1 GRÁFICOS DE DISPERSÃO

Os gráficos de dispersão permitem visualizar a tendência, a força e a forma de cada ponto ou conjunto de pontos distribuídos no gráfico. Cada ponto representa uma observação, e a localização desse ponto depende do valor das duas variáveis x e y. Este gráfico permite compreender como cada variável independente afeta o aumento ou a diminuição da variável dependente, e possibilita também verificar anomalias, outliers e conjuntos de pontos que não seguem o padrão. Para a criação dos gráficos foi utilizada a biblioteca Matplotlib.

Abaixo, apresenta-se os gráficos de dispersão separados pelos grupos definidos anteriormente.

Grupo 1: Variáveis que apresentam correlação linear positiva:

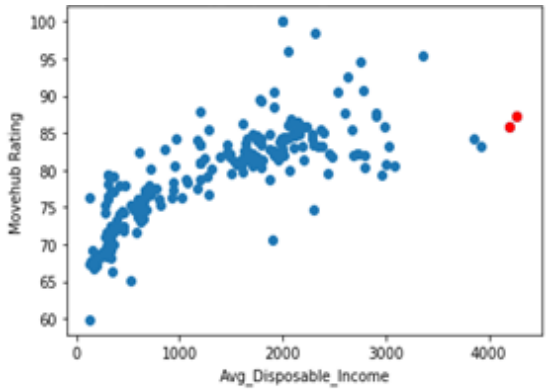

Figura 4: Relação entre Avg_Disposable_Income X Movehub_Rating

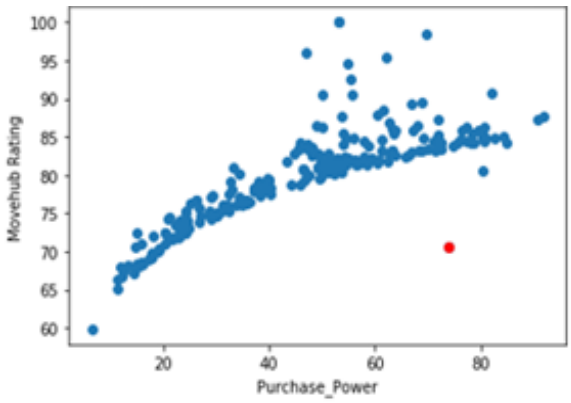

Figura 5: Relação entre Purchase_Power X Movehub_Rating

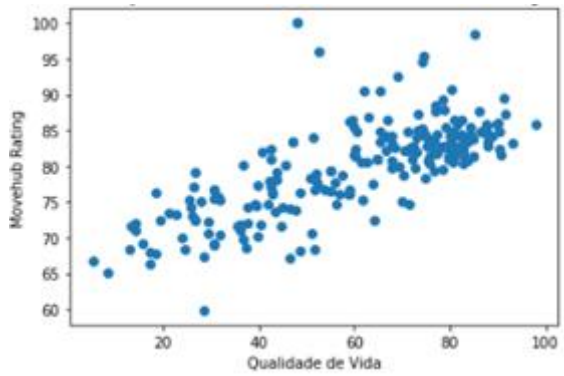

Figura 6: Relação entre Qualidade de Vida X Movehub_Rating

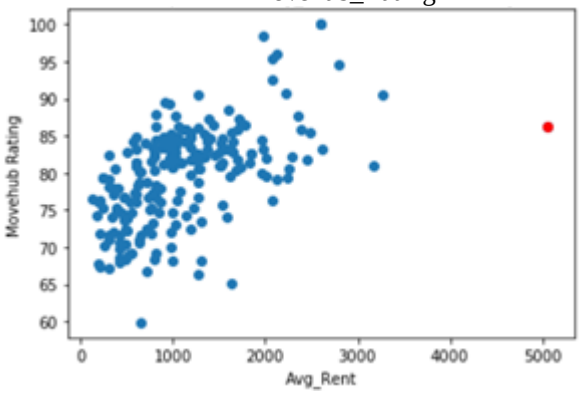

Figura 7: Relação entre Avg_Rent X Movehub_Rating

Pode-se observar que a tendência de 'Movehub Rating' é aumentar gradativamente ao aumento da variável independente, como no caso de Qualidade de vida, que se constata que: quanto melhor o nível de qualidade de vida melhor é a cidade para se viver.

Grupo 2: Variáveis que apresentam pouca correlação linear

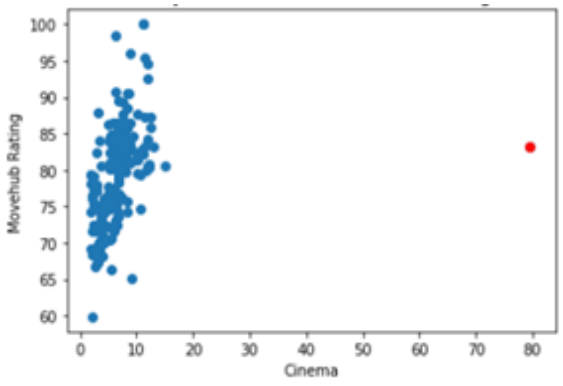

Figura 8: Relação entre Cinema X Movehub_Rating

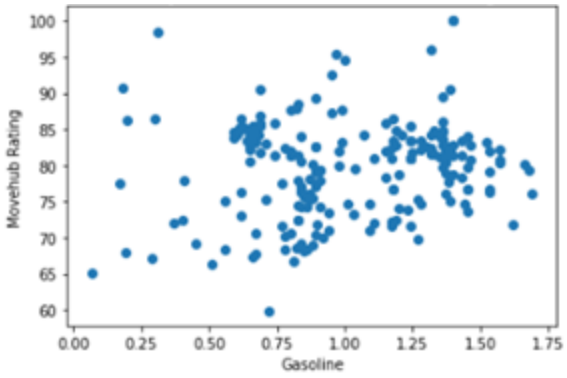

Figura 9: Relação entre Gasoline X Movehub_Rating

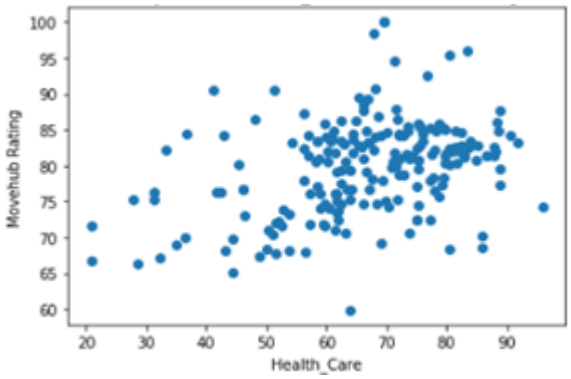

Figura 10: Relação entre Health_Care X Movehub_Rating

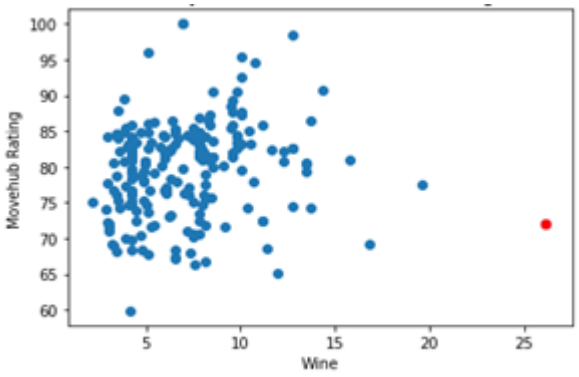

Figura 11: Relação entre Wine X Movehub_Rating

Grupo 3: Variáveis que apresentam correlação linear negativa:

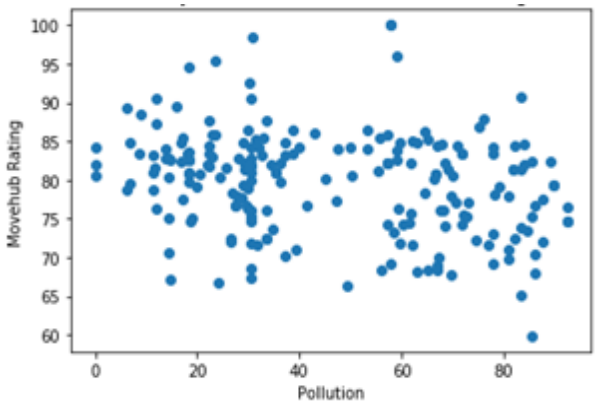

Figura 12: Relação entre Health_Care X Movehub_Rating

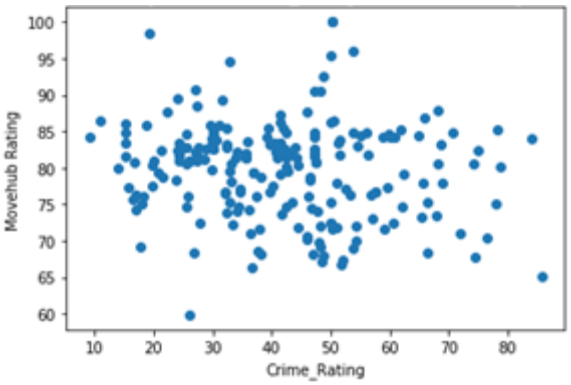

Figura 13: Relação entre Wine X Movehub_Rating

Pode-se observar que a tendência de 'Movehub Rating' é diminuir gradativamente ao aumento da variável independente. O aumento no valor de poluição e na taxa de crime cometidos, tendem a desvalorizar uma cidade, então consequentemente sua avaliação diminui.

2.4.2 REMOÇÃO DOS OUTLIERS

A visualização dos *Scatter Plots* permitiu a detecção da presença de *outliers*, como por exemplo, no caso da variável 'cinema' que possui um ponto de dado com o valor maior que 70. Este valor será considerado *outlier* pelo motivo de não se encontrar dentro do padrão do conjunto de pontos de dados. No gráfico abaixo, percebe-se a diferença ao remover *outliers* desse tipo, os pontos de dados se espalharam mais, tornando possível uma correlação linear positiva, como é confirmado ao calcular o coeficiente de Pearson. Conclui-se assim que, a detecção e remoção de *outliers* é essencial para análise da correlação linear entre variáveis numéricas.

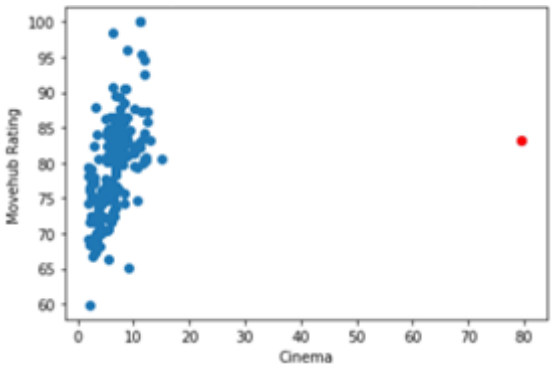

Figura 14: Relação entre Cinema X Movehub_Rating

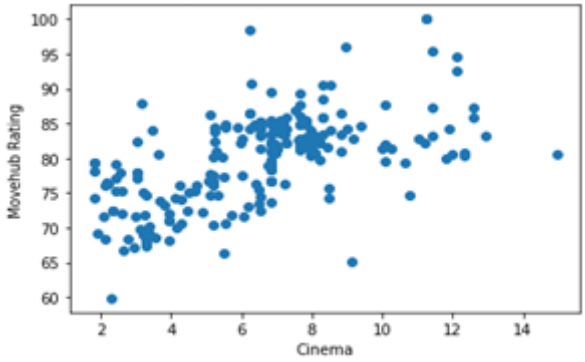

Figura 15: Relação entre Cinema X Movehub_Rating

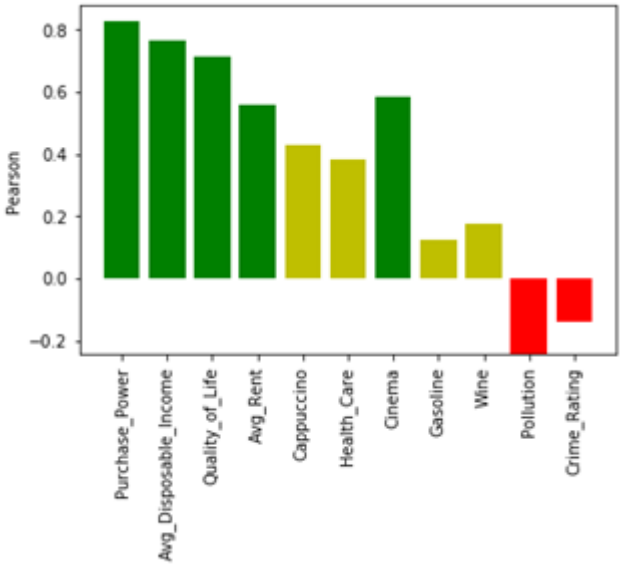

Figura 16: Resultado do Coeficiente de correlação linear após remoção de outliers

A partir dessa análise de *outliers*, foi efetuada a remoção de todos os pontos de dados considerados fora do padrão.

2.5 MODELAGEM DOS DADOS

A modelagem dos dados será feita com a utilização do Scikit-learn, biblioteca de *Machine Learning*, e o algoritmo escolhido foi LinearRegression(). Para esse modelo foi utilizada todas as variáveis, pois a partir de testes percebeu-se que a avaliação do modelo diminuía sem a presença da variável retirada.

O modelo de dados foi construído com 80% por cento de dados de treinamento e 20% por cento de dados para teste. Divide-se o *dataset* em conjunto de dados de treino para construir o modelo, e o outro conjunto de dados de teste para avaliar se o modelo terá um bom desempenho com novos dados (*que não foram vistos antes). Desse modo, temos o benefício de se ter dados de teste para avaliação do desempenho do modelo para evitar, por exemplo, o problema de *overfitting* dos dados. Se não houver esta divisão dos *datasets*, o modelo pode não ter precisão e passar a ser inadequado para a determinada solução, pois não seria capaz de prever novos dados, e por isso não iria atender ao problema de negócio.

2.6 AVALIAÇÃO DO MODELO

A avaliação do modelo é importante para saber se o modelo irá generalizar para dados não vistos ainda. Para realização dessa etapa, será utilizada algumas métricas de avaliação de regressão do Scikit-Learn. São elas:

2.6.1) COEFICIENTE DE DETERMINAÇÃO R^2

O coeficiente de determinação R^2 foi utilizado para quantificar o desempenho do modelo, ou seja, mostrar o quão bom é o modelo em fazer predições. Os valores variam de 0 até 1, quanto mais próximo de 1 melhor será a predição do valor da variável dependente.

O resultado obtido para o conjunto de treinamento foi de: 0.752 e do conjunto de teste foi de 0.902.

2.6.2) O PERCENTUAL DE VARIÂNCIA EXPLICADA

O percentual de variância explicada é o percentual de variação total explicada pela melhor partição para cada grupo. Quando mais próximo de 1, melhor.

O resultado obtido foi de 0.917.

2.6.3) ERRO MÉDIO ABSOLUTO

O erro médio absoluto é a diferença entre o valor previsto e o valor real.

O resultado obtido foi de 1.62

2.6.4) GRÁFICO VALORES REAIS X VALORES ESTIMADOS

Com o intuito de facilitar a grandeza de previsibilidade do modelo, gerou-se um gráfico com os valores estimados, que são aqueles do conjunto de teste e os valores reais, com o objetivo de demonstrar que os valores estimados se encontram próximos dos reais, permitindo assim a mesma avaliação no ranking.

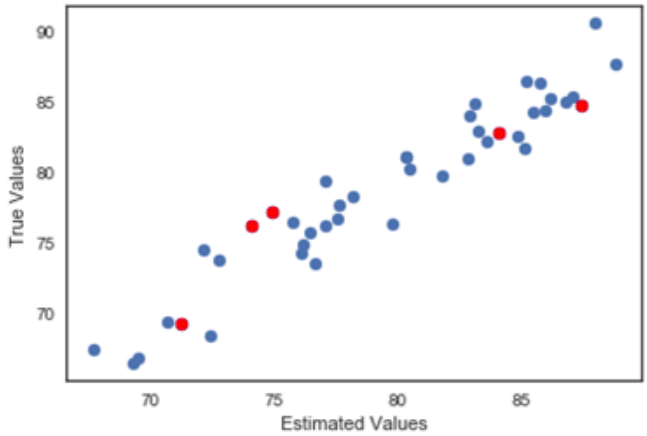

Figura 17: Gráfico que mostra a proximidade

entre os valores estimados e reais.

3.CONCLUSÃO

Pretendeu-se com esse artigo apresentar um modelo preditivo para a competição Movehub City Ranking, por meio da estruturação de um projeto de Ciência de Dados, com o intuito de aprofundamento no entendimento de cada etapa para entendimento do problema proposto.

Machine Learning foi utilizada para a análise de dados e automatização do desenvolvimento da modelagem de dados. O algoritmo de utilizado, possibilitou a obtenção de um resultado de 90,2% de precisão, o que demostra a eficácia do modelo na generalização de dados ainda não vistos.

Conclui-se assim que, modelo preditivo construído atingiu o seu propósito de realizar a predição do *ranking* das melhores cidades para se viver.

BIBLIOGRAFIA CONSULTADA

Guido, S. and Müller, A. *Introduction to Machine Learning with Python*. United States of America, O'Reilly Media, pp.376, 2017.

McKinney, Wes. *Python for Data Analysis.* United States of America, O'Reilly Media, pp.451, 2013.