**INSPER**

Amanda Rufino

André Dambry

Giovanna Cabral

**RELATÓRIO PROJETO FINAL DE CIÊNCIA DOS DADOS**

São Paulo

17 de novembro de 2018

**Introdução**

Neste trabalho de Ciência dos Dados iremos predizer, a partir dos métodos Linear Regression e Random Forest Regression, o ‘Happiness Score’ de países. Esse, por sua vez, foi realizado a partir de perguntas com tópicos que caracterizam o nível de felicidade. Assim, para cada pergunta os entrevistados responderam com um número na escala de 0 a 10, com 0 significando que a pessoa é infeliz no assunto e com 10 que é muito feliz. Com essas respostas, que serviram como fonte de pesquisa para o trabalho original, analisou-se quanto por cento cada variável contribuiu para o valor final do “Happiness Score” do país.

O objetivo do projeto é predizer esse mesmo valor de “Happiness Score” a partir de variáveis contínuas de cada país. Assim, a partir de dados como GDP per capita e expectativa de vida, iremos comparar os resultados obtidos pela regressão feita, com o valor experimental do dataset original. Como consequência, será verificado se as entrevistas realizadas com um público restrito se aproximam com dados reais medidos.

O projeto consiste em um problema supervisionado de regressão. É supervisionado, porque temos tanto os dados relacionados com o nível de felicidade, quanto o próprio “Happiness Score” que queremos predizer. Além disso, essa é uma tarefa de regressão porque o valor que queremos estimar é contínuo.

Para descobrir outras variáveis para o cálculo do modelo serão recolhidos dados do Gapminder e World bank. Para verificar se a variável é relevante para ser utilizada, a correlação dela com os dados do “Happiness Score” será feita. Assim, será possível encontrar um modelo de alta confiabilidade.

**Roteiro**

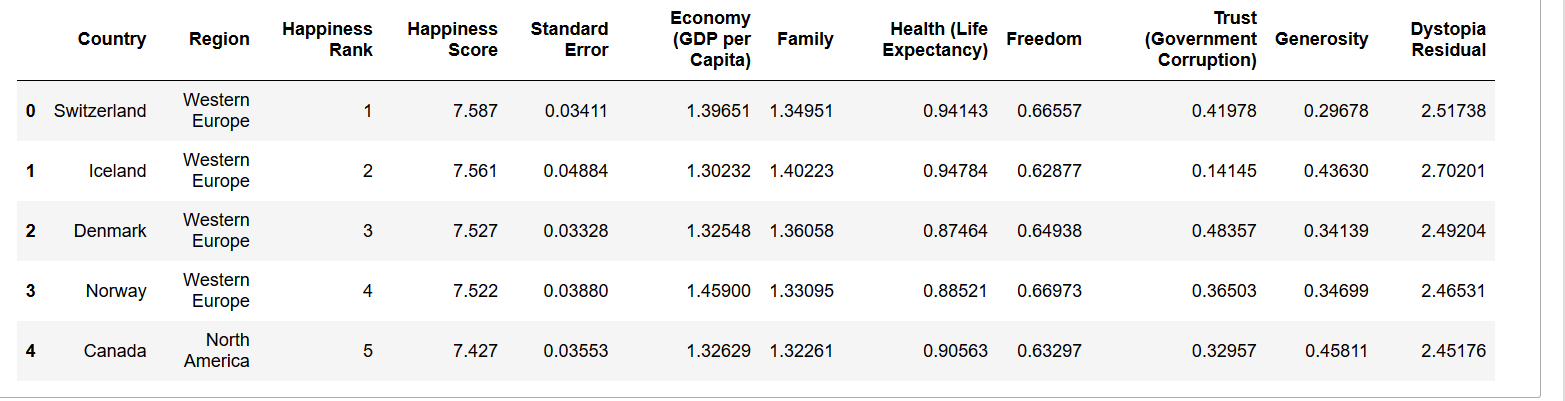
Para realizar projetos de machine learning, quando se tem um problema e modelo pensados, é necessário ter passos básicos definidos. Assim, para o nosso projeto utilizaremos a sequência abaixo:

1. Definir uma pergunta e determinar dados necessários.
2. Adquirir e analisar os dados
3. Identificar e corrigir anomalias se necessário
4. Estabelecer um modelo inicial
5. Treinar o modelo nos dados de treinamento
6. Fazer predições nos dados teste
7. Comparar as predições com dados já conhecidos e ver a performance pelo R quadrado
8. Se a performance não for boa, ajustar o modelo.
9. Interpretar o modelo, ou seja, explicar as metodologias

O primeiro passo já foi feito, a pergunta a ser analisada será: “É possível predizer o “Happiness Score” de países, com base em dados quantitativos?”.

**Análise do Dataset experimental**

No dataset experimental do Happiness Score, de 2015, algumas variáveis foram consideradas para o valor dessa pontuação, como é possível observar abaixo (imagem 1).



*Imagem 1 – dataset experimental*

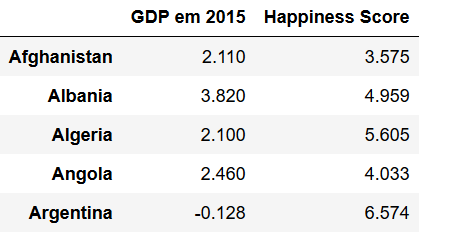
Como o dataset calcula a influência do valor da variável, os dados presentes na tabela não são os valores reais da variável, mas sim sua relação com o Score. Assim, optou-se por pegar os dados reais quantitativos do GapMinder. Como observado na imagem acima, que contém o dataset original, as únicas variáveis contínuas são “GDP per capita” e “Expectativa de vida”, por isso foram nossas primeiras escolhas de variáveis para o modelo.

Para escolher as demais variáveis utilizadas no modelo pegamos dados que tem relação com os temas do dataset experimental. Ou seja, foram escolhidos dados contínuos muito parecidos com o que foi perguntado na pesquisa.

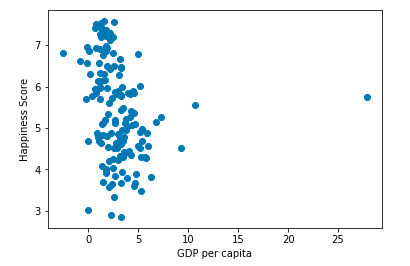
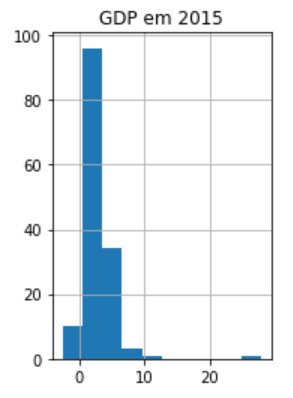
Como conseguinte, montamos o nosso dataframe com os dados de 2015, para os países presentes no dataset original, mantendo a coluna de “Happiness Score”, para poder comparar os dados do modelo com os dados reais.

**Análise dos dados - GDP**

Primeiramente analisamos GDP per capita, como verificado no dataframe abaixo, em que o GDP foi colocado junto com “Happiness Score”

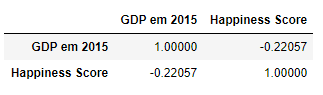


Como os dados foram obtidos no Gapminder, eles foram filtrados para apenas 2015 e para os países correspondentes com o “Happiness Score”. Assim, realizou-se a análise dessa variável.



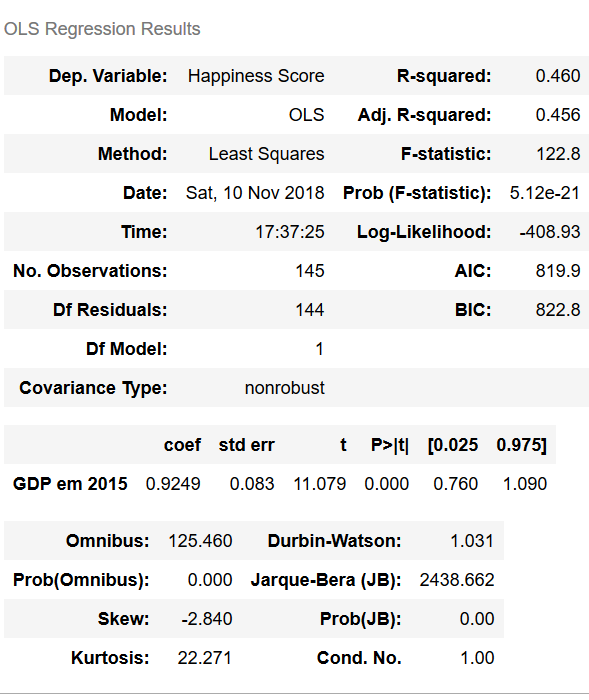
*Imagem 2 – Gráfico de dispersão Imagem 3 – Histograma GDP*

Observamos ao plotar o gráfico de dispersão da variável (imagem 2), que há uma correlação negativa entre os dados. No entanto, essa não é facilmente observada por causa da presença de outliers, os quais comprometem a visualização dos pontos. Além disso, era esperado que o histograma da variável se comportasse como uma distribuição normal, o que não ocorre devido à extensão do gráfico na cauda direita, o que também pode ter sido ocasionado pela presença do outlier.



*Imagem 4 – Correlação entre GDP e Happiness Score*

A correlação entre as variáveis foi de -0.22 (imagem 4). Com isso, montamos a regressão entre as variáveis e obtivemos o seguinte relatório:

‘

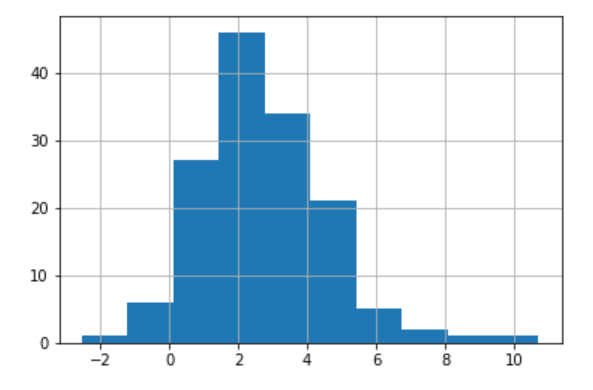
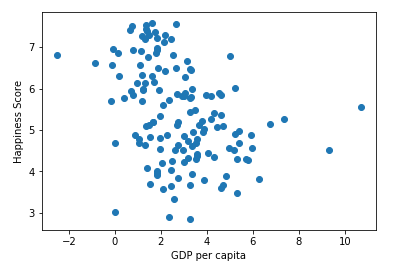
Ao analisar os dados estatísticos, observamos que:

* como o r-square indica o quão o modelo é bom para prever algo a partir dele, o nosso tem 46% de probabilidade de prever algo.
* O p-value do modelo é menor que o α de 5%, indicando que as variáveis x e y são realmente correlacionadas.
* A estatística de teste f indica que pelo menos um dos betas é não nulo e, assim, o modelo é estatisticamente significativo.
* O teste Omnibus indica a baixa probabilidade de os resíduos serem uma normal, propondo como uma iteração a ampliação do tipo de modelo, já que este se comporta para distribuições normais.

**Anomalias**

A partir do gráfico de dispersão do GDP pelo Happiness Score (imagem 2) e do histograma (imagem 3) percebe-se que o outlier dificulta a visualização e análise dos dados. Além disso, poderia ter atrapalhado o cálculo da correlação, pois o valor do GDP desse país é discrepante em relação aos outros. Portanto, o país cujo GDP é maior que 27 não será considerado, a fim de melhorar a relação entre as variáveis.

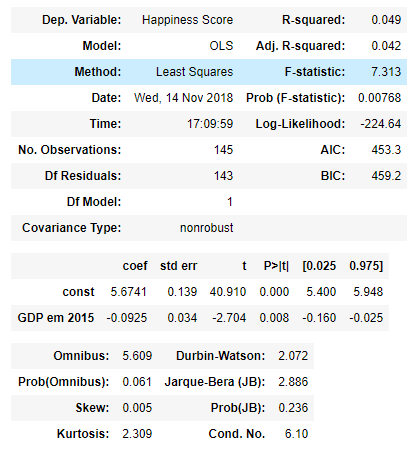
Os gráficos a seguir mostram as relações do GDP com o Happiness Score, sem o outlier (imagem 5), e o histograma do GDP (imagem 6), também sem o mesmo.



*Imagem 5 –Gráfico de dispersão GDP sem outlier Imagem 6 – Histograma do GDP per capita*

Como pode-se ver nos gráficos sem outlier, os dados ficaram melhores distribuídos e a visualização ficou mais clara. Além disso, pelo cálculo da correlação verifica-se que essa aumentou para -0.36, que é consideravelmente maior do que na última análise. Como consequência, a distribuição da variável é observada com mais clareza como uma normal.

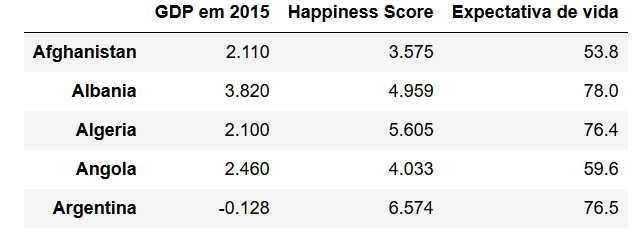
Com os dados sem outlier foi montada novamente a regressão entre as variáveis e obtivemos o seguinte relatório:



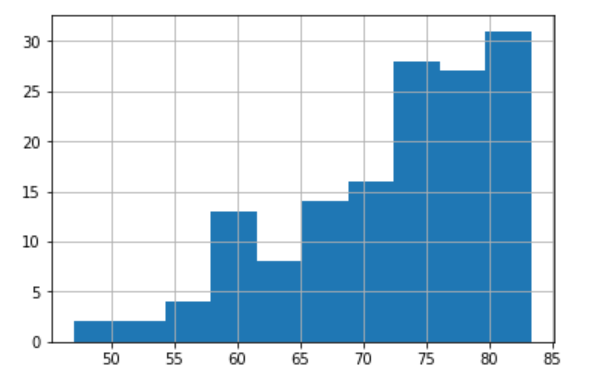
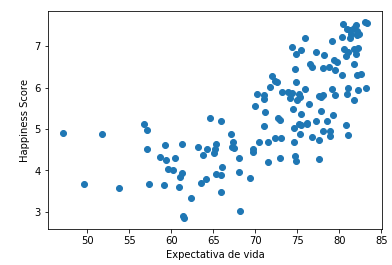
A partir desse resultado, observa-se que o r-squared também aumentou de 0.46 para 0.55 de confiabilidade, demonstrando a melhora no modelo.

**Análise dos dados – Expectativa de Vida**

A segunda variável escolhida para o modelo foi a expectativa de vida, com dados coletados no Gapminder. Assim, juntamos esses dados no dataframe com o GDP e o Happiness Score, obtendo o resultado abaixo.



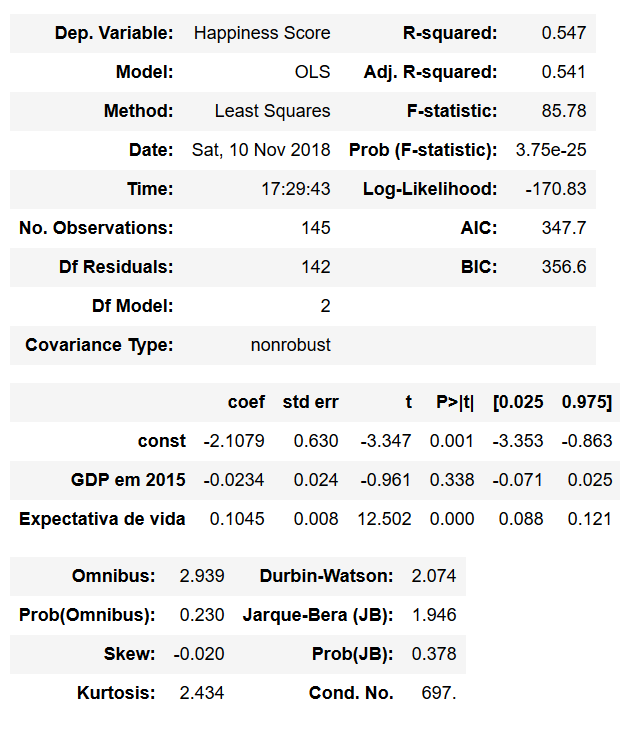
Para a análise e melhor visualização dos dados foram feitos o gráfico de dispersão da Expectativa de Vida pelo “Happiness Score” (imagem 7) e o histograma da expectativa de vida (imagem 8).



*Imagem 7 – Gráfico de dispersão Imagem 8 – Histograma da expectativa de vida*

Ao analisar o gráfico de dispersão percebe-se que a reta que relaciona pontos do gráfico é clara. Assim, ao calcular a correlação percebemos que essas duas variáveis tem o valor de 0.73 de coeficiente de correlação.

A regressão do modelo, com a segunda variável escolhida, expõe o seguinte sumário:

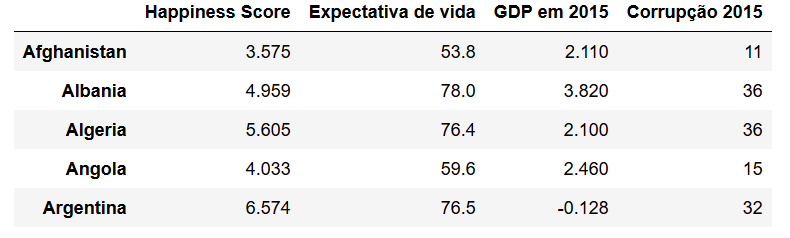


Com os valores expostos verifica-se que a confiabilidade é de 54%. Ademais, a probabilidade de F-statistic é muito baixa, afirmando que pelo menos um dos betas é diferente de zero. Porém, ao observar o p-value de cada uma das variáveis, a probabilidade é baixa em relação ao beta do 'Happiness Score' e da 'Expectativa de vida', quebrando a hipótese nula de que o beta é igual à zero. Mas para o p-value do 'GDP em 2015' não é possível afirmar nada, considerando que o p-value não é um valor baixo, e assim, não quebra a hipótese nula.

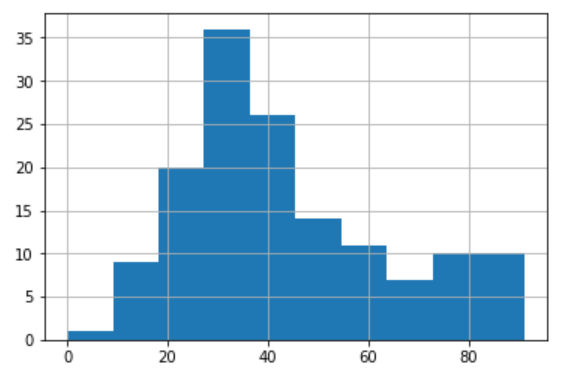
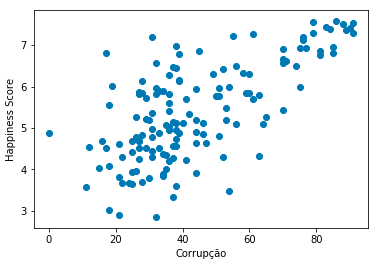
Em relação aos resíduos, tanto pelo método de Omnibus, quanto Jarque-Bera, demonstram que é baixa a probabilidade dos resíduos se comportarem como uma normal.

**Análise dos dados – Corrupção**

A terceira variável escolhida para implementar no modelo foi a corrupção, que também foi obtida a partir do Gapminder. Assim, os seus dados foram adicionados no dataframe com as demais variáveis, como observado na imagem abaixo.

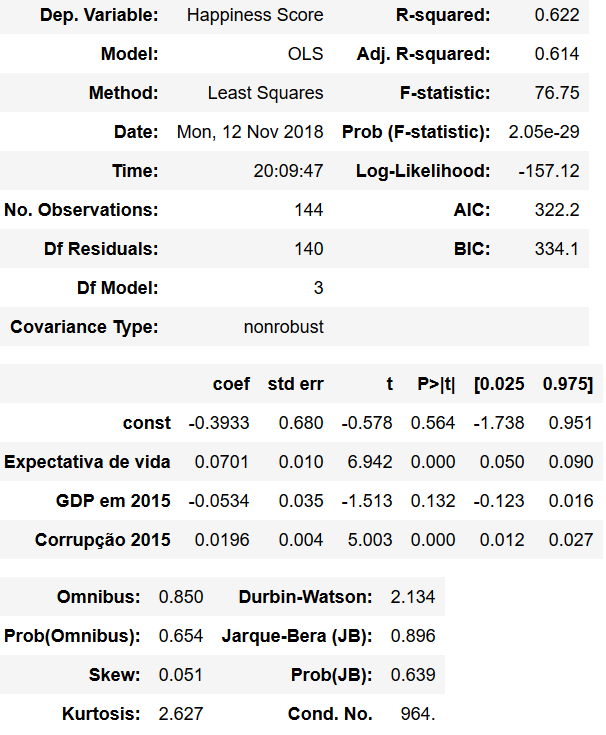


Ao plotar o gráfico, relacionando a Corrupção com o “Happiness Score” (imagem 9), e o histograma da corrupção (imagem 10), os seguintes resultados foram obtidos:



*Imagem 9 – Gráfico de dispersão Imagem 10 – Histograma da corrupção*

Entre essas duas variáveis tem-se uma correlação de 0.68. Ademais, no histograma percebe-se que não é uma curva normal completamente homogênea, mas se aproxima de uma. Assim, ao realizar a regressão do modelo com as múltiplas variáveis escolhidas, seguinte sumário é obtido:



A confiabilidade do modelo aumentou para 62% e a hipótese nula do F-statistic continua sendo quebrada, indicando que não é resultado de uma coincidência.

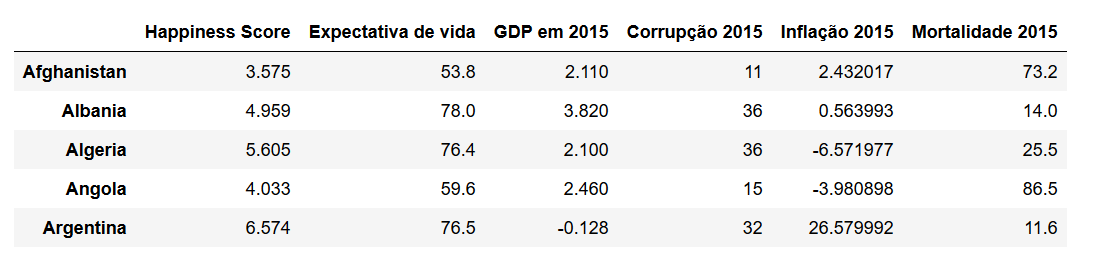
No p-values observamos algumas mudanças também:

* Para o Happiness Score, a probabilidade aumentou significamente, logo não podemos mais quebrar a hipótese nula com relação à essa variável.
* Expectativa de Vida: Continua quebrando a hipótese nula, devido ao baixo valor de p-value.
* GDP em 2015: Seu valor aumentou, e ainda nada se pode afirmar sobre o Beta
* Corrupção 2015: Como seu p-value apresenta um valor muito baixo se recusa a hipótese nula de que seu beta é igual à zero

Em relação a normalidade dos resíduos a partir dos testes Omnibus e JB, os resíduos apresentam ser normais, já que essa estatística de teste demonstra que a hipótese nula, de que é uma normal, não foi quebrada.

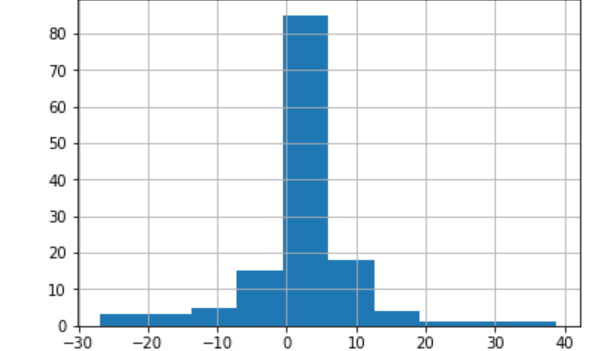
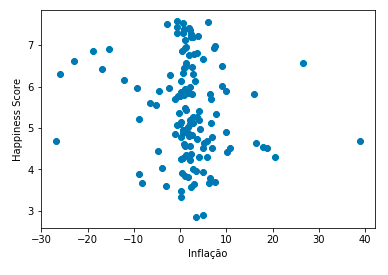
**Análise dos dados – Mortalidade e Inflação**

As últimas duas variáveis escolhidas para a análise foram Mortalidade e Inflação e essas foram adicionadas no dataframe com as outras variáveis, visto na imagem abaixo.



**Inflação**

Para analisar os dados de inflação foram feitos o seu gráfico de dispersão com o Happiness Score (imagem 11) e o seu histograma (imagem 12).

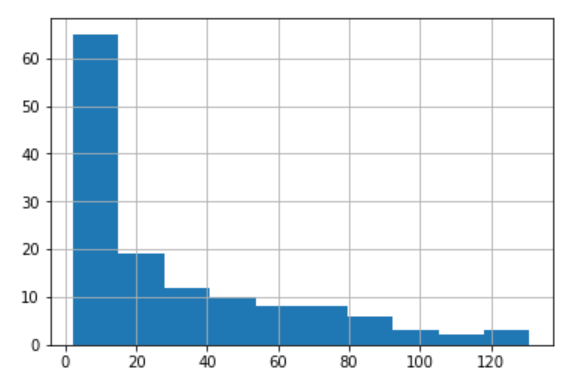
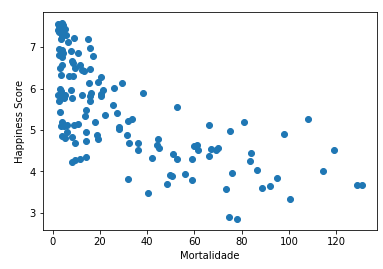


*Imagem 11 – Gráfico de dispersão Imagem 12 – Histograma da corrupção*

Observando o gráfico de dispersão percebemos que há uma concentração considerável de pontos no gráfico, os quais se aproximam de uma reta, porém há uma correlação de -0.15, e esse valor baixo ocorre possivelmente pela grande quantidade de outliers. No histograma é possível observar uma curva normal bem definida, com poucos resíduos em suas extremidades.

**Mortalidade**

Para analisar os dados da mortalidade foram feitos o seu gráfico de dispersão com o Happiness Score (imagem 13) e o seu histograma (imagem 14).



*Imagem 13 – Gráfico de dispersão Imagem 14 – Histograma da mortalidade*

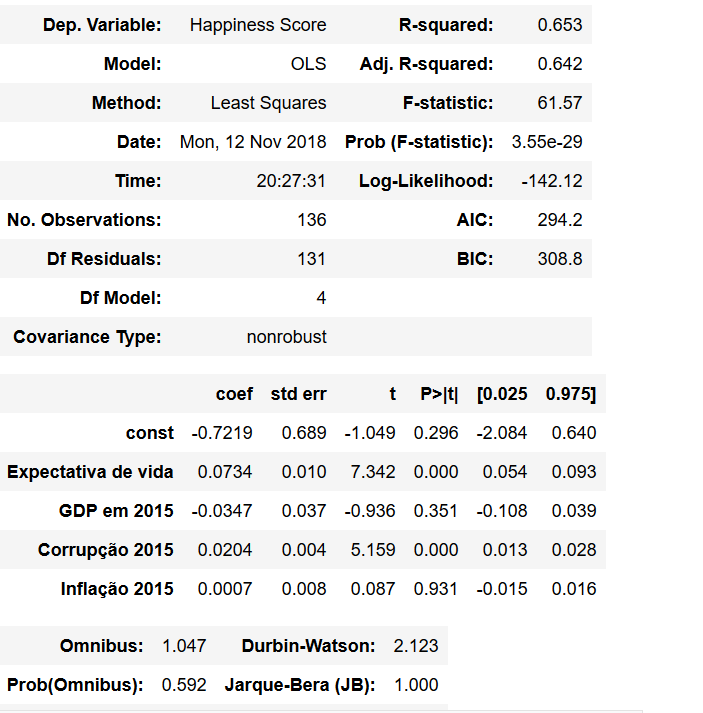
O gráfico de dispersão da Mortalidade apresenta seus pontos concentrados na parte superior do gráfico, com uma correlação de -0.69 com o Happiness Score e o seu histograma apresenta uma grade concentração na esquerda com cauda na direita.

Porém, observa-se na imagem 15 que a mortalidade apresenta uma correlação considerável com a expectativa de vida, cerca de 0.9, o que demonstra que essa variável pode ser dependente da expectativa de vida. Assim, talvez poderia atrapalhar os dados estatísticos, então decidimos retirar essa variável do modelo.



*Imagem 15 – Tabela com as correlações de todas as variáveis*

Como consequência, foi realizada a regressão final com todas as variáveis que serão implementadas no modelo. O resultado dessa foi o relatório a seguir.



A partir dessa análise, e com a implementação da inflação no modelo, a confiabilidade aumentou para 65%, com a probabilidade de F-statistic ainda sendo baixa e assim rejeitando sua hipótese nula.

Sobre os p-values:

* Happiness Score: Diminuiu significamente, e ainda não podemos afirmar nada sobre seu Beta
* Expectativa de Vida: Seu valor continua baixo, ainda rejeitando a hipótese nula
* GDP em 2015: Sua probabilidade aumentou, não conseguindo quebrar a hipótese nula
* Corrupção 2015: Sua probabilidade continua sendo baixa, rejeitando a hipótese nula
* Inflação 2015: Seu valor é muito alto, não rejeitando a hipótese nula e nada se pode afirmar em relação ao seu beta

As estatísticas de teste dos resíduos também apresentam que estes são normais, sendo que a cada nova implementação, esse fator se reforça.

**Metodologia**

Para predizer o valor do “Happiness Score” dos países, as partes de treinamento e de teste foram separadas. Durante o treinamento, nós deixamos o modelo “ver” as respostas, nesse caso o valor do “Happiness Score”, para assim ele aprender a predizer esse valor por meio das variáveis. Então, quando vamos avaliar o modelo, pedimos para que ele faça predições em uma parte de teste, no qual o modelo não tem acesso as respostas, ou seja, ao valor do nível de felicidade. Para o nosso projeto, a base de treinamento será uma parte dos países e o teste será a parte restante, ou seja, o modelo irá predizer o valor do “Happiness Score” para os próximos países do dataset.

**Regressão Linear**

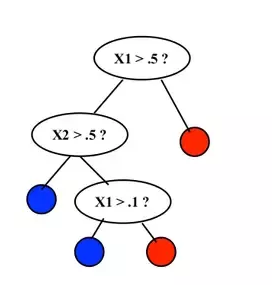
É uma técnica que prevê os resultados a partir da observação da relação entre as variáveis independentes e dependente, montando uma equação que modela o evento. Assim, é como se déssemos os dados, e o modelo tentasse fazer a melhor reta possível com esses dados. Para fazer essa reta melhor possível, há o uso do método dos mínimos quadrados, que compara o quadrado dos erros entre a reta e os dados reais, a fim de que esse erro seja o menor possível, fazendo o ajuste da reta. Como para a parte do treinamento, nós fornecemos os dados do “Happiness Score”, o modelo monta a equação com os dados, e para os outros valores que queremos predizer, ele extrapola a reta obtida com o treinamento, e gera os valores.

**Random Forest Regression**

Para aumentar a performance da predição, o random forest regression foi utilizado. Assim, sabemos que o modelo consegue predizer o valor do “Happiness Score” e de uma forma melhor que a regressão linear, mas é preciso entender como o random forest regression funciona para entender o porquê.

Para entender esse modelo de regressão é necessário entender primeiro o que é uma árvore de decisão (imagem 16). Quando os problemas são de regressão, tentamos predizer números reais na árvore de decisão. Nessa, a ideia é dividir o dataset baseado na homogeneidade dos dados. Assim, ela é formada por vários nós e esses separam os dados em subseções que contém valores similares.

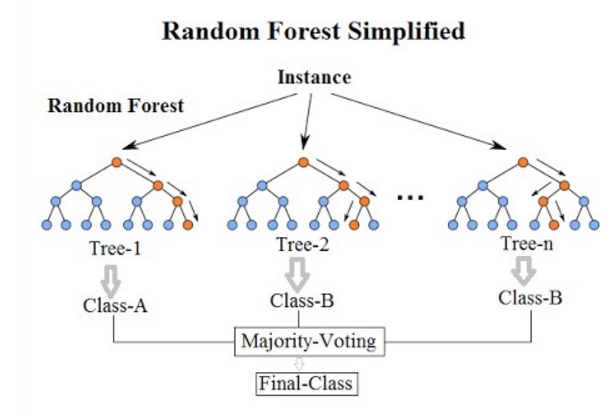
No entanto, em uma árvore de regressão, já que a variável que queremos estimar é um valor real, nós fazemos um modelo de regressão para a variável dependente usando cada uma das variáveis independentes. Então, para cada variável independente, os dados são divididos em vários pontos de divisão. Assim, calculamos a soma do erro ao quadrado em cada ponto de decisão da árvore, entre o valor predito e o valor real. A variável com o menor erro é escolhida para cada nó.



*Imagem 16 – árvore de decisões.*

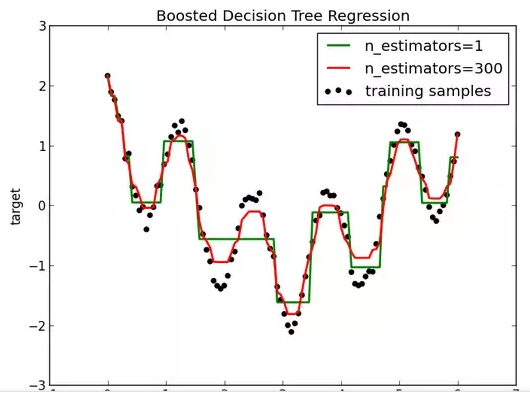
*Fonte: https://www.quora.com/How-does-random-forest-work-for-regression-1*

O Random Forest Regression utiliza várias árvores de decisão e tira a média delas. Esse método é uma aplicação bootstrap da árvore de decisões. Assim, uma explicação visual do Random Forest Regression pode ser visualizada nas images 17 e 18 abaixo.



*Imagem 17 – simplificação do Random Forest.*

*Fonte: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0*



*Imagem 18 – Como o Random Forest aproxima os dados.*

*Fonte: https://www.quora.com/How-does-random-forest-work-for-regression-1*

**Resultados**

Para obtermos o resultado, usamos a biblioteca “linear regression” do sklearn que executa os cálculos e monta a modelagem. Assim, para os nossos dados, o r-square é de 0,54475.

Assim, para melhorar a predição do modelo, decidimos implementar uma nova biblioteca do conjunto sklearn, a random forest, que permite que a regressão se adapte a outras formas, o que aumenta a acurácia da predição.

Verificamos que com essa nova biblioteca, o r-square aumentou para 0,58535. Assim, houve uma melhora, mas não muito discrepante, o que pode ser explicado pelo fato de que os dados tendem a ser lineares.

**Conclusão**

Portanto, pode-se concluir que o método Random Forest Regression é mais eficiente para predizer o valor do Happiness Score do que a regressão linear. No entanto, é possível observar que a pouca diferença entre a performance usando linear regression e o random forest deve ser decorrente do fato de que o modelo se aproxima de uma reta, o que faz com que a regressão do random forest fique melhor, mas não absurdamente, já que o método é de aproximação por “escadinhas”, similar à imagem 18.

Assim, observamos que é possível predizer razoavelmente bem o Happiness Score com dados contínuos, mas não é tão fiel aos dados da pesquisa qualitativa antes obtida. Como partimos de dados quantitativos tivemos algumas discrepâncias nos resultados, mesmo escolhendo as variáveis com maior correlação. Isso pode ser fruto do fato de que a felicidade é mais próxima de um estado de espírito das pessoas, e não necessariamente de uma relação contínua entre fatores econômicos e sociais.

**Referências bibliográficas**

Random Forest in Python. *Disponível em: <https://towardsdatascience.com/random-forest-in-python-24d0893d51c0>. Acesso em: 14 nov. 2018.*

Simple and Multiple Linear Regression in Python. *Disponível em: <https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9>. Acesso em: 14 nov. 2018.*

*How does random forest work for regression? Disponível em: <https://www.quora.com/How-does-random-forest-work-for-regression-1 >. Acesso em: 19 nov. 2018.*

*Random Forest Simple Explanation. Disponível em: <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d >. Acesso em: 19 nov. 2018.*

*Gap Minder. Disponivel em: <https://www.gapminder.org/tools/#$chart-type=bubbles> Acesso em 13 nov. 2018.*