Análise resultados Trabalho 3 IIA 2019.2 UNB

Pedro Luis Chaves Rocha - 18/0054635

Novembro 2019

1 Introdução

Para o trabalho 3 de Introdução a Inteligência Artificial do ano de 2019.2 da UNB, foi realizado um programa que, dependendo de 36 atributos (partial faces, is female, baby, child, teenager, youth, middle age, senior, white, black, asian, oval face, round face, heart face, smiling, mouth open, frowning, wearing glasses, wearing sunglasses, wearing lipstick, tongue out, duck face, black hair, blond hair, brown hair, red hair, curly hair, straight hair, braid hair, showing cellphone, using earphone, using mirror, braces, wearing hat, harsh lightinge e dim lighting) , avaliar a popularidade de uma selfie de acordo com os dados adquiquiridos nesse SITE

O valor da popularidade é medido em uma escala de 0 a 10 em variáveis do tipo flutuante sendo que, nos dados baixados , 0 valor mínimo arredondado é 1 e o máximo arredondado é 7, isso é importante identificar pois, para esse projeto foi feitos 2 algoritmos de árvores randômicas um que se usa de classificação, e outro que usa de regressão utilizando uma biblioteca de python chamada de skicit learn.

2 Análise de Resultados

Os testes revelaram principalmente, quais atributos mais impactam na hora de realizas a classificação ou da regressão, sendo os mesmos para ambos os casos

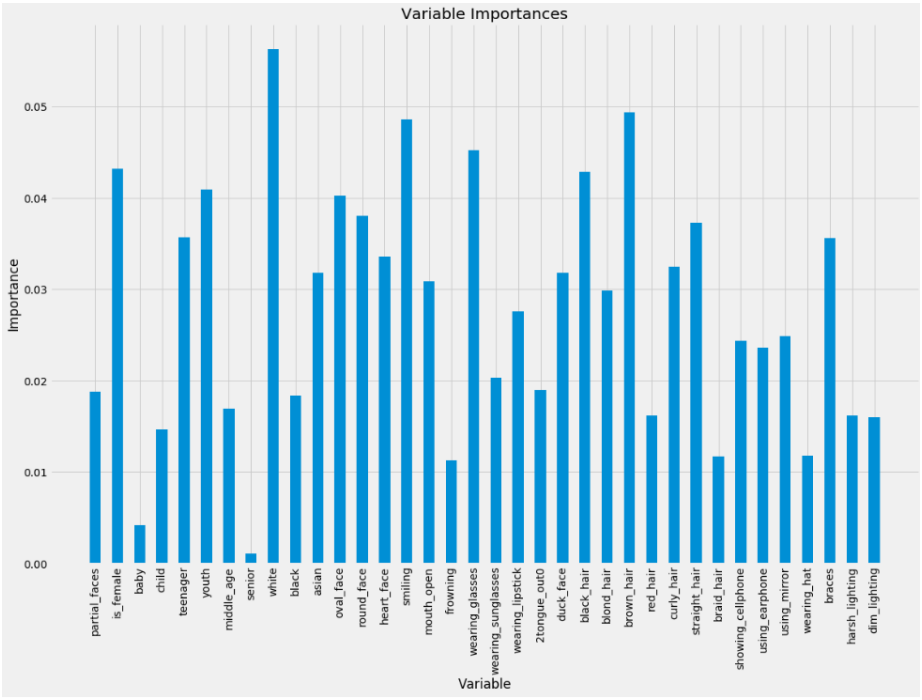

Figure 1: Importâncias das Variáveis

Pode-se observar que mesmo as variáveis com maior importância dente elas apresentam um valor relativamente baixo próximo de 0.06 o que indica que para os dados utilizados, não importa as variáveis, elas impactam muito pouco no resultado final, podendo assim, prever que o uso de classificadores pode não ser muito eficiente principalmente pelo uso de um numero grande de inputs e pelo fato de que esses inputs interferem pouco no resultado.

Mesmo após a análise das importâncias para cada variável, as análises continuaram sendo feitas para os Classificador e para a Regressão do Random Forest. Na implementação feita, foram feitas análises diferentes em que o modelo treinado pode ser salvo setando a variável isTrained como False, ou pode-se utilizar o modelo anteriormente treinado sentando a mesma isTrained como True, ou até retreinar uma outra floresta como um mínimo de importâncias em árvores setando o valor mínimo na variável threshhold e o retrain como True. Essas alterações servem principalmente para evitar o over fitting, problema mais comuns em árvores de decisão, no entanto, não houve melhora significativa na performance dos algoritmos, isso pode ser comprovado ao observar o score do trainning set, em que deu aproximadamente uma melhora de aproximadamente 3% na precisão.

No entanto, ao realizar uma poda na árvore de modo a avaliar somente os atributos com importância acima de 0.04 obtemos uma distribuição de im-

portâncias diferentes como mostra a imagem a seguir.

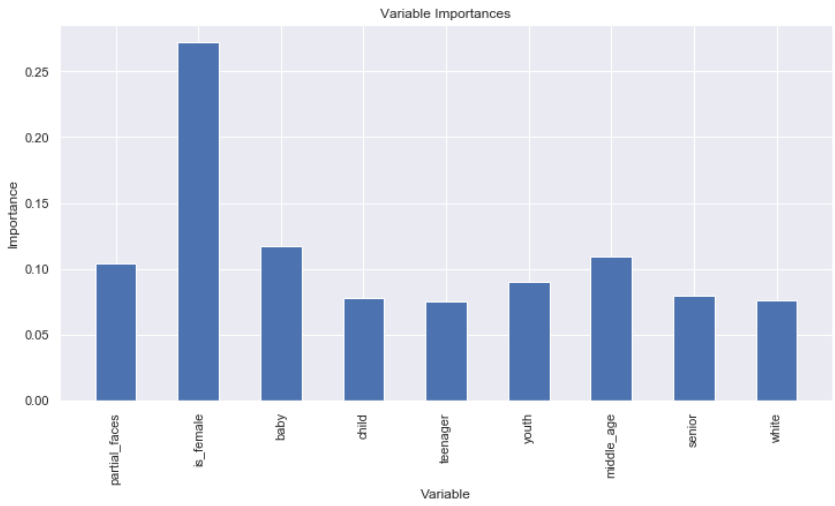

Figure 2: Importâncias das Variáveis

Isso indica que, quado não há muitas variáveis para dividirem a importância do resultado da popularidade, pode-se observar que quem se sobressai é exatamente o atributo (is female), pois mesmo podando, ainda foi observado uma precisão de aproximadamente 83% em ambos os casos, no de classificação e no de regressão.

Para algoritmo de classificação foram feito arredondamentos para os valores de ponto flutuante com o abjetivo de fazer uma classificação de 0 a 10, pois caso contrário, o sckit learn teria um erro , já que é impossível classificar algo em infinitas possibilidades como nos números de ponto flutuante. Para a medida do erro, foi utilizado o Mean Absolute Erro ou MAE, ele mede a diferença entre o valor real e o valor que foi previsto de acordo com os variáveis de teste (0.1 de todos os dados) e foi obtido um valor próximo de 0.4329 (pode variar já que o algoritmo é aleatório), isso indica que para uma classificação, que leva em conta variáveis discretas, se obteve um erro médio ao quadrado de 0.4 (pode variar devido a aleatoriedade do algoritmo) , que demonstra quase 40 % de erro, o que pode ser comprovado ao medir a precisão de acordo com os valores reais e os valores previstos a partir dos dados de testes (0.1 dos dados totais) e obtendo um valor aproximado de 62,1 % de acerto, ou seja, uma precisão muito baixa muito provavelmente por causa da baixa dependência do modelo em relação a suas variáveis, pois como é observado nessa matriz de confusão e com a informação das previsões, a maioria dos dados está próximo de 4 simplesmente por serem dados reais, ou seja, é normal ter maior concentração em torno da media, fazendo com que a maioria dos acertos sejam encima deles.

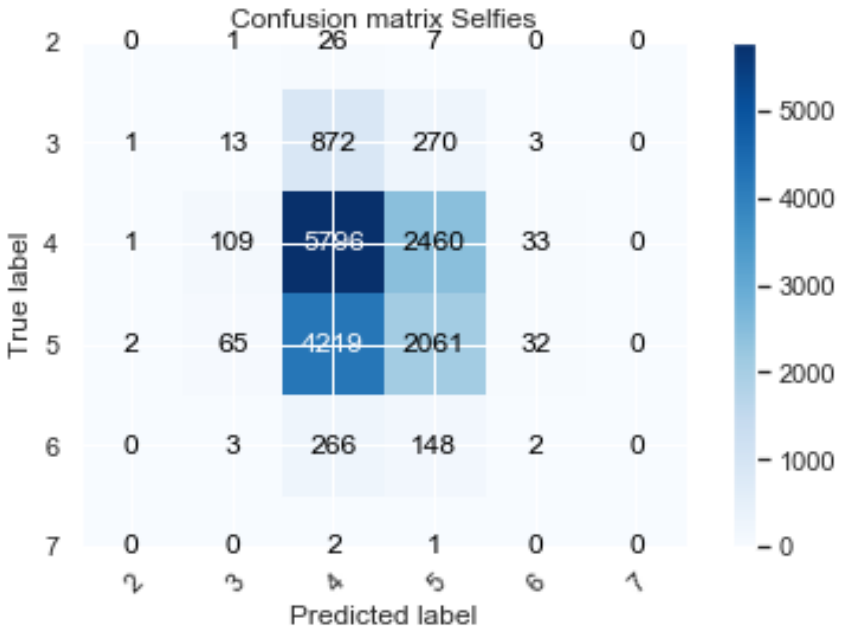

Figure 3: Matriz de Confusão

Tendo isso em vista, os valores das avaliações foram divididos em 5 subintervalos, esses intervalos foram divididos ao pegar o valor máximo, somado ao mínimo e dividido por 5, pois assim, é praticamente garantido que a média 4 fique próximo do terceiro intervalo, que intervalo). Assim, ao treinar a inteligência artificial, foi obtido uma precisão de aproximadamente 85% , o que já está de excelente tamanho para o problema de classificação em 5 classificações. Além disso, a importância medida das variáveis foi praticamente a mesma, com pequenas variações provavelmente pelo caráter randômico do algoritmo.

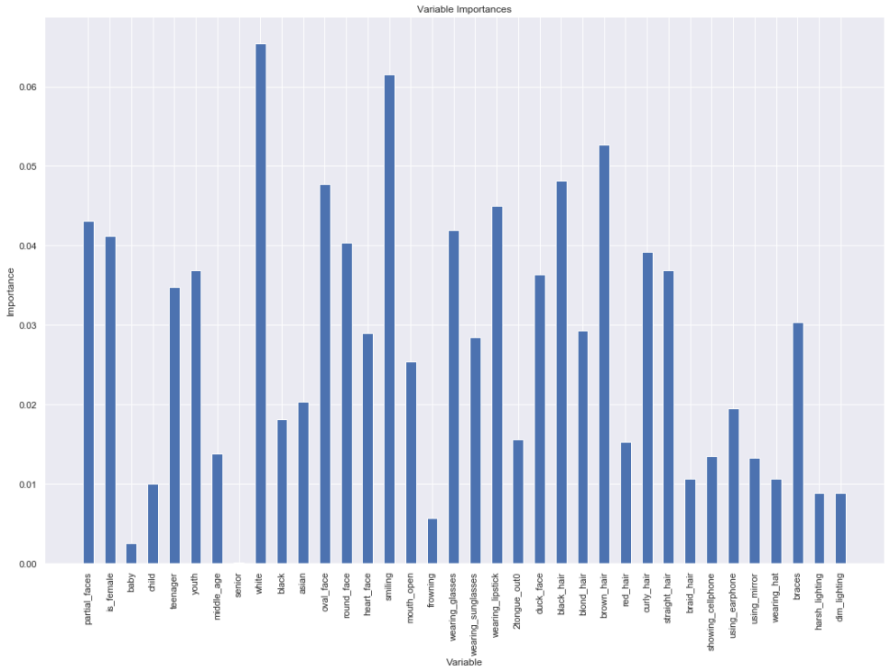

Figure 4: Importâncias das Variáveis Classificador

Já para o algoritmo de regressão, é feita uma aproximação de função para as variáveis que representam as características das fotos retornando sempre sua popularidade, podendo assim, assumir valores de ponto flutuante. O regressor é treinado também com 0.9 dos dados utilizados e testado com 0.1. Para medir o erro também foi utilizado o Mean Absolute Error ou MAE e foi obtido um valor próximo de 0.4, o que indica que a diferença media entre os valores previstos e os valores reais é de 0.4, ou seja em média, uma popularidade de, por exemplo 5, é avaliada ente 4.6 e 5.4 o que para uma avaliação é uma distância muito grande. Mas, por serem casos contínuos, a precisão foi medida de a cordo com o Mean Absolute Percentage Error ou MAPE que tem como fómula o MAPE (True Value/Pred Value)/True Value * 100 que retornou um valor próximo de 10 % e ao pegar seu inverso, indica uma precisão de 90 % ou seja em média, a previsão está correta em 90 % ou seja, temos um modelo nem ruim e nem ótimo, mas aceitável para indicar a popularidade de uma foto de acordo com 36 características, exatamente por apresentar uma variação relativamente aceitave dentre as previsões feitas.

3 Conclusão

Como conclusão, foi possível observar que a dependência da popularidade de uma foro depende muito pouco das 36 variáveis apresentadas como ilustra o gráfico, assim dificultando a utilização de um modelo de classificação que foi somente possível ao arredondar os valores de popularidades e classificá-los de 0 a 10, no entanto os resultados obtidos não foram nada satisfatórios por haver um acerto de apenas 60 % dente os 1% dos dados reservados para testes, no entanto ao reduzir as classificações para 5 possíveis, foi possível obter um melhor resultado de 85% de precisão, também pelo fato de ser um numero ímpar, já que a maioria do dataset está classificado na média de 4 em avaliação. Já o modelo de regressão foi mais satisfatório por não medir acertos e erros, mas aproximar uma função que ao inserir as características das imagens, ele retorna um valor de popularidade que, de acordo com os 1% de dados testados, ouve um erro médio de 0.4, ou seja, os valores diferem dos reais em média 0.4, e ao levar em conta a precisão, de acordo com o MAPE (Mean Absolute Percentage Error), pôde-se observar que em média as previsões são corretas em 90% do real.