Aplicação de Técnicas de Aprendizado de Máquina para Prever o Desempenho das Escolas Na Avaliação Nacional do Ensino Básico

Erico André[1], **Everton Veloso**[1], **Kimbelly Ferraz**[1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brasil

Abstract. *This paper presents analyzes through the use of infrastructure data from the Brazilian Schools and the SAEB test obtained from the School Census, with deep learning techniques, to infer if the performance of 5th grade students is directly or indirectly related to infrastructure of schools in the Brazilian regions.*

Resumo. *Este artigo apresenta análises através da utilização de dados de infraestrutura das Escolas Brasileiras e da prova do SAEB obtidos do Censo Escolar, com técnicas de aprendizado profundo, para inferir se desempenho dos estudantes do 5 ano do ensino fundamental, esta relacionada diretamente ou indiretamente com a infraestrutura das escolas das regiões brasileiras brasileira.*

1. Introdução

Há diversos estudos que relacionam o desempenho escolar à um conjunto de fatores, dentre estes fatores, é possível citar questões como infraestrutura, segurança, entre outros. Dos então citados é preciso dar um destaque especial à infraestrutura das escolas, fator base para a promoção de um ensino e aprendizado efetivo. Para melhorar a infraestrutura das escolas é também buscar melhorar a presença dos estudantes, a motivação dos funcionários, e aumentar o desempenho e conquistas estudantis. O processo de classificação das melhores escolas da região brasileira utilizando parâmetros de infraestrutura, e técnicas de Deep Learning, com a utilização de três algorítimos: SVM "Maquinas de vetor de Suporte", Random Forest "Floresta Aleatória"e Rede neural artificial "Multilayer Perceptrons"MLP, conjunto de dados foi dividido primeiro quartil e quarto quartil, atributo usado para a divisão media de língua portuguesa do quinto ano. Onde comparamos a acurácia dos modelos SVM vs ANN.

SVM "Maquinas de vetor de Suporte"é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para desafios de classificação ou regressão. Que pega os dados como uma entrada e gera uma linha que separa essas classes, se possível. Ele encontra os pontos mais próximos de ambas as classes. Esses pontos são chamados vetores de suporte. O kernel seleciona o tipo de hiperplano usado para separar os dados. O "linear" usará um hiperplano linear (uma linha no caso de dados 2D). 'rbf' e 'poly' usam um hiperplano não linear. o valor padrão é "rbf", Gama é um parâmetro para hiperplanos não lineares. Quanto maior o valor de gama, mais ele tenta ajustar-se exatamente ao conjunto de dados de treinamento, C é o parâmetro de penalidade do termo de erro. Ele controla o trade off entre o limite de decisão suave e a classificação correta dos pontos de treinamento. O aumento dos valores de C pode levar a um overfitting dos dados de

treinamento, degree é um parâmetro usado quando o kernel é definido como “poly“. Basicamente, é o grau do polinômio usado para encontrar o hiperplano para dividir os dados [Fluente 2019].

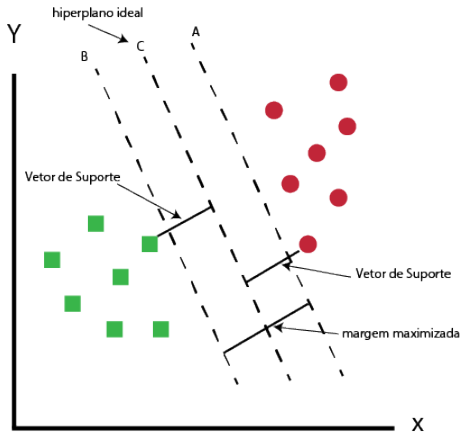

Figura 1. Hiperplano usando o algoritmo SVM.

Por fim, Redes Neurais Artificiais ou MLPs (Multilayer Perceptrons). A primeira camada é a entrada e a última camada é a saída. Se houver mais de uma camada oculta, nós as chamamos de redes neurais ”profundas”(ou Deep Learning). Esses tipos de redes neurais calculam uma série de transformações que alteram as semelhanças entre os casos. As atividades dos neurônios em cada camada são uma função não-linear das atividades na camada anterior.é um modelo preditivo motivado pela forma como o cérebro funciona, consistem de neurônios artificiais, que desenvolve caculos similares sobre suas entradas.[Academy 2019]

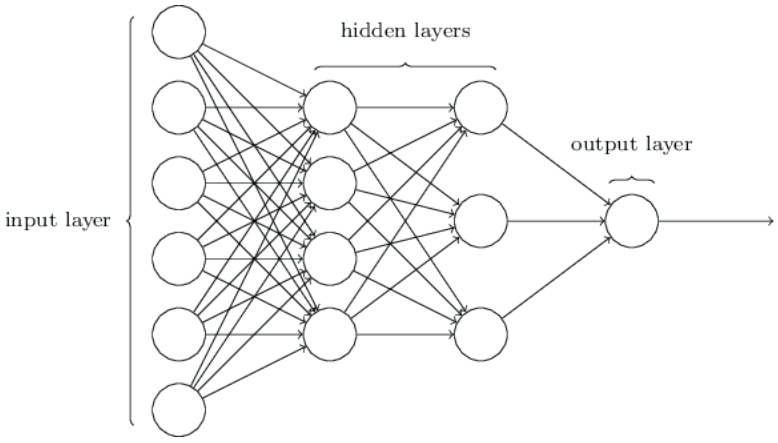

Figura 2. Rede Multilayer Perceptrons.

1.1. Objetivo

Este trabalho tem como objetivo desenvolver um classificador com base infraestrutura das Escolas Brasileiras e da prova do SAEB obtidos do Censo Escolar, as escolas das regiões brasileira que necessitam de melhorias em suas infraestruturas para poder manter o nível do ensino no pais.

2. Metodologia

Foi criado um novo conjunto de dados a partir dos dados de escolas, região, da prova do SAEB, originando um novo conjunto de dados contendo todas as informações chamado de base_completa, que foi dividida em primeiro e quarto quartil, onde o primeiro quartil representa as escolas que precisa de atenção, e as que se encontram no quarto são escolas considerada desempenho ”ótimo”, esse novo conjunto de dados foi utilizado em todos os algorítimos testados neste artigo.

Para realizar a classificação, foi utilizada Keras e sklearn, uma API de alto nível de redes neurais, maquinas de vetor de suport ”SVM”e Florestas aleatórias , capaz de executar em cima de TensorFlow , CNTK , ou Theano [Keras.io 2019].

2.1. Pré-processamento

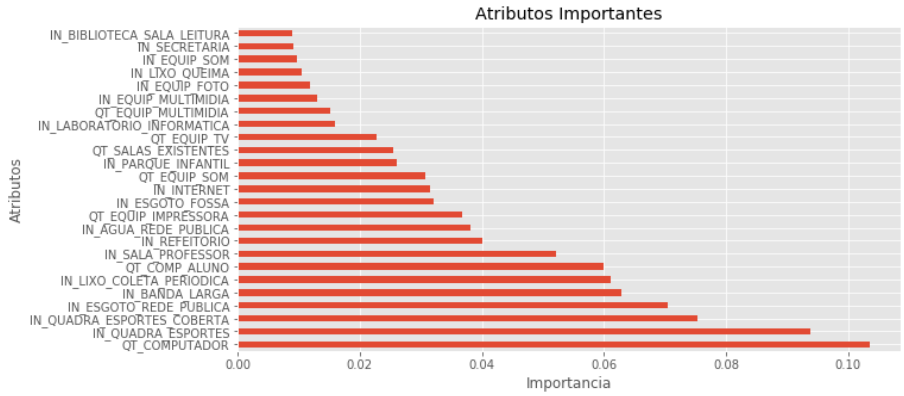

Figura 3. Atributos Selecionados Através do Random Forest.

Google Colab, plataforma usada para executar os algoritmos de treinamento da rede [Google 2019].

2.2. Treinamento

3. Resultados

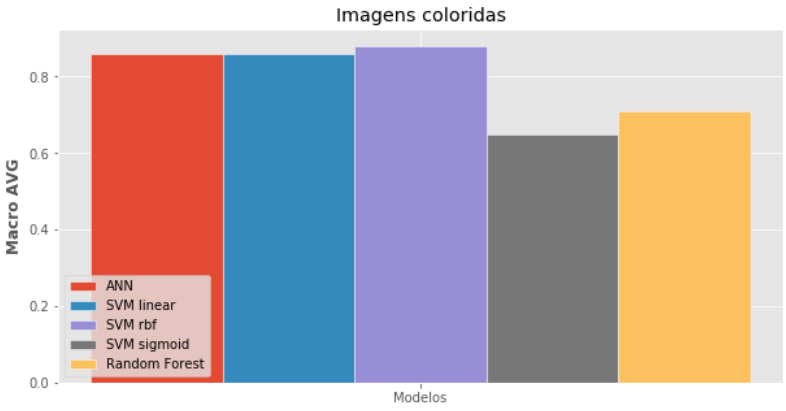

Figura 4. Comparações dos diferentes métodos usados.

4. Conclusão

Referências

[Academy 2019] Academy, D. S. (2019). Deep learning book. http://deeplearningbook.com.br/a-arquitetura-das-redes-neurais/. Accessed: 2019-11-26.

[Fluente 2019] Fluente, C. (2019). Código fluente - um blog sobre linguagem de programação c, python, r, django, ciência de dados. https://www.codigofluente.com.br/aula-08-scikit-learn-maquina-de-vetores-de-suporte/. Accessed: 2019-11-22.

[Google 2019] Google (2019). Google colaboratory. https://colab.research.google.com/. Accessed: 2019-11-18.

[Keras.io 2019] Keras.io (2019). Keras: The python deep learning library. https://keras.io/. Accessed: 2019-11-18.