RELATORIO CIENCIA DE DADOS

Aluno:Matheus Ribas

O nosso grupo recebeu a dataset para prever os valores das casas, nosso objetivo principal era entender como as variáveis podiam e faziam diferenças no valor da média final das casas, tínhamos variáveis como criminalidades, distâncias das áreas de emprego, renda da população.
Entendemos desde o princípio que isso era um problema relacionado à regressão, pois estávamos tentando prever um valor numérico contínuo.
O Dataset tinha colunas que tinham muitas e poucas relações, foi usado alguns gráficos para análises disso, como o gráfico quente e frio, que conseguimos identificar variáveis que eram ou não eram mais informativas e tinham relações altas com os valores da média das casas, existiam poucas linhas que estavam com valores nulos, fizemos uma média pela coluna e inserimos no local, pois talvez seriam dados importantes a serem levados em consideração para excluirmos.
Conseguimos com os dados das colunas fazer análises e descobrir quais as influências maiores no valor das casas, descobrindo coisas como locais com maior taxa de criminalidade tendem a ter suas casas com menores valores ou também que casas com mais cômodos tendem a ser mais caras, analisando isso conseguimos ver como tudo pode afetar no valor das casas ao final do projeto.

Avaliando os modelos com as métricas para ver quais têm mais erro.

Vemos que a Random Forest teve melhor desempenho que o modelo RNA sendo a mais precisa para a previsão dos valores das casas.

Acreditamos que o Random Forest seja mais interpretável pois conseguimos entender suas decisões, que são sempre pegar subconjuntos aleatórios, e pegando ao final a média de todas as árvores e tendo um valor final.

A RNA demorou mais a ser treinada para chegar em um resultado um pouco melhor que foi 0.75 com menos erros, tivemos que inserir mais neurônios na segunda camada para conseguir ter um bom resultado sendo preciso 1000 vezes treinada para um resultado bom, já a RANDOM FOREST após a criação de 100 árvores já teve um resultado bom.

Para a RNA funcionar teve que ser alterado alguns dados, como novas variáveis, adição de uma segunda camada de neurônio e na RANDOM FOREST foi algo mais simples sem precisar de mudanças apenas as que já tinham sido feitas, porém ficamos em 100 árvores para não correr risco ao invés de melhorar, piorar.

.