**Relatório Trabalho Aprendizado de Máquina**

**Jéssica Martins de Oliveira**

**Objetivo**

Assim como ensinado nas aulas de Aprendizado de Máquina, os alunos baixaram o Anaconda. O professor disponibilizou dois conjuntos de dados, onde o primeiro (*train.csv*) apresentava nas colunas o Id, algumas características de imóveis, sendo a última coluna o preço referente a cada um. O segundo (*test.csv*) não continha a coluna “preço”, que faria parte da resposta enviada para o site Kaggle, escolhido para submeter os trabalhos. A ideia é que, a partir das características do arquivo de treino, pudéssemos predefinir o preço dos imóveis que faziam parte do arquivo de teste.

**Metodologia**

O tratamento do arquivo de treino consiste em alguns passos. Primeiro, foi verificado as colunas que apresentavam *string* em sua composição, e para elas, foi usado o comando *drop* para eliminar as colunas que tinham esse tipo de dados, que não seriam tratados, pois precisamos que as colunas tivessem números. Colunas como bairro e diferenciais de um imóvel, são difíceis de serem tratadas na regressão linear, pois podem ser representadas por infinitos números, que para concluir um resultado pertinente, deveriam ser representados pelos mesmos bairros tanto no arquivo de teste, quanto no arquivo de treino. Para que não houvesse ambiguidades, essas colunas foram eliminadas do processo.

Em segundo lugar, foi verificado as linhas com imóveis que fugiam do padrão, com preços muito altos ou muito baixos se comparados com imóveis semelhantes. Essas linhas da tabela também foram eliminadas, pois causavam uma distorção na verificação de dados, se caracterizando pincipalmente como exceção.

Após isso, as colunas do arquivo foram identificadas como atributos, com exceção da última, identificada como alvo. Os atributos foram identificados com a letra “X”, enquanto o alvo, com a letra “y”, e representando as características dos imóveis e seu preço, respectivamente. Os atributos e alvos também foram divididos em grupos de treino e teste, levando os nomes *X\_train, X\_test, y\_train, y\_test*, com separação de 0,4 do total para o grupo de testes.

Foi usada a função *LinearRegression().fit* para estimar a função responsável por correlacionar os atributos *X\_train* com os alvos *y\_train*. Enquanto para o conjunto de dados separados como teste, foi usada a função *LinearRegression(). predict*, para estimar os valores do alvo a partir dos dados de *X\_test*.

Na segunda parte do trabalho, foi usado o arquivo chamado *teste.csv*, com os dados que não tem a última coluna “preço”. Para esse conjunto de dados, foi feito um tratamento semelhante ao anterior: as colunas com *string* também foram eliminadas, e todas as colunas (já que elas não têm o alvo), foram chamadas de atributos. Assim, o alvo foi estimado pela função *LinearRegression(). predict,* e os resultados foram guardados em um arquivo *resultado.csv* que contém o Id e o preço estimado.

**Resultados**

Assim que o arquivo final *resultado.csv* foi enviado para o kaggle, este comparou os preços de achados por todos os alunos, e ordenou de acordo com os melhores resultados aproximados da resposta original. Como resultado, o Kaggle definiu Score = 0,42860.