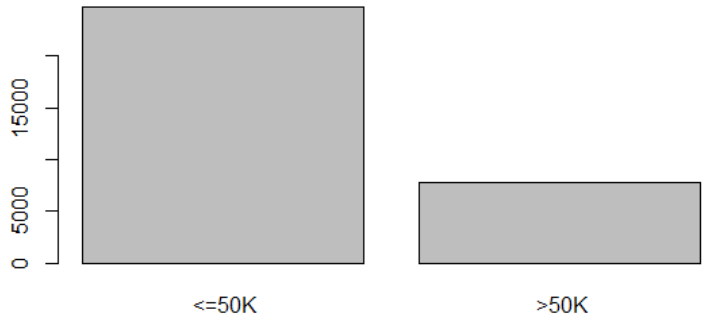
Candidato: Jhaidan Ribeiro Cruz

## 1 DESCRIÇÃO GRÁFICA E ESTATÍSTICA DOS DADOS

A descrição gráfica e estatística dos dados foi realizada por meio da linguagem R. A principal variável do conjunto, a ser predita (yearly_wage) foi identificada como qualitativa nominal, ou categórica (*factor* no R). Esta classifica a amostra em dois tipos salários: maiores que 50 mil e menores ou iguais que 50 mil. Na Figura 1 é apresentado o gráfico de barras no qual pode-se identificar as frequências absolutas. Este tipo de gráfico foi escolhido pois descreve adequadamente a variáveis qualitativas. Nota-se que a maioria das amostras indicam salário anual menor ou igual a 50 mil. As frequências relativas são: 0,76 para <=50K e 0,24 para >50K.

**Figura 1** – Gráfico de Barras de yearly_wage (Salário Anual)




Dentre as outras variáveis do conjunto, abaixo são apresentadas as frequências absolutas (gráficos de barras) das demais variáveis categóricas.

**Figura 2** – Gráfico de Barras de workclass (Classe de Trabalho)

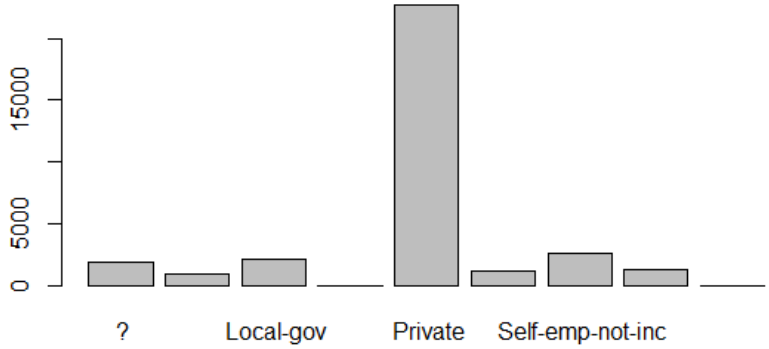


**Figura 3** – Gráfico de Barras de education (Educação)

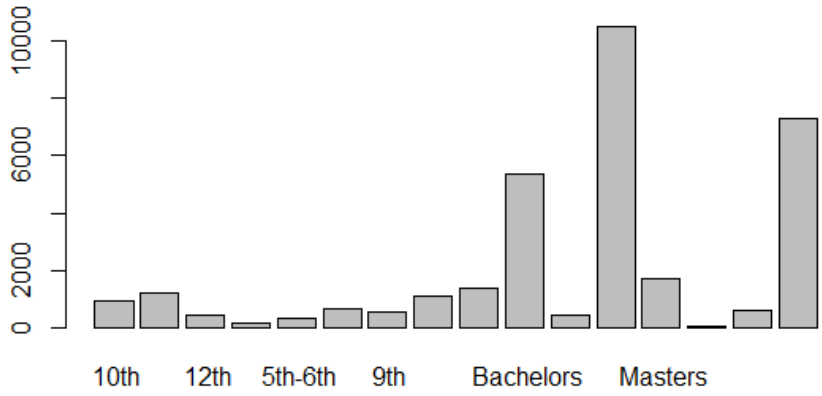



**Figura 4** – Gráfico de Barras de marital_status (Estado Civil)


**Figura 5** – Gráfico de Barras de occupation (Profissão)

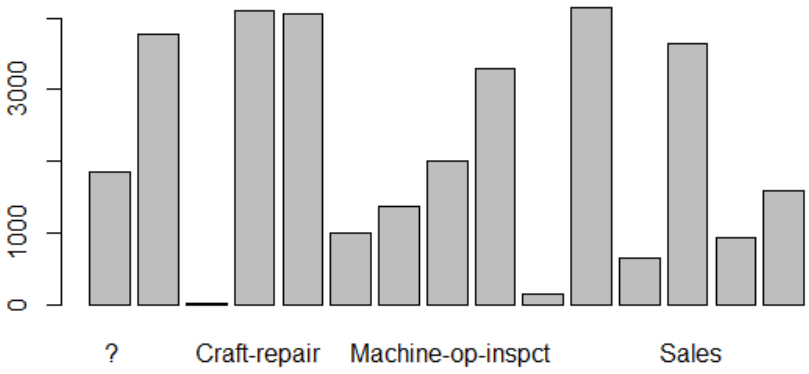


**Figura 6** – Gráfico de Barras de relationship (Relacionamento)

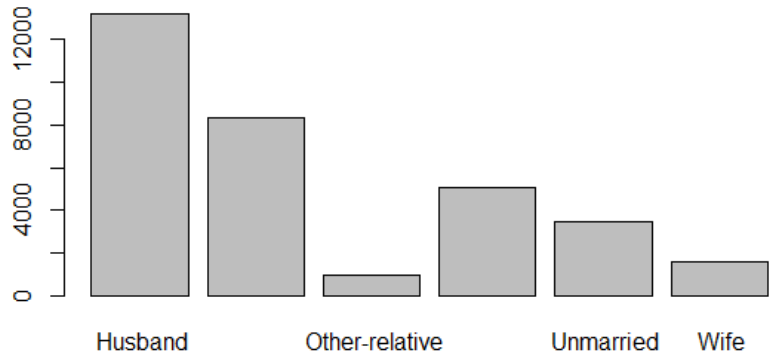



**Figura 7** – Gráfico de Barras de race (Raça)

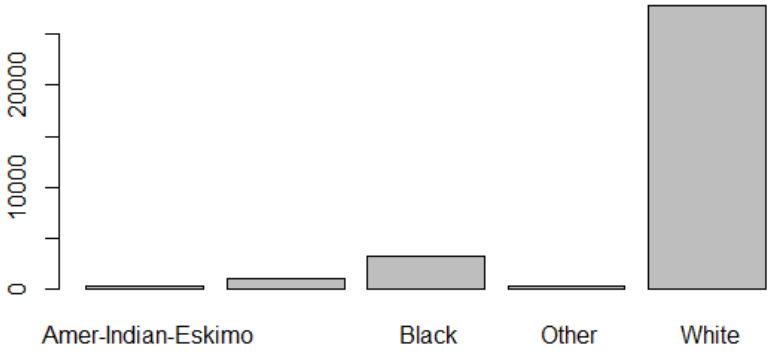



**Figura 8** – Gráfico de Barras de sex (Sexo)

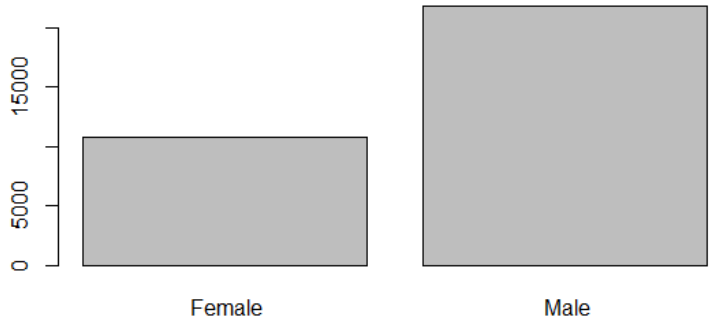
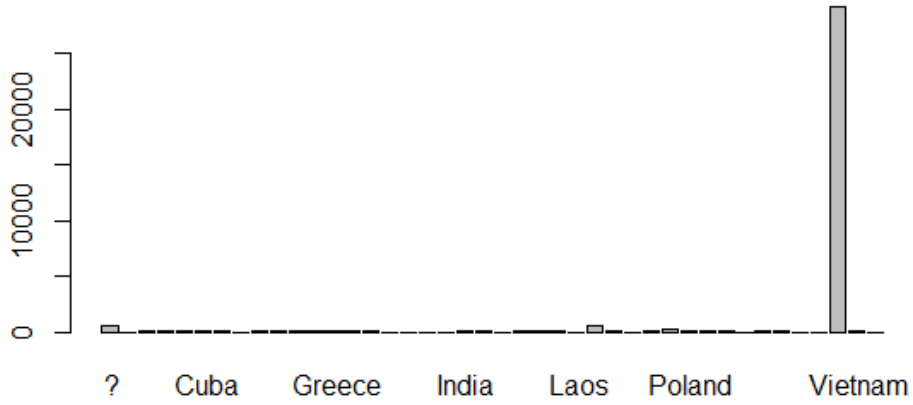


**Figura 9** – Gráfico de Barras de native_country (País Nativo)




Além das variáveis acima, as demais variáveis são quantitativas (*numeric* no R). São elas: age, fnlwgt, education_num, capital_gain, capital_loss, e hours_per_week. Foram escolhidas duas, devido as seus significados perante o conjunto de dados, para descrição estatística e gráfica: age e hours_per_week. Foram escolhidos o gráfico de frequências relativas (Figuras 10 e 11) e as medidas de posição e dispersão (Tabela 1), por apresentar e descrever adequadamente variáveis do tipo quantitativa discreta.

**Figura 10** – Gráfico de Frequências Relativas de age (Idade)

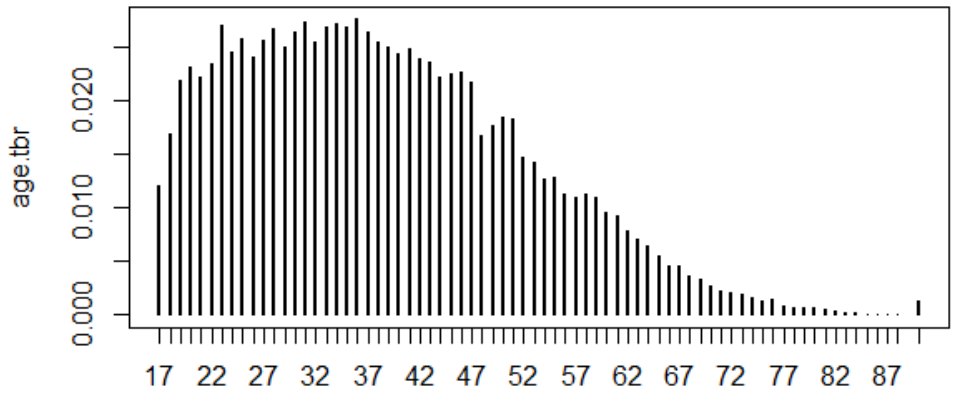


**Figura 11** – Gráfico de Frequências Relativas de hours_per_week (Horas por Semana)

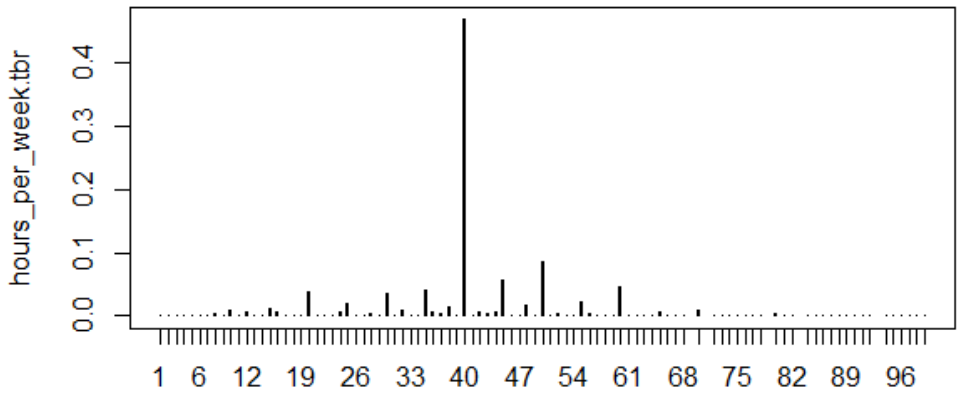



**Tabela 1** – Medidas de posição e dispersão de age (Idade) e hours_per_week

| Variável | Média | Desvio Padrão | Moda | Mínimo | 1º Quartil | Mediana | 3º Quartil | Máximo |
|---|---|---|---|---|---|---|---|---|
| age | 38,58 | 13,64 | 36 | 17 | 28 | 37 | 48 | 90 |
| hours_per_week | 40,44 | 12,35 | 40 | 1 | 40 | 40 | 45 | 99 |

## 2 PREVISÃO DE SALÁRIO ANUAL

A predição em questão foi realizada por meio do *software* (e também linguagem de programação) Matlab. Trata-se de um problema de classificação, pois a variável a ser predita (yearly_wage) é do tipo qualitativa nominal. Esta é não numérica (e consequente não contínua) e possui duas classes (<=50K e >50K). Sendo assim, a medida de performance adequada para escolha e treinamento do modelo é o *accuracy* (comumente traduzido como acurácia, ou precisão), pois esta representa a relação entre o total de acertos e o total de classificações, e não apenas acertos para uma das classificações. No caso da predição da variável em questão, não foi especificada importância maior para uma das duas classes.

A primeira etapa da predição inclui a transformação das variáveis e a escolha do modelo, realizada na primeira parte do código do arquivo "wage_predict.m". Inicialmente, as variáveis foram transformadas para seus tipos adequados, semelhante às transformações apresentadas na seção 1. As variáveis transformadas para qualitativas (*categorical* no Matlab) foram: workclass, education, marital_status, occupation, relationship, race, sex, native_country e a principal yearly_wage. As variáveis mantidas como quantitativas (*double* no Matlab) foram: age, fnlwgt, education_num, capital_gain, capital_loss, e hours_per_week. A escolha do modelo foi feita por meio do aplicativo "Classification Learner" da barra de ferramentas "Machine Learning and Deep Learning" do Matlab. Na Figura 12 é apresentada a seleção e configuração dos dados para o treino dos modelos. Na Figura 13 é mostrado a escolha de todos os modelos para serem treinados e observados (selecionando "all" na aba "model type"), além da ativação do PCA (Análise de Componentes Principais), o qual reduz a dimensionalidade e, consequentemente, dificuldades da predição devido às correlações entre as variáveis de entrada.

Os modelos estão foram testados e, como mostrado na Figura 14, o modelo com melhor *accuracy* (0,833 ou 83,3%) foi o Ensemble. Este se trata da união de diferentes modelos de aprendizado de máquina para uma predição, que neste caso foram árvores de decisão impulsionadas. Escolhido o modelo, este foi treinado novamente, agora configurado como um Ensemble otimizável, que otimiza os parâmetros (escolhendo "*Optimizable Ensemble*" na aba "*model type*"), como mostra a Figura 15. Em seguida, o modelo foi exportado (denominado "wage_ensemble_trainClassifier.m") como uma função para ser utilizado no código principal ("wage_predict.m").

**Figura 12** – Seleção e Configuração dos Dados

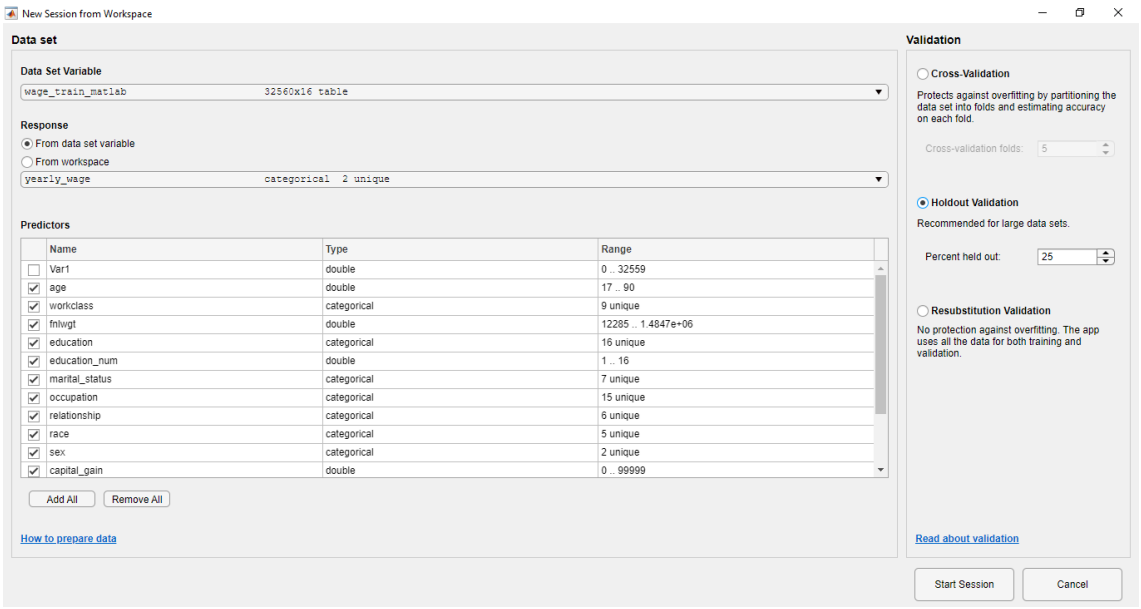



**Figura 13** – Configuração dos Modelos para Serem Treinados

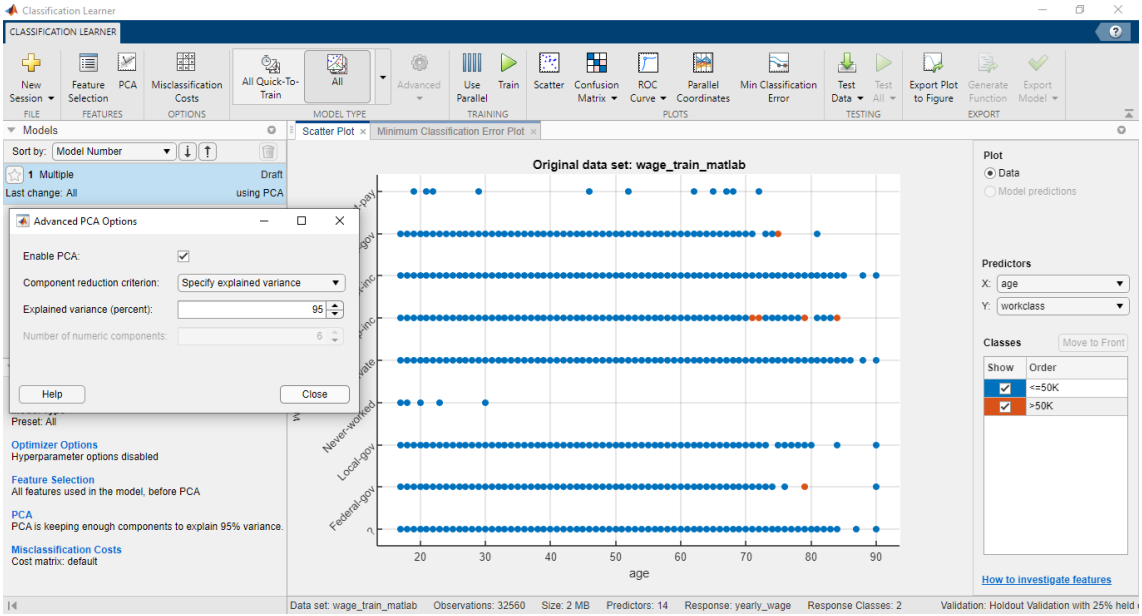


**Figura 14** – Modelos Testados e Respectivos Valores de *Accuracy*

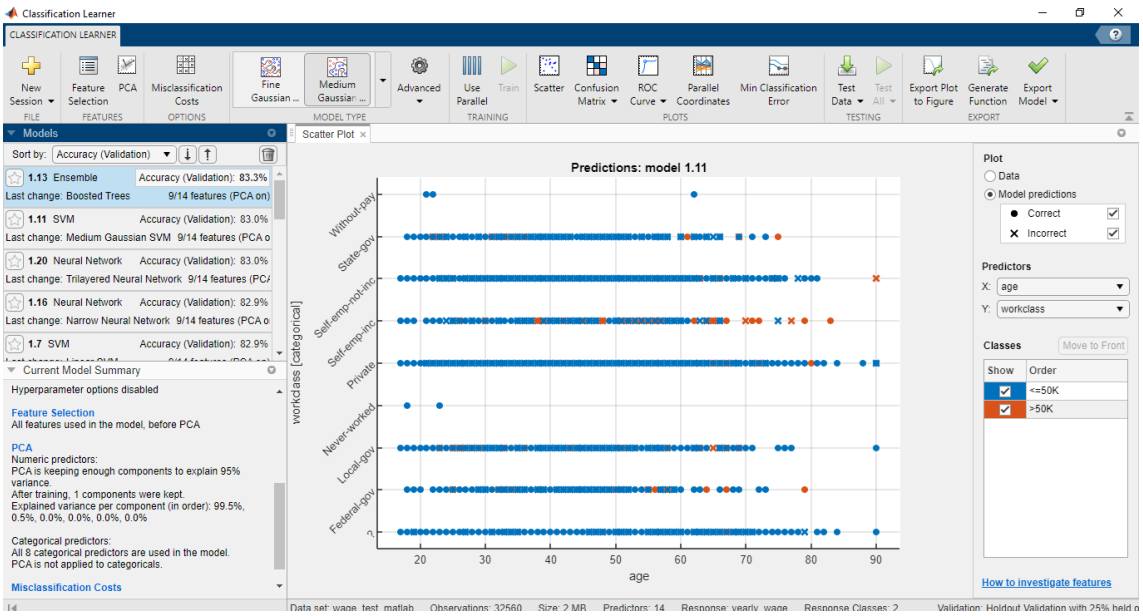
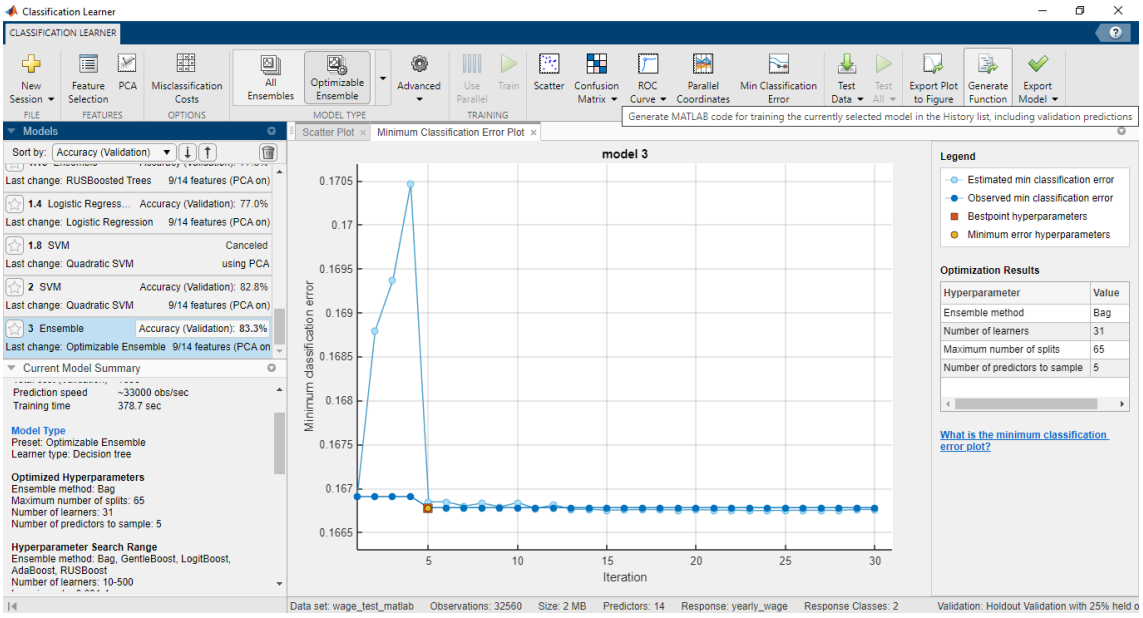



**Figura 15** – Novo Treinamento e Exportação do Modelo Escolhido




Sendo assim, realiza-se a segunda etapa, que se trata da predição propriamente dita, no código “wage_predict.m”. Primeiramente, realiza-se o treinamento do modelo, a partir dos dados de treinamento (“wage_train.csv”) e em seguida a predição, por meio do modelo treinado, a partir dos dados de entrada para a predição (“wage_test.csv”).