**POS IESB 2020**

**Disciplina: Fundamentos de IA**

**Alunos: Eduardo Gomes Piza**

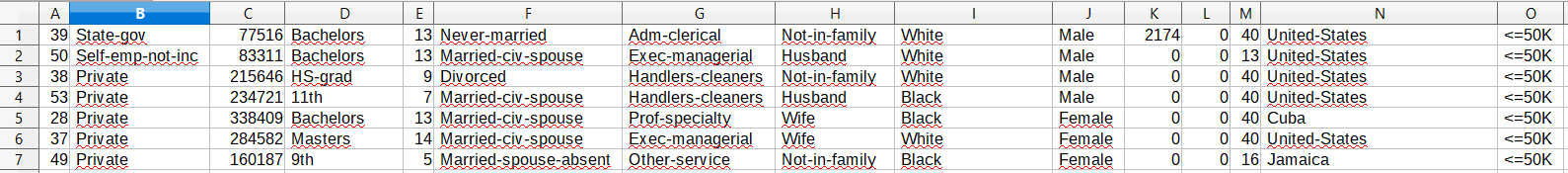
**Valter Takechi Hada**

**Data: 02/09/2020**

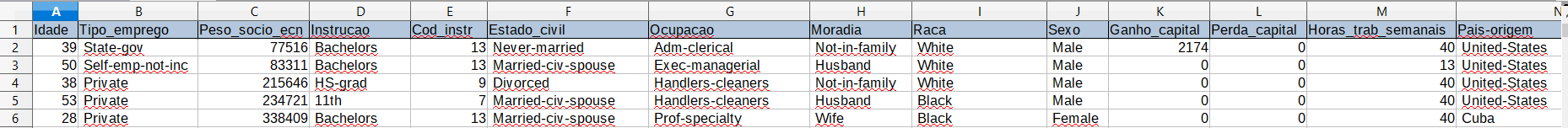
**Projeto Final – Determinar se um trabalhador americano(ou nos EUA) ganha mais de U$50.000 ao ano**

A partir do dataset disponibilizado vamos treinar uma rede neural PMC visando estimar a partir de uma série de informações (idade, grau de escolaridade, estado civil, raça, país de origem, etc.) se a pessoa “testada” pois uma renda anual superior a U$50,000.

O dataset utilizado (disponibilizado no site http://archive.ics.uci.edu/ml/datasets/Adult) possui 32.561 regis-tros, cada um deles contendo 15 atributos (sendo o último deles a indicação da renda superior ao valor supra citado). O dataset (***adult.data***) não apresentava originalmente um label para cada uma das colunas:



Nossa primeira intervenção no arquivo em termos de pré-processamento foi acrescentar as colunas (conforme documentação constante no site que o disponibiliza). Após esta intervenção o arquivo assumiu esta forma:



Com a leitura da documentação do dataset (para compreender os atributos) surgiram alguns pontos importantes para a análise:

. em todo as features cujo valor não é numérico havia um “espaço” antes do conteúdo, que foi retirado no pré processa- mento ainda na planilha;

. há atributos nulos no dataset e para o primeiro “ensaio” vamos apenas desprezar estes registros (excluindo os registros diretamente na planilha);

. o atributo *Peso\_socio\_ecn* é importante pois pessoas que vivem em locais com características demográficas similares possuirão este atributo com valor próximo;

. o atributo *Instrucao* (que representa o grau de escolaridade) possui no atributo seguinte (cod\_instr) um código que o representa, então durante o pré-processamento que será realizado no código python vamos desconsiderar este atributo e tratar apenas o seu código de referência;

. o atributo *Perda\_capital* representa “redução”, então no primeiro “ensaio” não vamos alterá-lo, mas faremos testes posteriores substituindo suas informações por “0” e “1”, significando que não houve perda ou que houve perda, respectivamente.

Como o último procedimento descrito nossa base de dados foi reduzida pois excluímos: **1836** registros com ***Tipo\_emprego*** nulo, **7** registros cuja *Ocupacao* esta nula (na verdade, verificando o Tipo\_emprego destes registros, estão sinalizados como pessoas que nunca trabalharam, e entendemos que estes registros não deveriam constar pois se não trabalham, não recebem proventos), e **555** registros com atributo Pais\_origem nulo. Estas exclusões representam **7,4% da base original** (o que nos deixou com **30162** registros para treino).

A continuidade do pré-processamento, já no programa jupyter, vai substituir todas as “strings” por códigos conforme separação abaixo:

. **Tipo\_emprego**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| Federal-gov | Governo Federal | 00 |
| Local-gov | Governo Local | 01 |
| Private | Empresa Privada | 02 |
| Self-emp-inc | Empresário | 03 |
| Self-emp-not-inc | Empresário | 04 |
| State-gov | Governo Estado | 05 |
| Without-pay | Trabalho Voluntário | 06 |

**. Instrucao**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| Preschool | Pré escola | 00 |
| 1st-4th | Primeiro ao quarto ano | 01 |
| 5th-6th | Quinto ao sexto ano | 02 |
| 7th-8th | Sétimo ao oitavo ano | 03 |
| 9th | Nono ano | 04 |
| 10th | Décimo ano | 05 |
| 11th | Décimo primeiro ano | 06 |
| 12th | Décimo segundo ano | 07 |
| HS-grad |  | 08 |
| Some-college | Alguma faculdade | 09 |
| Assoc-voc |  | 10 |
| Assoc-acdm |  | 11 |
| Bachelors | Bacharelado | 12 |
| Masters | Mestrado | 13 |
| Prof-school | Professor Universitário | 14 |
| Doctorate | Doutorado | 15 |

Obs.: na verdade excluiremos a coluna ***Instrucao*** e manteremos apenas ***Cod\_instr***, mas os dados acima representam o “de X para” existente no database.

. **Estado\_civil**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| Divorced | Divorciado | 00 |
| Married-civ-spouse | Casado Civil | 01 |
| Married-spouse-absent | Casado Cônjuge Ausente | 02 |
| Never-married | Solteiro | 03 |
| Separated | Divorciado | 04 |
| Widowed | Viúvo | 05 |

. **Ocupacao**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| Adm-clerical | Administrador | 00 |
| Armed-forces | Forças Armadas | 01 |
| Craft-repair | Mecânico veículos | 02 |
| Exec-managerial | Gerencial | 03 |
| Farming-Fishing | Agricultor/Pescador | 04 |
| Handlers-cleaners | Serviços de Limpeza | 05 |
| Machine-op-inspct | Operador de Máquinas | 06 |
| Other-service | Serviços Gerais | 07 |
| Priv-house-serv | Serviço Privado | 08 |
| Prof-specialty | Especialista Profissional | 09 |
| Protective-serv | Serviço de Proteção | 10 |
| Sales | Vendas | 11 |
| Tech-support | Suporte Técnico | 12 |
| Transport-moving | Transporte | 13 |

. **Moradia**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| Husband | Marido | 00 |
| Not-in-family | Não mora com a família | 01 |
| Own-child | Filhos | 02 |
| Unmarried | Mora sozinho | 03 |
| Wife | Esposa | 04 |
| Other-relative | Mora com outros parentes | 05 |

. **Raca**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| Amer-Indian-Eskimo | Ameríndio/Esquimó (Nativos) | 00 |
| Asian-Pac-Islander | Asiático/Ilhas do Pacífico | 01 |
| Black | Negro | 02 |
| White | Branco | 03 |
| Other | Outros | 04 |

. **Sexo**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| Female | Feminino | 00 |
| Male | Masculino | 01 |

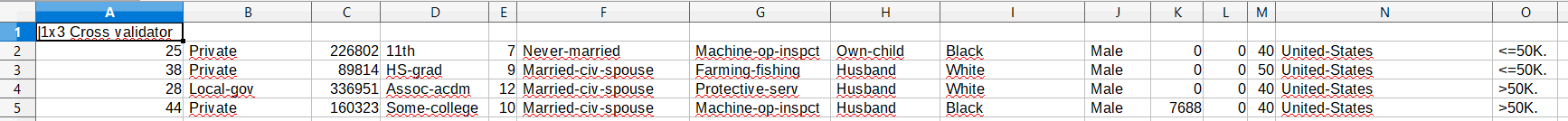
. **Pais\_origem**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| Cambodia | Camboja | 00 |
| Canada | Canadá | 01 |
| China | China | 02 |
| Columbia | Colômbia | 03 |
| Cuba | Cuba | 04 |
| Dominican-Republic | Republica Dominicana | 05 |
| Ecuador | Equador | 06 |
| El-Salvador | El-Salvador | 07 |
| England | Inglaterra | 08 |
| France | França | 09 |
| Germany | Alemanha | 10 |
| Greece | Grécia | 11 |
| Guatemala | Guatemala | 12 |
| Haiti | Haiti | 13 |
| Holand-Netherlands | Holanda | 14 |
| Honduras | Honduras | 15 |
| Hong | Hong-Kong | 16 |
| Hungary | Hungria | 17 |
| India | India | 18 |
| Iran | Irã | 19 |
| Ireland | Irlanda | 20 |
| Italy | Itália | 21 |
| Jamaica | Jamaica | 22 |
| Japan | Japão | 23 |
| Laos | Laos | 24 |
| Mexico | México | 25 |
| Nicaragua | Nicaragua | 26 |
| Outlying-US(Guam-USVI-etc) | Americanos morando fora | 27 |
| Peru | Peru | 28 |
| Philippines | Filipinas | 29 |
| Poland | Polônia | 30 |
| Portugal | Portugal | 31 |
| Puerto-Rico | Porto-Rico | 32 |
| Scotland | Escócia | 33 |
| South | Sul? | 34 |
| Taiwan | Taiwan | 35 |
| Thailand | Tailândia | 36 |
| Trinadad&Tobago | Trinidad e Tobago | 37 |
| United-States | Estados Unidos | 38 |
| Vietnam | Vietnã | 39 |
| Yugoslavia | Iugoslávia | 40 |

. **Renda**

|  |  |  |
| --- | --- | --- |
| Termo no Dataset | Tradução | Código |
| <=50K | Menor igual a 50k anuais | 00 |
| >50K | Maior que 50k anuais | 01 |

Para este dataset foi disponibilizado um outro exclusivamente para testes. O conteúdo do outro dataset foi verificado e ele passou/passará pelos mesmos ajustes e conversões realizados no dataset de treino. A única alteração adicional que será aplicada no dataset de testes é a exclusão da primeira linha da planilha pois apresenta um conteúdo não tratável:



O dataset de testes possuía originalmente **16281 registros** mas teve o tamanho total reduzido para **15060 registros** em virtude das exclusões por nulidade (uma **redução de 7,4% da base de testes**, o mesmo percentual reduzido da base de treinamento).

Para evitarmos problemas com a execução do código Jupyter os 2 arquivos foram disponibilizados em pasta de trabalho pública no GitHub, conforme links abaixo:

**Treinamento:**

<https://github.com/EduPiza/POS_IESB_FUND_IA/blob/master/adult_without_null.csv>

**Teste:**

<https://github.com/EduPiza/POS_IESB_FUND_IA/blob/master/adult_test_without_null.csv>

A primeira execução do treinamento utilizou os parâmetros abaixo:

. epocas = 100

. hidden\_layer\_sizes=(128,128,)

. activation='relu'

. max\_iter=epocas

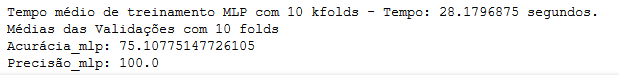
. solver='sgd'

. n\_iter\_no\_change=10

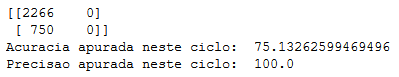
. random\_state=1

. learning\_rate\_init=.001

Este foi o resultado:

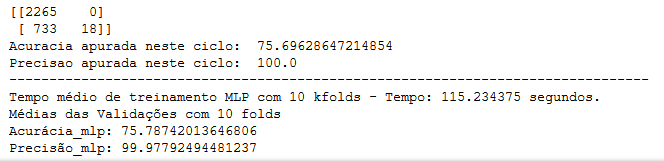


Verificando a matriz de confusão de cada uma das “k” folds (10) verifiquei que todas apresentavam o mesmo resultado:

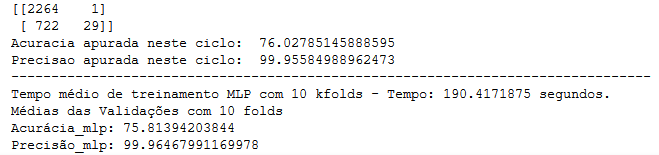


Independente da informação passada o resultado da validação era sempre 1 (que representa salário anual superior a 50k.

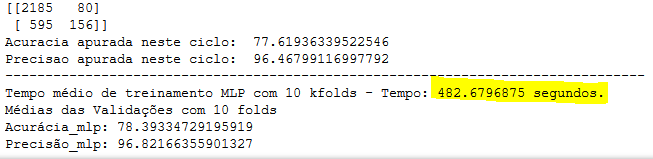
Iniciamos então algumas mudanças nos parâmetros, começando pela função retificadora (substituímos a ‘*relu*’ por ‘*tanh*’). Apesar de levar cerca de 4 vezes mais tempo de processamento esta mudança trouxe impactos positivos e nosso classificador passou a identificar algumas combinações com resultado 0 (salário anual inferior a 50k). Segue o resultado do último ciclo de validação e seu resumo:



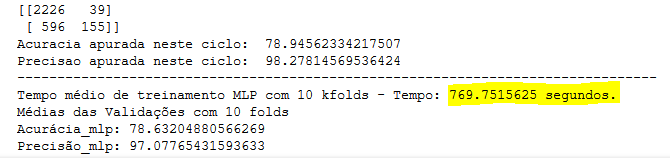
A próxima mudança arquitetural realizada foi a adição de mais uma camada de neurônios, onde passamos a informar *hidden\_layer\_sizes=(128,128,128,)*. Percebemos uma pequena melhoria na distribuição do resultado, e tivemos um incremento de 50% no tempo de execução. Segue o resultado do último ciclo e o resumo de todo o processamento:



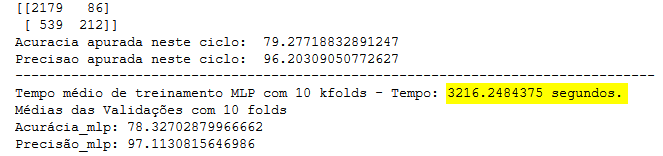
A mudança arquitetural seguinte na rede ocorreu no algoritmo otimizador, onde passamos a informar ‘*solver='lbfgs'*’. Com esta mudança atingimos um nível de acurácia satisfatório, mas ainda apresentando uma quantidade maior de “1s” como resultado (segundo a documentação do database cerca 25% dos registros são de combinações cujo resultado é “0” porém estamos com apenas 8-9% de registros nesta situação, o que mostra que ainda há pontos a melhorar). A mudança do algoritmo aumentou exponencialmente o tempo de execução (cerca de 2 vezes maior que com o parâmetro anterior). Nestas execuções recebemos “warnings” que, consultando sites do assunto, indicaram que seria em virtude do reduzido número de épocas informadas (apenas 100, até esta execução). Segue o resultado do último ciclo e o resumo de todo o processamento:



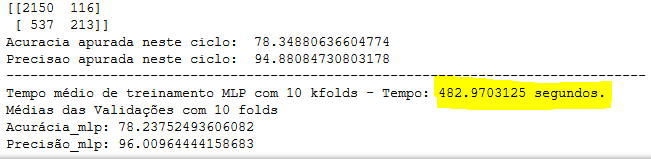
Para a próxima execução o único parâmetro alterado foi o número de épocas, onde o aumentamos de 100 para 400. Percebemos uma nova melhoria, mas ainda sensível. E o tempo de processamento foi de 2 vezes o do processamento anterior:



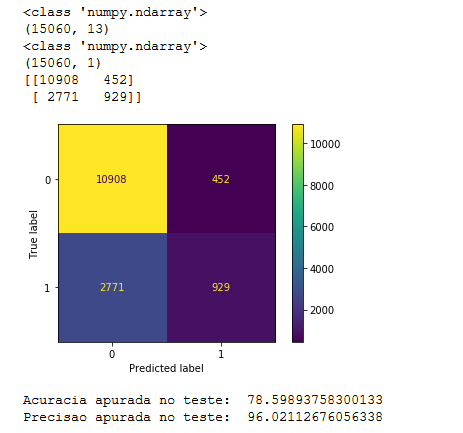
Para a próxima execução alteramos novamente o número de níveis da rede – *hidden\_layer\_sizes=(256,256,256,). O tempo de treinamento subiu para 1h e levou quase 1h para fazer a validação dos dados de teste, e a diferença no resultado foi mínima. Segue o resultado do último ciclo e o resumo de todo o processamento:*

**

*Vamos retornar o número de épocas para 200 e o parâmetro referente às camadas da rede para 128,128,128. Uma última execução com estes novos parâmetros antes de iniciarmos mudanças no dataset utilizado (avaliar a retirada de colunas). O resultado deste último processamento:*



*Mantendo os mesmos parâmetros da rede executamos novamente a predição usando o arquivo de testes. Segue o resultado:*

**

O arquivo “***Projeto\_Final\_Eduardo\_Valter.ipynb***” e os datasets utilizados no estudo foram disponibilizados no GitHub, conforme links abaixo:

**Código Jupyter:**

*https://github.com/EduPiza/POS\_IESB\_FUND\_IA/blob/master/Projeto\_Final\_Eduardo\_Valter.ipynb*

**Datasets:**

***Treinamento*** *-* *https://github.com/EduPiza/POS\_IESB\_FUND\_IA/blob/master/adult\_without\_null.csv*

***Teste***

[*https://github.com/EduPiza/POS\_IESB\_FUND\_IA/blob/master/adult\_test\_without\_null.csv*](https://github.com/EduPiza/POS_IESB_FUND_IA/blob/master/adult_test_without_null.csv)