Universidade São Judas Tadeu

Inteligência Artificial

Profª Gabriela Oliveira Biondi
SIN3AN-MCA3

Henrique Olo Bardeja	81815270
Gabriel Cardoso da Silva	818136132
Iuri Moura da Silva	818143167
Wagner Alves de Melo	818137692
Lucas Venceslau S. Soares	818221111
Danilo Yuudi Hirata	819228171

Projeto - Base Census

❑ Informações sobre a base de dados census:

- **Objetivo:** Prever se a receita das pessoas ultrapassa US $ 50 mil / ano com base nos dados do census.

- **Escopo:** Foram coletados os seguintes atributos para servirem como base para o objetivo do census:

  - Age
  - Workclass
  - Final.weight
  - Education
  - Education.num
  - Marital.status
  - Occupation
  - Relationship
  - Race
  - Sex
  - Capital.gain
  - Capital.loos
  - Hour.per.week
  - Native.country
  - Income

- **Site da Base Census:** UCI Machine Learning Repository

- **Data base :** 01/05/1996

❑ Informações sobre a variável target escolhida e sua justificativa:

- Utilizamos a variável target **"native country"** que diz respeito ao país nativo da pessoa, pois com ela podemos realizar análises demográficas, culturais e socioeconômicas, identificando e projetando resultados estatísticos através das análises geográficas.

❑ Informações sobre o problema e a solução proposta (comitê de classificadores):

- Com base em pesquisas feitas pelo grupo identificamos um grande problema relacionado a desigualdade de renda nos países, enquanto alguns tem uma economia em grande ascensão outros enfrentam diversos tipos de desigualdades seja ela por renda, cultural ou de classe social.

  Através desse levantamento tivemos como objetivo verificar e analisar a qualidade de vida dos moradores nativos dos EUA, que atualmente é a maior economia do mundo em relação aos moradores que nasceram em outros países, com essa premissa conseguimos ter uma amostra significante de casos que refletem o total desnível econômico global entre os países.

  A solução para essa desigualdade são comitês entre os países desenvolvidos, emergentes e subdesenvolvidos para entender qual a melhor estratégia econômica mundial que melhor beneficia os países como um todo.

❑ Informações sobre os procedimentos de pré-processamento realizados:

- Antes de trabalhar com a base, utilizamos os seguintes procedimentos:

- A remoção de dados inválidos e a transformação de dados de texto em representações numéricas.

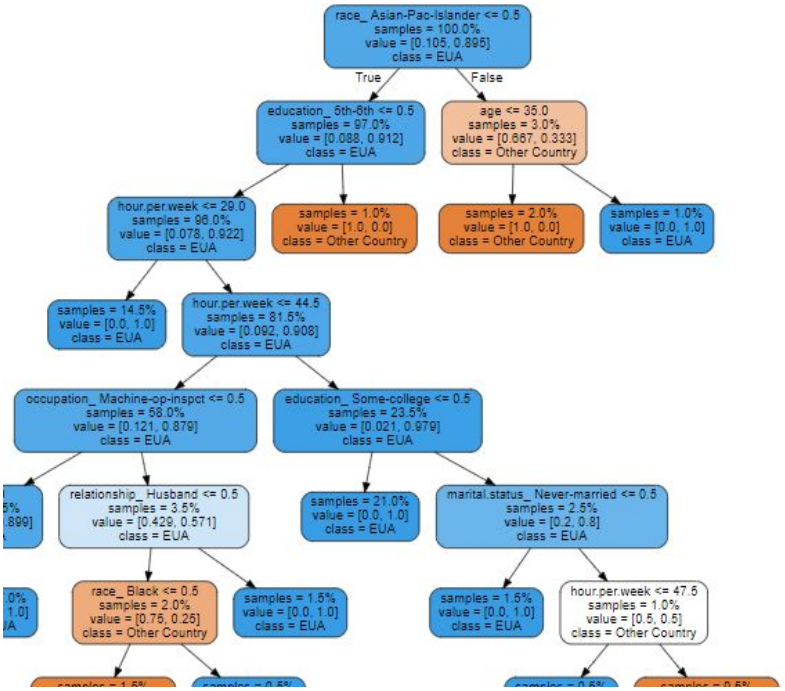
```
df = df[(df.astype(str) != ' ?').all(axis=1)]
df = df.loc[:, ~df.columns.str.contains('^Unnamed')]

native_country = le.fit_transform(df['native.country'])
education = le.fit_transform(df['education'])
workclass = le.fit_transform(df['workclass'])
marital = le.fit_transform(df['marital.status'])
occupation = le.fit_transform(df['occupation'])
relationship = le.fit_transform(df['relationship'])
race = le.fit_transform(df['race'])
sex = le.fit_transform(df['sex'])
income = le.fit_transform(df['income'])
```

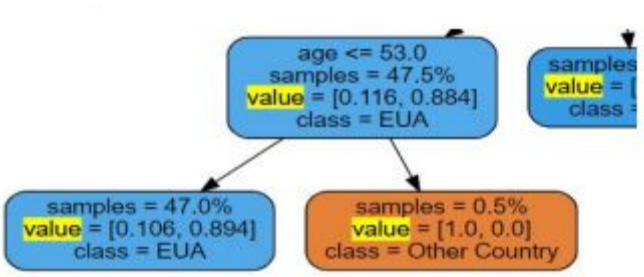
- Criação de uma coluna que classifica se a pessoa é ou não nascida nos EUA

```
df['is_eua'] = df.apply(lambda row: 1 if 'United-States'in row['native.country'] else 0, axis=1)
```

❑ Informações sobre cada um dos algoritmos usados no projeto:

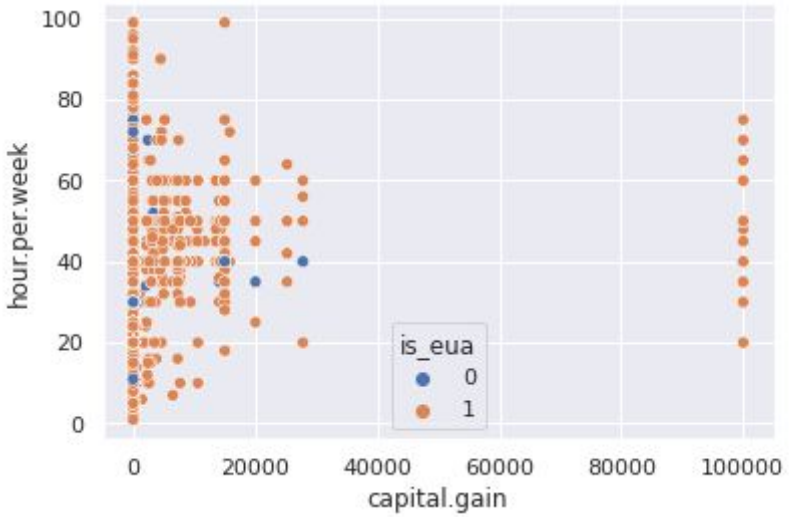
- **Árvore de decisão**
  Foi elaborada uma árvore de decisão, que consiste em um método para classificação e identificação dos registros da base. Segue os exemplos abaixo:





Com base na nossa variável target, no primeiro teste, foi possível notar que Estadunidenses formados no ensino superior, casados, com a idade menor ou igual a 38 anos e meio tem uma média salarial maior do que 50 mil no ano.

- **KNN**
  É um algoritmo de aprendizagem supervisionada que é utilizada nos campos de data mining e machine learning, ele é um classificador onde o aprendizado é baseado “no quão similar” é um dado do outro.



Com esse gráfico gerado a partir do algoritmo, podemos analisar a média de capital ganho e horas trabalhadas dos moradores dos EUA, sendo os círculos laranja pessoas que nasceram nos Estados Unidos e os círculos azuis pessoas de outras nacionalidades.

- **Regressão Logística**
  A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias.

  Utilizamos para verificar a média salarial dos moradores dos EUA, se eles possuem a renda maior ou menor do que 50 mil dólares.

```
from sklearn.metrics import classification_report
print(classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

       <=50K       0.82      0.91      0.86      9115
        >50K       0.59      0.39      0.47      2950

    accuracy                           0.78     12065
   macro avg       0.71      0.65      0.67     12065
weighted avg       0.77      0.78      0.77     12065
```

- **Naive Bayes**
  É uma técnica de classificação baseado no teorema de Bayes com uma suposição de
  independência entre os preditores., ou seja, baseia-se na probabilidade de cada evento ocorrer, desconsiderando a correlação.

  Foram utilizados 3 modelos, sendo eles: Gaussiano, Multinomial e Bernoulli.

  Modelo Gaussiano:

```
from sklearn.naive_bayes import GaussianNB
gaussian = GaussianNB()
gaussian.fit(d_train_att, d_train_gt50)
```

  Modelo Multinomial:

```
from sklearn.naive_bayes import MultinomialNB
multinomial = MultinomialNB()
multinomial.fit(d_train_att, d_train_gt50)
```

  Modelo Bernoulli:

```
# Modelo 3 classificador Bernoulli
from sklearn.naive_bayes import BernoulliNB
bernoulli = BernoulliNB()
bernoulli.fit(d_train_att, d_train_gt50)
```

❑ Tabela comparativa com o desempenho de todos os algoritmos:

Árvore de decisão	93%
KNN	91%
Regressão logística	78%
Naive Bayes	98.39%

❑ Informações sobre a estratégia de avaliação:

- Utilizamos como avaliação dos algoritmos a acurácia para identificar algoritmos que melhor se sobressaem em relação a precisão dos dados, resultados obtidos e estruturação do algoritmo.

❑ Discussão da comparação e conclusão:

- Após debate em grupo conseguimos concluir que todos os algoritmos tem suas particularidades e vantagens, alguns são mais eficientes e mostram uma maior performance em algumas tarefas, já outros demonstram uma precisão e performance menor em comparação com a mesma tarefa.

  Observamos que o algoritmo de classificação Naive Bayes tem ótimo desempenho e obteve acurácias altas em boa parte das tarefas, ele teve maior desempenho em tarefas que não exigiam extrema correlação entre os atributos,

  Entendemos que ele pode ser mais utilizado para classificação de textos, análise de sentimento as redes sociais (identificar se determinado usuário está triste ou feliz através de determinado texto), filtragem de SPAM e etc.

  Os demais algoritmos tiveram bons desempenhos mas como nós havíamos dito, cada um tem sua exceção, por exemplo, algoritmos de árvore de decisão são simples de entender e não precisam de normalização dos dados pois aceitam dados categóricos, número e até faltantes, já o KNN tem vantagem de ter uma implementação simples e trazer resultados muito bons para tarefas classificação e regressão e o algoritmo de regressão logística tem sua vantagem na facilidade de lidar com variáveis dependentes e independentes categóricas, trazendo alta confiabilidade nos seus resultados probabilísticos.

  Portanto, para nossa aplicação da base census definimos os algoritmos através do nível de acurácia que melhor se destaca, mostrando suas vantagens e desvantagens dependendo de qual tarefa vão executar.