UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE INFORMÁTICA

TECNICAS DE MACHINE LEARNING EM PREVISÃO DA POBREZA DOMÉSTICA NA COSTA RICA

TRABALHO 2

Gabriel Thiago - RA: 107774
Sergio Alvarez - RA: 115735

APREND. MAQ. E MODEL. CONHECIM. INCERTO
PROF. DR. WAGNER IGARASHI

SUMÁRIO

SOBRE O PROBLEMA

- Base de dados fornecida pelo Banco Interamericano de desenvolvimento
- Objetivo de prever prever a pobreza no nível familiar
- Classificando em:
  - 1 = pobreza extrema
  - 2 = pobreza moderada
  - 3 = famílias vulneráveis
  - 4 = famílias não vulneráveis

SOBRE O PROBLEMA

- Conjunto de treinamento possui
  - 9.557 linhas
  - 143 colunas,
- Conjunto de teste
  - 23.856 linhas
  - 142 colunas

- Linha: Indivíduo

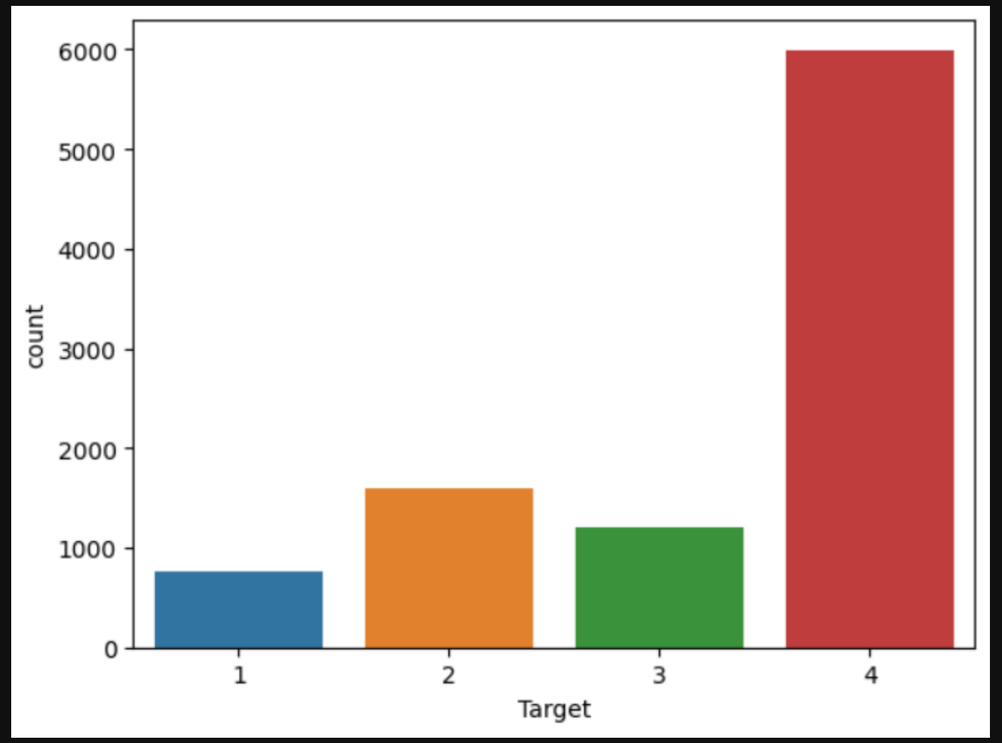
- Coluna: Característica
  - exclusiva do indivíduo ou da família do indivíduo

DADOS

- Base de treinos fornecida pelo Kaggle contém a informação de 9557 pessoas
- Base de testes do kaggle contém 23856 pessoas
- Dados desbalanceados

  1 = POBREZA EXTREMA
  2 = POBREZA MODERADA
  3 = FAMÍLIAS VULNERÁVEIS
  4 = FAMÍLIAS NÃO VULNERÁVEIS

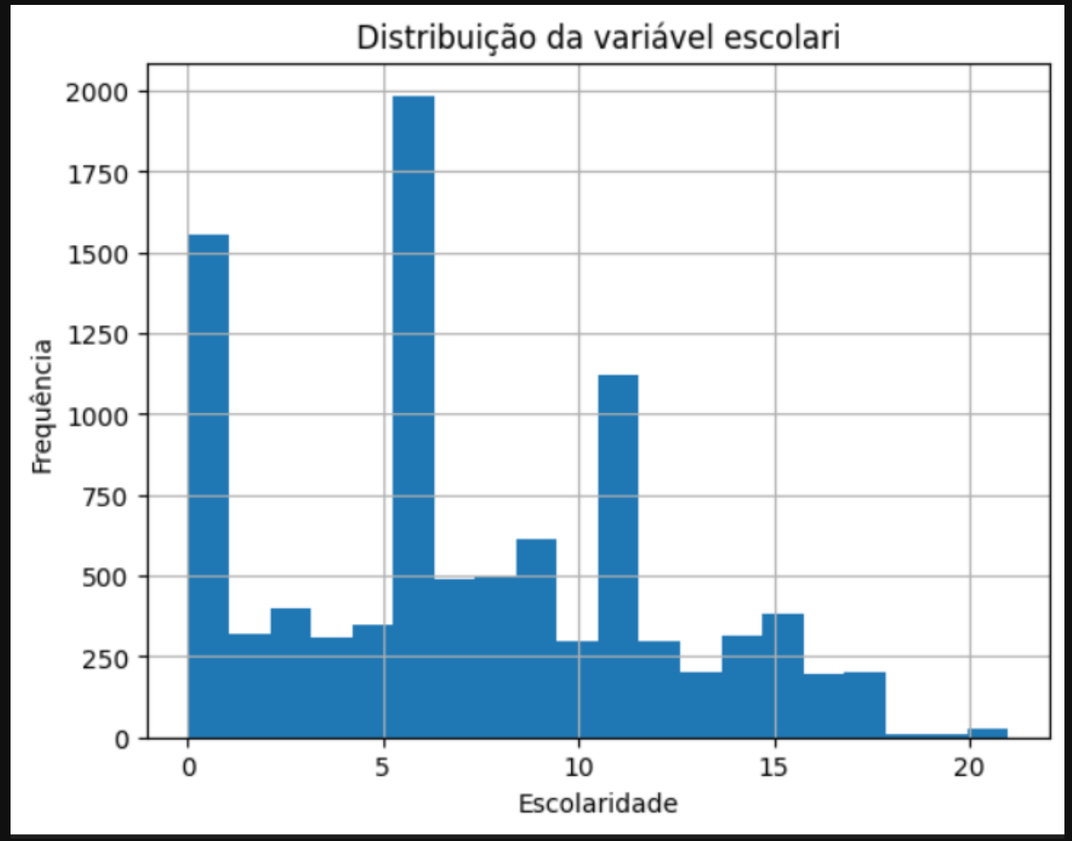
- Dados desbalanceados
  - F1 Score Macro



DADOS

Exemplo de dados:

- v2a1: Valor mensal do aluguel pago pela habitação, 0 se é própria.
- hacdor: Se a casa foi construída pelo dono ou não
- rooms: Número de quartos na habitação
- hacapo: Se a casa é de propriedade da família ou não
- v14a: Se há (ou não) abastecimento de água na casa
- refrig: Se a casa possui refrigerador ou não
- paredblolad: Se a habitação tem parede de bloco/bloco de cimento ou não
- paredzocalo: Se a habitação tem parede de pedra, barro ou madeira ou não
- paredpreb: Se a habitação tem paredes de material pré-fabricado ou não



INFORMAÇÕES

Sobre a solução no Kaggle

- Utilizado para submeter nosso Teste de cada modelo
- Simple Submission = 0.19
- Máximo Score de submissão no problema
  - 0.44

#	△	Team	Members	Score	Entries
1	—	Eric Antoine Scuccimarra		0.44878	140
2	—	yulin pan		0.44785	9
3	—	Reployer		0.44694	76
4	—	CHEN		0.44644	9
5	—	tgvuhn		0.44606	34
6	—	Konstantin V. Grishanov		0.44603	80
7	—	Rashid		0.44510	19
8	—	Mads		0.44508	10
9	—	Ruby Shepard		0.44503	37

PROCEDIMENTO

PASSO 1 - PREPARAÇÃO DOS DADOS

- Células nulas e/ou NAN
  - Filtro com funcão "isnull" do pandas para listagens
    - train.isnull().sum().sort_values(ascending=False

  - Correção com "fillna"
    - 0 para somatórios e negação
      - Ex.: famílias que não pagam alugues
    - -1 quando não existir
      - Ex. média de escolaridade dos que não estudaram

PROCEDIMENTO

PASSO 1 - PREPARAÇÃO DOS DADOS

- Células com strings e números
  - Podendo ser transformadas em binários

  - Nova variável binária com cada célula
    - Se sim: 1
    - Se não: 0

  - Exemplo:
    - edjefa: se o possui escolaridade
    - dependentes: se possui dependentes na familia

PROCEDIMENTO

PASSO 1 - PREPARAÇÃO DOS DADOS

- Células com strings e números
  - Não podendo ser transformadas em binários

  - Exemplo:
    - idhogar: ID do domicilio que o indivíduo vive

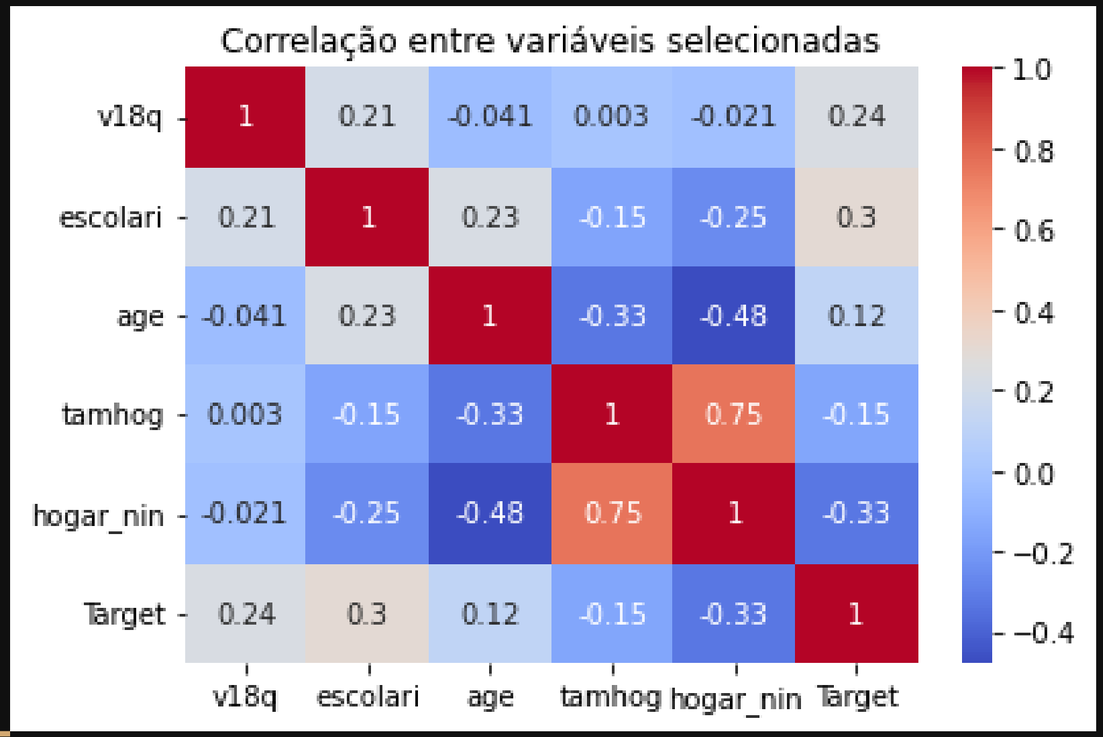
  - Nova variável com ".groupby"
  - Agrupado e contato para cada ID
  - E armazenando o tamanho da familia desde indivíduo nas células correspondentes

PROCEDIMENTO

PASSO 2 - ANÁLISES PARA SELEÇÃO

- Correlação entre variáveis

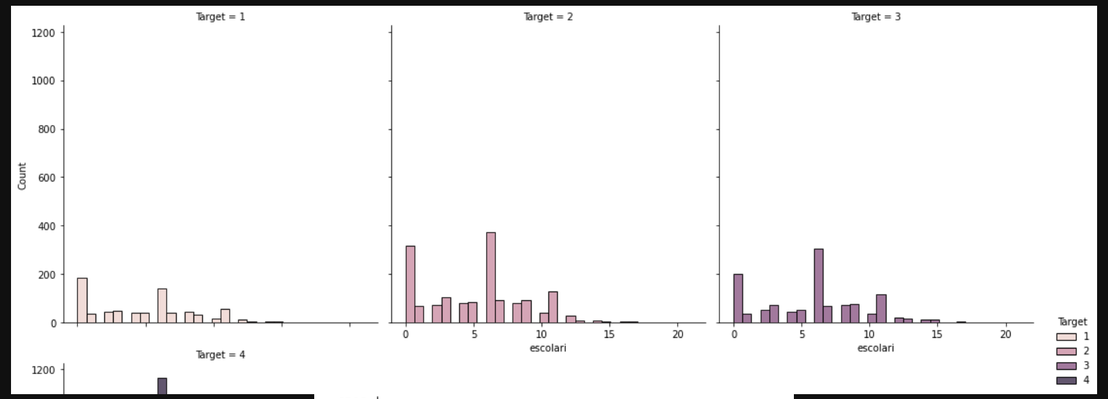
- Variável "hogar_min", "tamhog", "hogar_max", "hhsize" e "hogar_total"
  - Altamente correlacionadas
  - 0.75 correlação
  - Redundante!
    - Podemos deixar apenas uma para uso (escolhida a "hogar_total")



PROCEDIMENTO

PASSO 2 - ANÁLISES PARA SELEÇÃO

- Visualizar histogramas das variáveis separadas por classe para tentar encontrar algum padrão



PROCEDIMENTO

PASSO 3 - DATAFRAMES PARA O MODELO

- Criação de 2 Dataframe a partir do .CVS
  - X: contem apenas as variáveis que desejemos como preditora (features) do modelo
  - y: contem a variável de predição ('Target')
- Normalização dos dados

PROCEDIMENTO

PASSO 4 - APLICAÇÃO DOS MODELOS

RANDOM FOREST

LOGISTIC REGRESSION

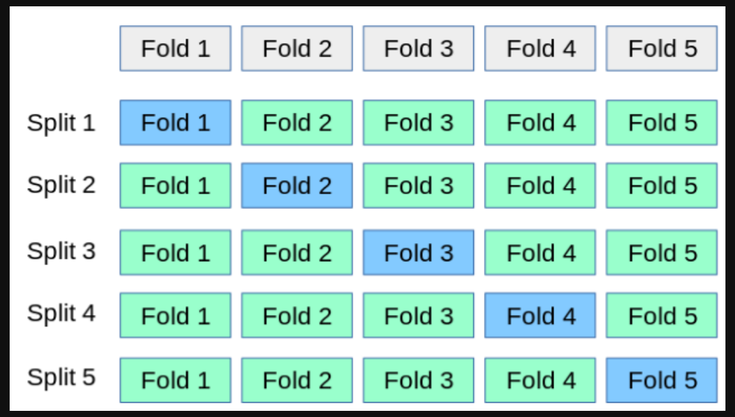
PASSO 5 - AVALIAÇÃO E AJUSTE DO MODELO

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE INFORMÁTICA

AVALIAÇÃO

CROSS VALIDATION E MÉTRICA

- Consiste em dividir o conjunto de dados em k subconjuntos (chamados de folds) e usar k-1 subconjuntos para treinamento e o fold restante para validação.



- O F1 Score é a média harmônica da precisão e do recall e varia entre 0 e 1, onde 1 indica uma performance perfeita do modelo.
- O F1 Score é uma métrica útil quando há um desequilíbrio nas classes

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

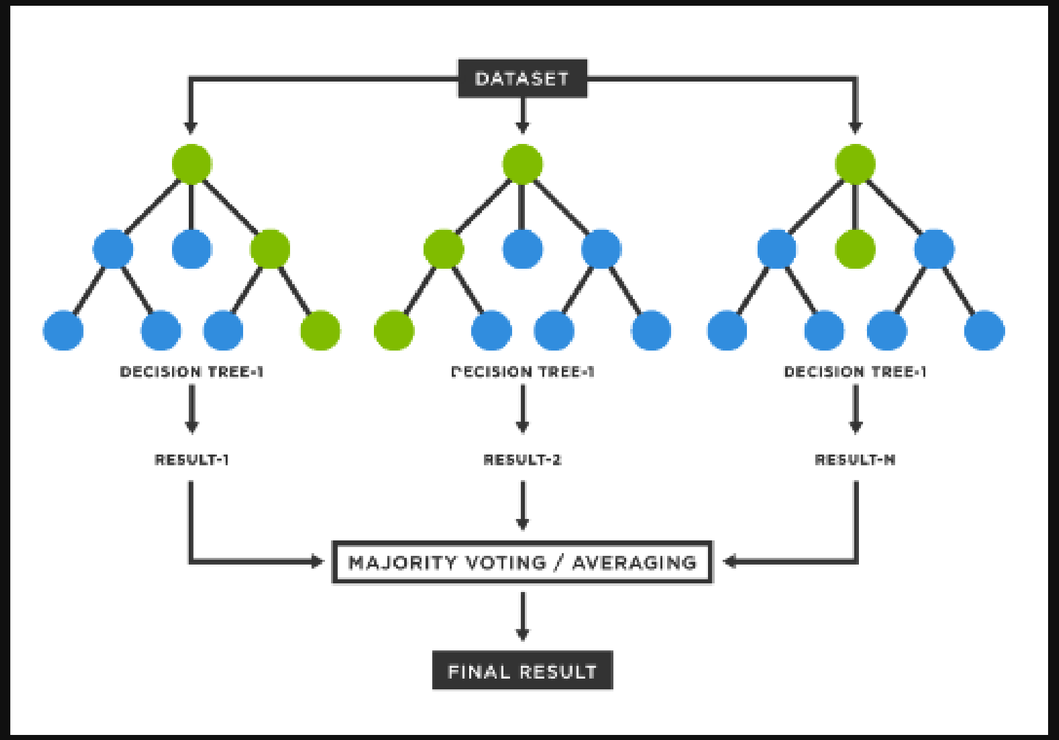
SOLUÇÃO ORIGINAL

- Tecnicas utilizadas de acordo com a pesquisa
  - Support Vector Classifier (Linear SVC) = Score: 0.28346
  - Gaussian Naive Bayes (GaussianNB) = Score: 0.17935
  - Multi-layer Perceptron classifier (MLPClassifier) = Score: 0.28674
  - Linear Discriminant Analysis (LDA) = Score: 0.32217
  - Ridge Classifier CrossValidation = Score: 0.27896
  - KNN
    - KNN with 5 neighbors = Score: 0.35078
    - KNN with 10 neighbors = Score: 0.32153
    - KNN with 20 neighbors = Score: 0.31039
  - Extra-trees classifier = Score: 0.32215

TÉCNICA 1

RANDOM FOREST

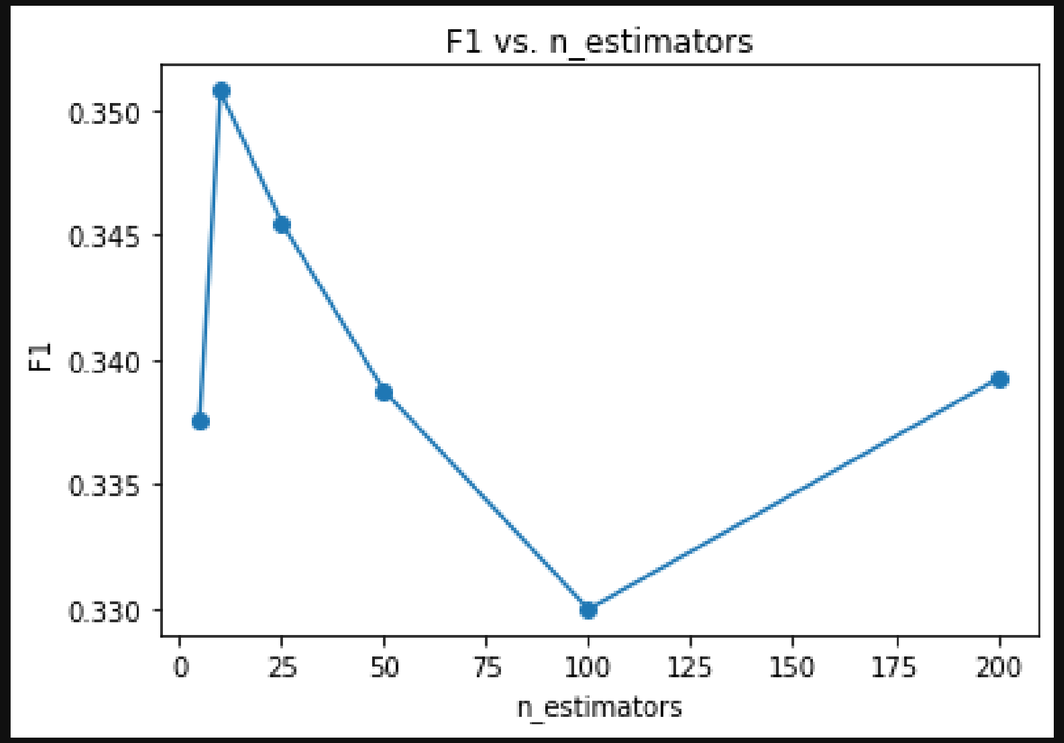
- Combina vários modelos de árvore de decisão para criar um modelo mais poderoso e preciso.
- Seleção aleatória de subconjuntos de dados é feita para reduzir a correlação entre os modelos de árvore de decisão.
- Cada árvore é construída usando um algoritmo de árvore de decisão padrão.
- As previsões de todas as árvores de decisão são combinadas para chegar a uma previsão final.



TÉCNICA 1

RANDOM FOREST

- Melhor Configuração
  - N_estimators = 10
    - [5, 10, 25, 100, 200]
  - Max_Depth = None
    - [None, 5, 10]
- K-fold = 3 F1 Score: 0.30915, Std: 0.02397
- K-fold = 5 F1 Score: 0.33290, Std: 0.05407
- K-fold = 10 F1 Score: 0.35287, Std: 0.04929
- Score final de 0.37798 no Kaggle



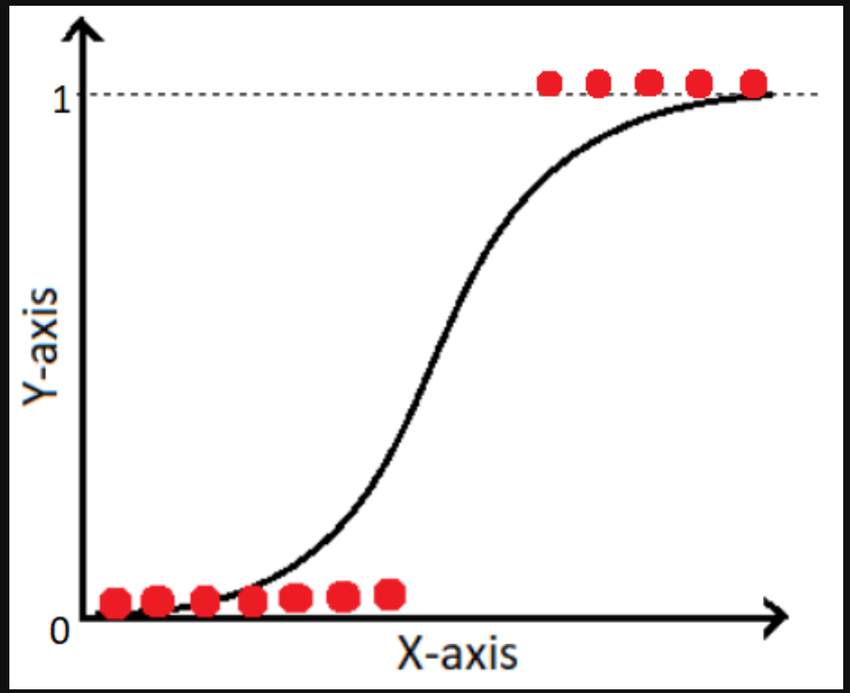
TÉCNICA 2

REGRESSÃO LOGÍSTICA

- Modelo estatístico que permite prever a probabilidade de um evento binário ocorrer
  - Baseado em um conjunto de variáveis preditoras (features)

- Prever se uma família é pobre ou não com base em um conjunto de variáveis explicativas
  - E então a classifica-las

- Simples e pratico de se aplicar



TÉCNICA 2

REGRESSÃO LOGÍSTICA

- Melhores configurações
  - C : 0.3
    - [.001, 0.3, .1]
  - solver : newton-cg
    - ['newton-cg', 'lbfgs', 'liblinear', 'saga']
  - max_iter: 1000
    - [1000, 2000, 5000, 1000]

- F1 (score) = 0.33

K-Folds usado
3, 5 e 10

```
Fitting 3 folds for each of 48 candidates, totalling 144 fits
Tuned Hyperparameters : {'C': 0.3, 'max_iter': 1000, 'solver': 'newton-cg'}
Accuracy : 0.3312214371056125
```

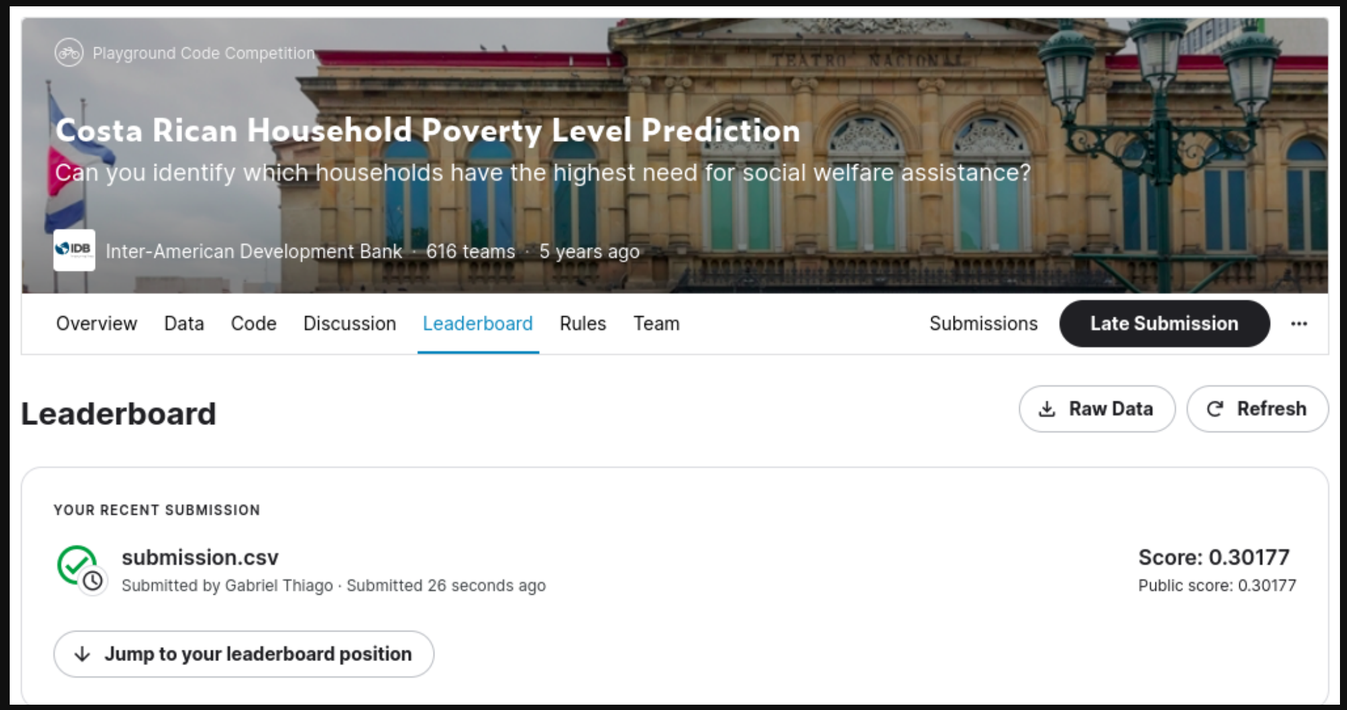
```
Fitting 5 folds for each of 48 candidates, totalling 240 fits
Tuned Hyperparameters : {'C': 0.3, 'max_iter': 1000, 'solver': 'lbfgs'}
Accuracy : 0.32887231031433445
```

```
Fitting 10 folds for each of 48 candidates, totalling 480 fits
Tuned Hyperparameters : {'C': 0.3, 'max_iter': 1000, 'solver': 'newton-cg'}
Accuracy : 0.3305188534056273
```

TÉCNICA 2

REGRESSÃO LOGÍSTICA

- F1 score macro treino
  - 0.33
- F1 score macro teste (Kaggle)
  - 0.30



COMPARAÇÃO

Técnicas Kaggle

Classificador	F1 macro
Support Vector (Linear SVC)	0.28346
Gaussian Naive Bayes	0.17935
Multi-layer Perceptron (MLPClassifier)	0.28674
Linear Discriminant Analysis (LDA)	0.32217
KNN with 5 neighbors	**0.35078**
Extra-trees	0.32215

Técnicas utilizadas

Classificador	F1 macro
Random Forest	0.35287
Regressão Logistica	0.33051

CONCLUSÃO

- Através dos métodos e técnicas de aprendizado de máquina utilizados, conseguimos realizar a classificação da pobreza na Costa Rica com resultados similares aos encontrados no Kaggle.

- Permitiu ampliar nosso conhecimento em tópicos como pré-processamento de dados, seleção de features, modelos de classificação e avaliação de desempenho de modelos.

REFERENCIAS

- Documentação Sklearn, Pandas e matplotlib
- Kaggle notebook "A Complete Introduction and Walkthrough"
- Regressão Logistica - https://www.hashtagtreinamentos.com/regressao-logistica-ciencias-dados
- Como usar o GridSearchCV - https://andersonuyekita.github.io/notebooks/blog/2019/03/21/como-usar-o-gridsearchcv/
- Resolva o Titanic Como um Campeão do Kaggle - https://www.youtube.com/playlist?list=PLwnip85KhroW8Q1JSNbgl06iNPeC0SDkx
- Slides de APREND. MAQ. E MODEL. CONHECIM. INCERTO

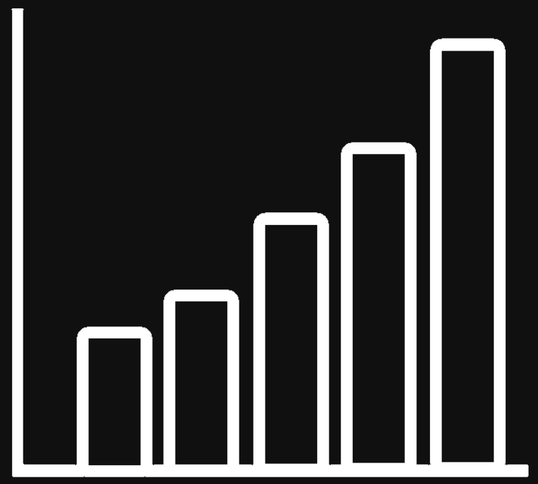
UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE INFORMÁTICA

=)

Obrigado!

Gabriel Thiago
Sergio Alvarez

RA<107774, 115735>@uem.br



APREND. MAQ. E MODEL. CONHECIM. INCERTO
PROF. DR. WAGNER IGARASHI