Modelos de Classificação para determinar a elegibilidade ao Programa Social Bolsa Família.

Discente: Emanuel Flavio Dos Santos Silva
Docente: Débora da Conceição Araújo
Ciência de dados

Programa Bolsa Família e problema abordado

- Bolsa Família
  - Maior programa de transferência de renda do Brasil
  - 2004
  - Combate à fome
- Problema
  - Existem pessoas em situação de vulnerabilidade que não recebem o benefício.
  - Tal como, pessoas que recebem sem está em situação de vulnerabilidade.

Objetivo

- Desenvolver, implementar e avaliar modelos de machine learning para verificar a elegibilidade de indivíduos para o programa social Bolsa Família.

Base de dados

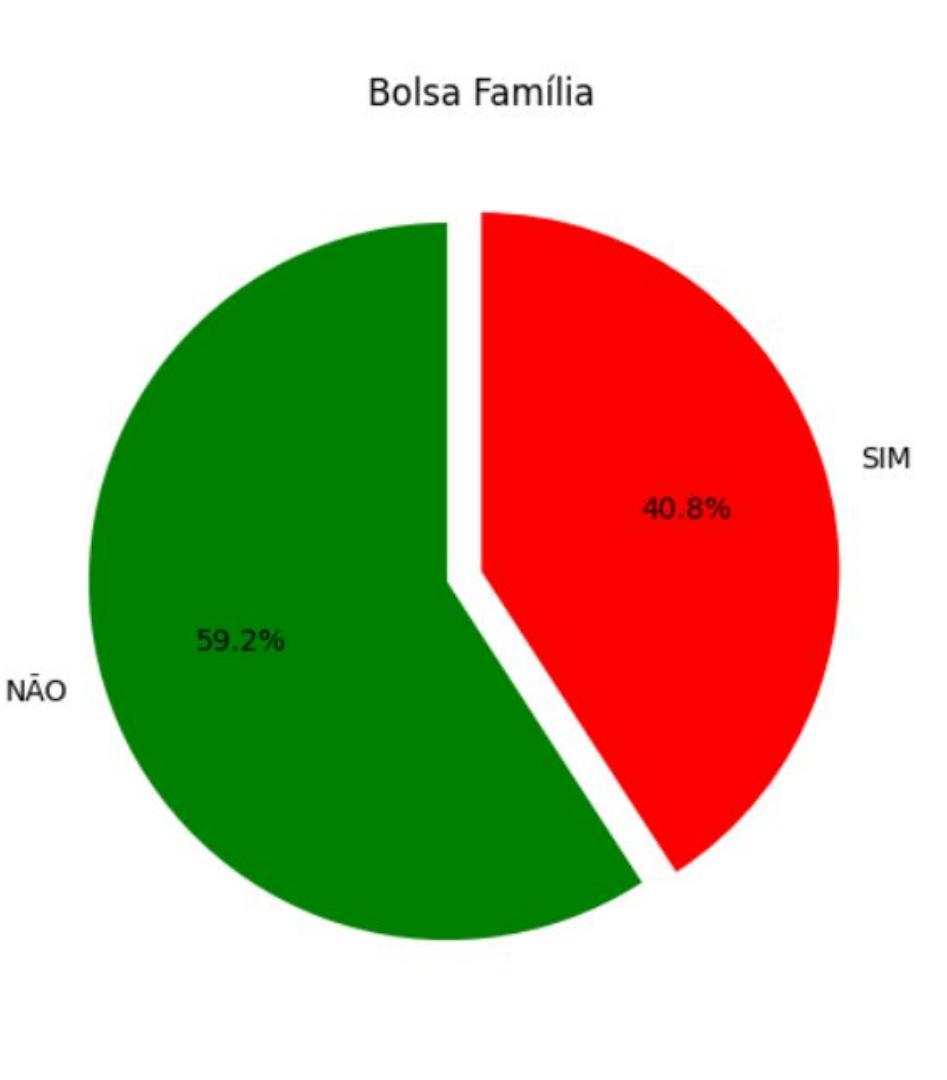
- CadÚnico
- 2019
- Dados de pessoas de Belo Horizonte
- 416.105 dados

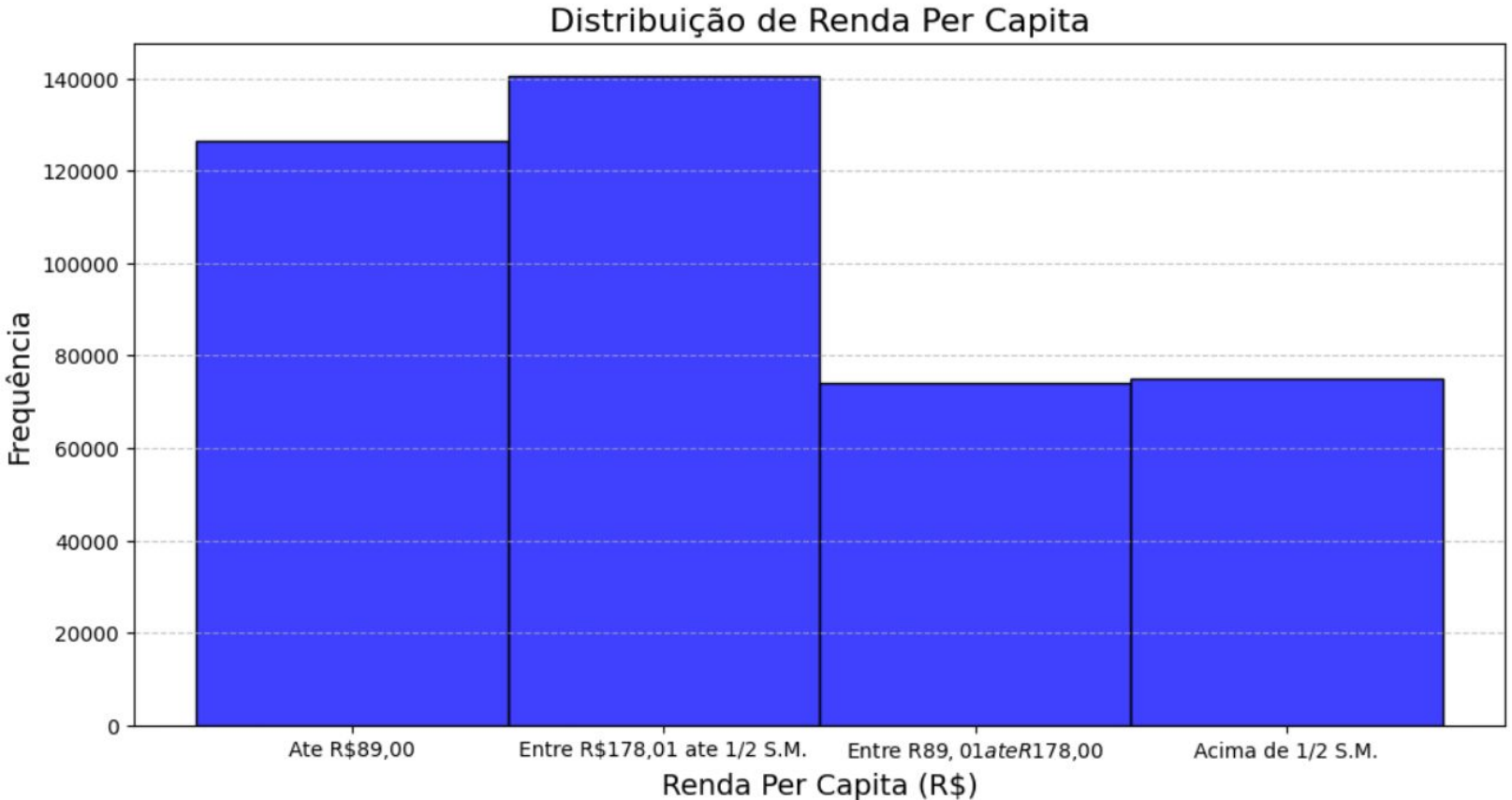
	PARENTESCO_RF	DATA_NASCIMENTO	IDADE	SEXO	BOLSA_FAMILIA	POP_RUA	GRAU_INSTRUCAO	COR_RACA	FAIXA_RENDA_FAMILIAR_PER_CAPITA	V
0	PESSOA RESPONSÁVEL PELA UNIDADE FAMILIAR - RF	20/03/1990 00:00	28	FEMININO	SIM	NAO	Medio incompleto	Parda	Ate R$89,00	
1	PESSOA RESPONSÁVEL PELA UNIDADE FAMILIAR - RF	18/07/1990 00:00	28	FEMININO	SIM	NAO	Medio incompleto	Parda	Ate R$89,00	
2	PESSOA RESPONSÁVEL PELA UNIDADE FAMILIAR - RF	30/01/1997 00:00	21	FEMININO	SIM	NAO	Fundamental incompleto	Preta	Ate R$89,00	
3	FILHO(A)	15/09/1998 00:00	20	FEMININO	SIM	NAO	Medio completo	Branca	Ate R$89,00	
4	PESSOA RESPONSÁVEL PELA UNIDADE FAMILIAR - RF	01/07/1977 00:00	41	FEMININO	SIM	NAO	Fundamental incompleto	Branca	Ate R$89,00	
...	...	...	...	...	...	...	...	...	...	
416103	PESSOA RESPONSÁVEL PELA UNIDADE FAMILIAR - RF	05/02/1955 00:00	63	FEMININO	NÃO	NAO	Sem instrucao	Parda	Acima de 1/2 S.M.	
416104	FILHO(A)	26/06/2000 00:00	18	MASCULINO	NÃO	NAO	Fundamental incompleto	Parda	Entre R$178,01 ate 1/2 S.M.	
416105	FILHO(A)	05/02/2002 00:00	16	MASCULINO	NÃO	NAO	Fundamental incompleto	Branca	Entre R$178,01 ate 1/2 S.M.	

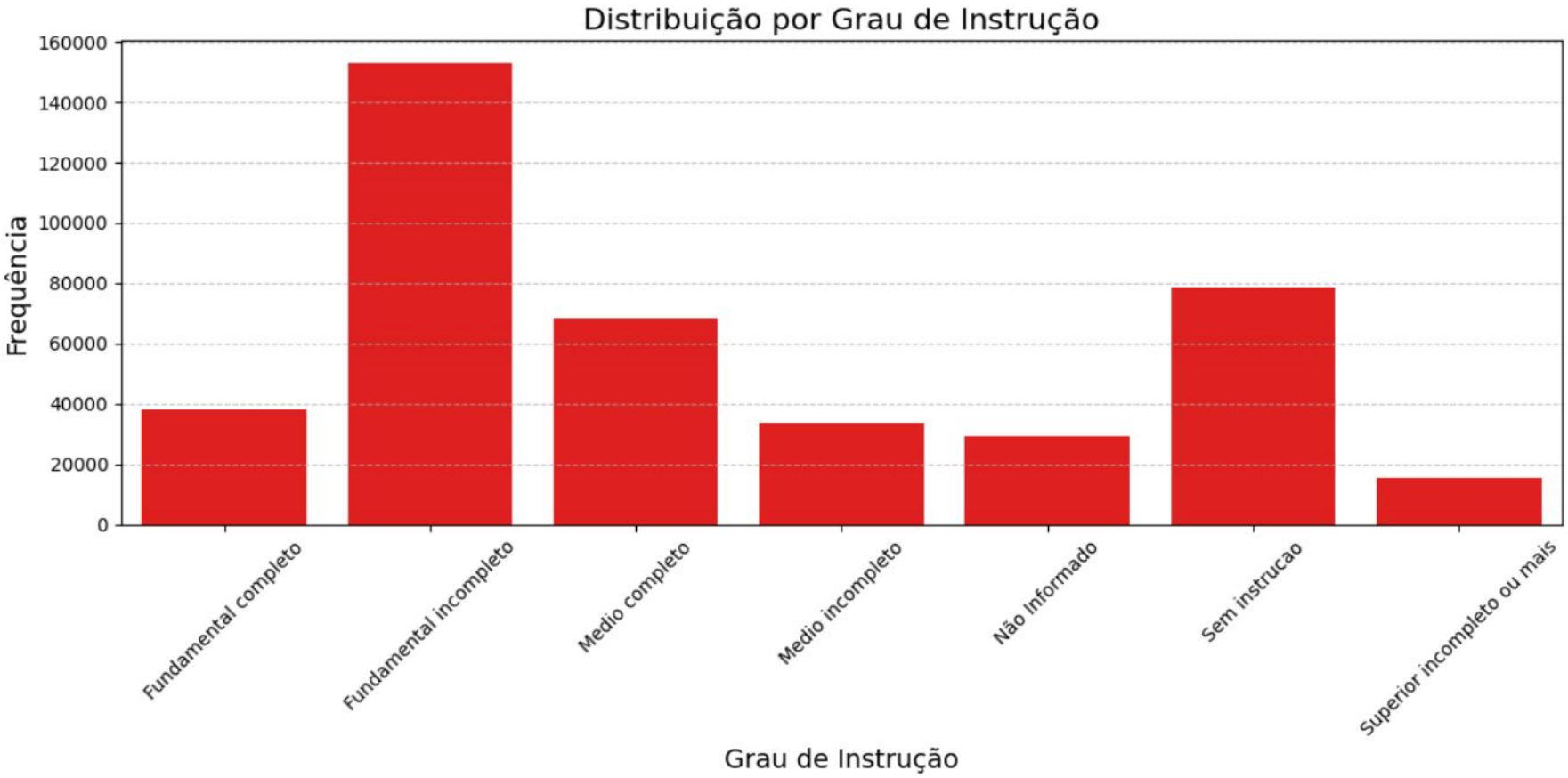
```
0   PARENTESCO_RF                      415136 non-null  object
1   DATA_NASCIMENTO                    416108 non-null  object
2   IDADE                              416108 non-null  int64
3   SEXO                               416108 non-null  object
4   BOLSA_FAMILIA                      416108 non-null  object
5   POP_RUA                            416108 non-null  object
6   GRAU_INSTRUCAO                     416108 non-null  object
7   COR_RACA                           416108 non-null  object
8   FAIXA_RENDA_FAMILIAR_PER_CAPITA    416108 non-null  object
9   VAL_REMUNERACAO_MES_PASSADO        289367 non-null  float64
10  CRAS                               416108 non-null  object
11  REGIONAL                           416108 non-null  object
12  FAIXA_DESATUALIZACAO_CADASTRAL     416108 non-null  int64
13  MES_ANO_REFERENCIA                 416108 non-null  object
```

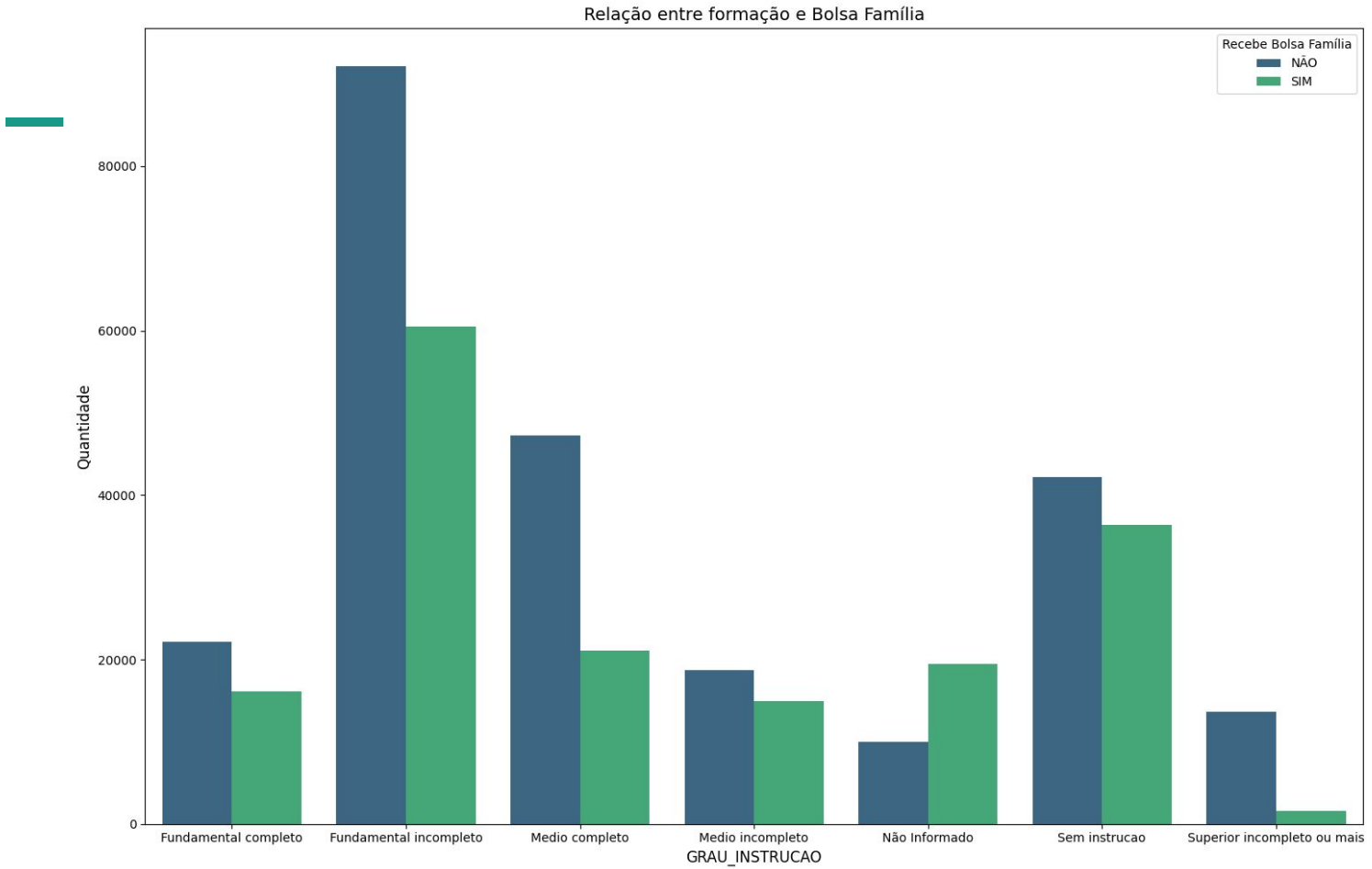
Análise de dados

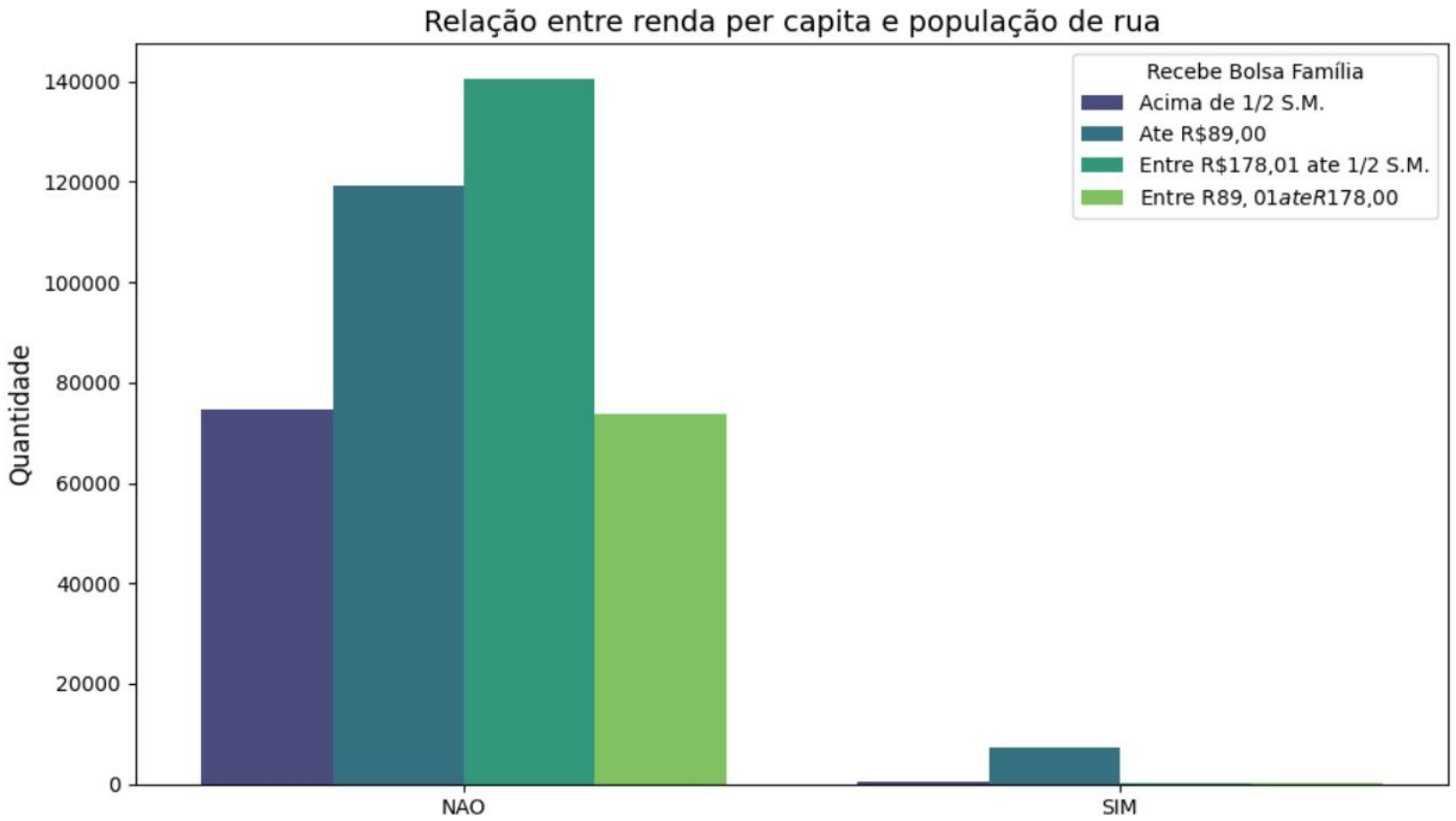
Bolsa família




Distribuição de Renda Per Capita
Frequência
140000
120000
100000
80000
60000
40000
20000
0
Ate R$89,00
Entre R$178,01 ate 1/2 S.M.
Entre R89, 01ateR178,00
Acima de 1/2 S.M.
Renda Per Capita (R$)


Distribuição por Grau de Instrução
Frequência
160000
140000
120000
100000
80000
60000
40000
20000
0
Fundamental completo
Fundamental incompleto
Medio completo
Medio incompleto
Não Informado
Sem instrucao
Superior incompleto ou mais
Grau de Instrução


Relação entre formação e Bolsa Família
Recebe Bolsa Família
NÃO
SIM
Quantidade
0
20000
40000
60000
80000
Fundamental completo
Fundamental incompleto
Medio completo
Medio incompleto
Não Informado
Sem instrucao
Superior incompleto ou mais
GRAU_INSTRUCAO


Relação entre renda per capita e população de rua
Recebe Bolsa Família
Acima de 1/2 S.M.
Ate R$89,00
Entre R$178,01 ate 1/2 S.M.
Entre R89, 01ateR178,00
Quantidade
140000
120000
100000
80000
60000
40000
20000
0
NAO
SIM

Técnicas Utilizadas

- Pré-processamento
  - Remoção de features irrelevantes ao problema
  - Codificação de labels categóricas
  - Não houve dados faltantes ou nulos nas features selecionadas
- Validação Cruzada
  - Stratified K-Fold
  - GridSearch
- Classificadores
  - Random Forest
  - Regressão Logística
  - Naive Bayes
  - KNN

Resultados

```
=== Random Forest ===
Fitting 5 folds for each of 72 candidates, totalling 360 fits

Melhores Hiperparâmetros para Random Forest: {'criterion': 'gini', 'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 200}

Cross-Validation Scores:
[0.68409579 0.68437044 0.68596687 0.68663634 0.68447344]

Average CV Score: 0.6851085743712986

Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.75      0.74     73856
           1       0.62      0.59      0.60     50977

    accuracy                           0.68    124833
   macro avg       0.67      0.67      0.67    124833
weighted avg       0.68      0.68      0.68    124833


Accuracy Score: 0.6843783294481427
```

Resultados

```
=== Regressão Logística ===
Fitting 5 folds for each of 6 candidates, totalling 30 fits

Melhores Hiperparâmetros para Regressão Logística: {'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}

Cross-Validation Scores:
[0.66370269 0.6671702  0.66243241 0.66572826 0.66413183]

Average CV Score: 0.6646330787056905

Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.75      0.73     73856
           1       0.60      0.53      0.56     50977

    accuracy                           0.66    124833
   macro avg       0.65      0.64      0.65    124833
weighted avg       0.66      0.66      0.66    124833
```

Resultados

```
=== Naive Bayes ===
Fitting 5 folds for each of 1 candidates, totalling 5 fits

Melhores Hiperparâmetros para Naive Bayes: {}

Cross-Validation Scores:
[0.62053043 0.61694275 0.607656   0.61330358 0.60829113]

Average CV Score: 0.6133447772723372

Classification Report:
              precision    recall  f1-score   support

           0       0.61      0.98      0.75     73856
           1       0.74      0.08      0.15     50977

    accuracy                           0.61    124833
   macro avg       0.67      0.53      0.45    124833
weighted avg       0.66      0.61      0.50    124833


Accuracy Score: 0.6129549077567631
```

Resultados

```
=== KNN ===
Fitting 5 folds for each of 12 candidates, totalling 60 fits

Melhores Hiperparâmetros para KNN: {'n_neighbors': 7, 'p': 2, 'weights': 'uniform'}

Cross-Validation Scores:
[0.65487941 0.6521157  0.65436443 0.65403828 0.65532572]

Average CV Score: 0.6541447086087031

Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.72      0.71     73856
           1       0.58      0.57      0.57     50977

    accuracy                           0.66    124833
   macro avg       0.64      0.64      0.64    124833
weighted avg       0.65      0.66      0.66    124833


Accuracy Score: 0.6559723790984756
```

Conclusão

- Para esse caso, e utilizando a métrica do F1-Score, temos que o modelo Random Forest teve melhores resultados.
- 74% para a classe 0.
- 60% para a classe 1.

Obrigado !