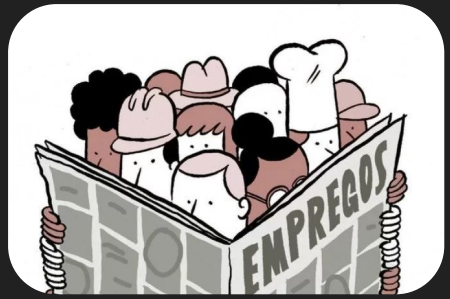
APRENDIZAGEM DE MÁQUINA APLICADA AO *CADASTRO GERAL DE EMPREGADOS E DESEMPREGADOS (CAGED)*

Aluno: Thiago G. Gonçalves

O QUE É O CAGED?



- *É uma atribuição do Ministério da Economia para permitir o auxílio governamental, por exemplo, com benefícios de seguro desemprego.*

- *Atualmente pode ser usado como fiscalização e como instrumento em auxílio de realocação do trabalhador no mercado de trabalho.*

INFORMAÇÕES SOBRE OS REGISTROS DISPONÍVEIS

Haviam ao todo +400M de registros desde 2003 até maio/2022. Dentre os registros, haviam variáveis como:

- ***Região***
- ***Tempo de trabalho***
- ***Salário***
- ***Raça/cor***
- ***Idade***
- ***Grau de Instrução/Escolaridade***
- ***Tipo de deficiência (se tiver)***
- ***Indicador de trabalho informal***
- ***Indicador de Admissão/Desligamento***
- ***Indicador do Setor de Atividade***
- ***...***

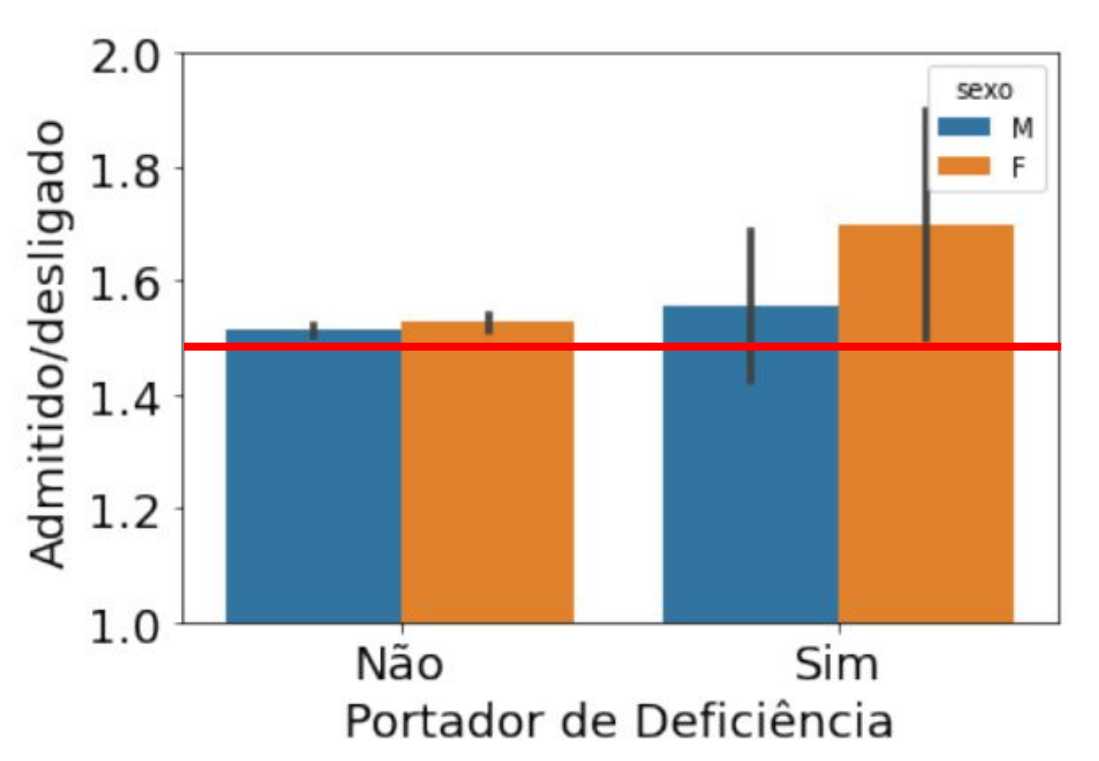
ANÁLISE E MÉTODOS

- *Alguns gráficos foram gerados a partir de uma amostra aleatória de ~15.5k de registros espalhados entre 2007 e 2019, apenas para uma pequena análise exploratória e planejamento dos próximos passos.*

- *Todo o código foi feito em Python no ambiente do Jupyter Notebook, utilizando scikit-learn, pandas, matplotlib, yellowbrick, …*

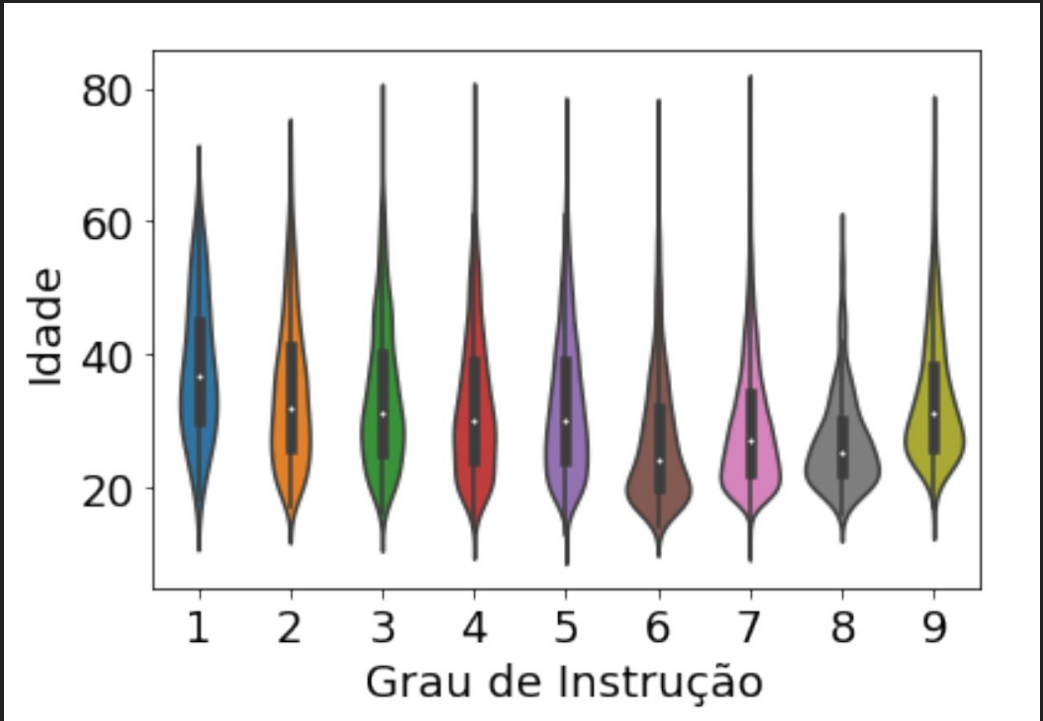
ANÁLISE E MÉTODOS

- Contagem/Soma das contratações (2.0) e desligamentos (1.0) para o total de registros entre 2007-2019.



ANÁLISE E MÉTODOS

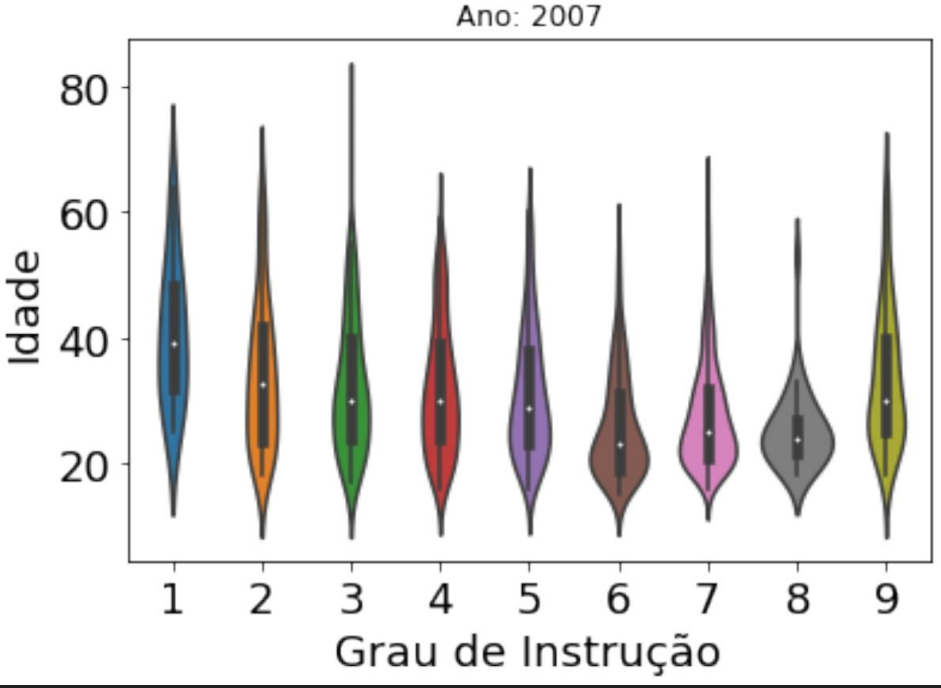
- Registros entre 2007-2019.

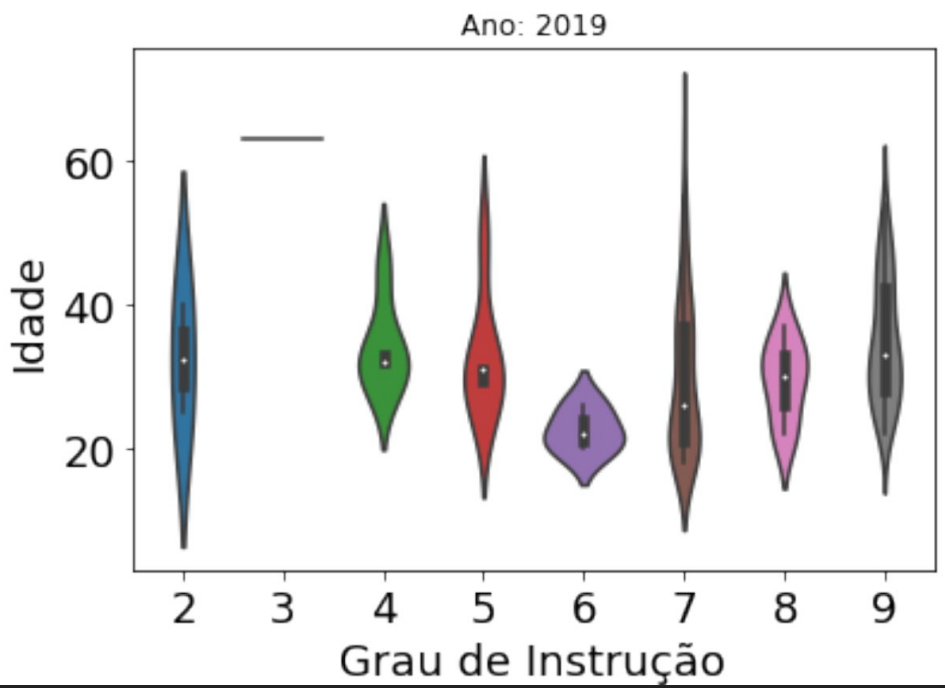


Descrição	Grau de Instrução
Analfabeto	1
Até 5° Ano Incompleto	2
5° Ano Completo	3
6° ao 9° Ano do Fundamental	4
Fundamental Completo	5
Médio Incompleto	6
Médio Completo	7
Superior Incompleto	8
Superior Completo	9
Mestrado	10
Doutorado	11
Ignorado	-1

ANÁLISE E MÉTODOS

- Comparativo filtrando apenas o ano de 2007 e o ano de 2019.





Mapa de calor para os dados ainda não pré-processados entre 2007-2019.

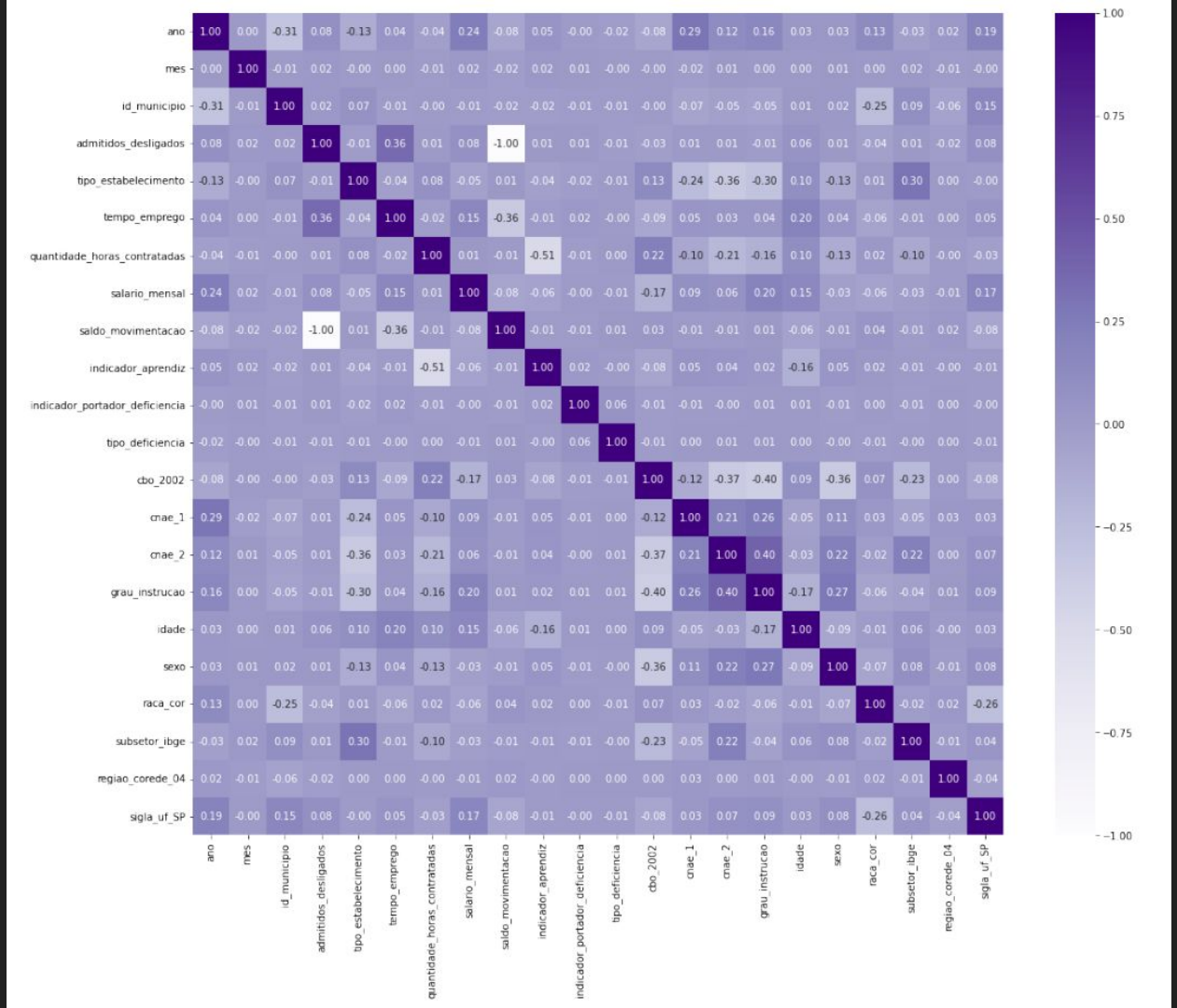
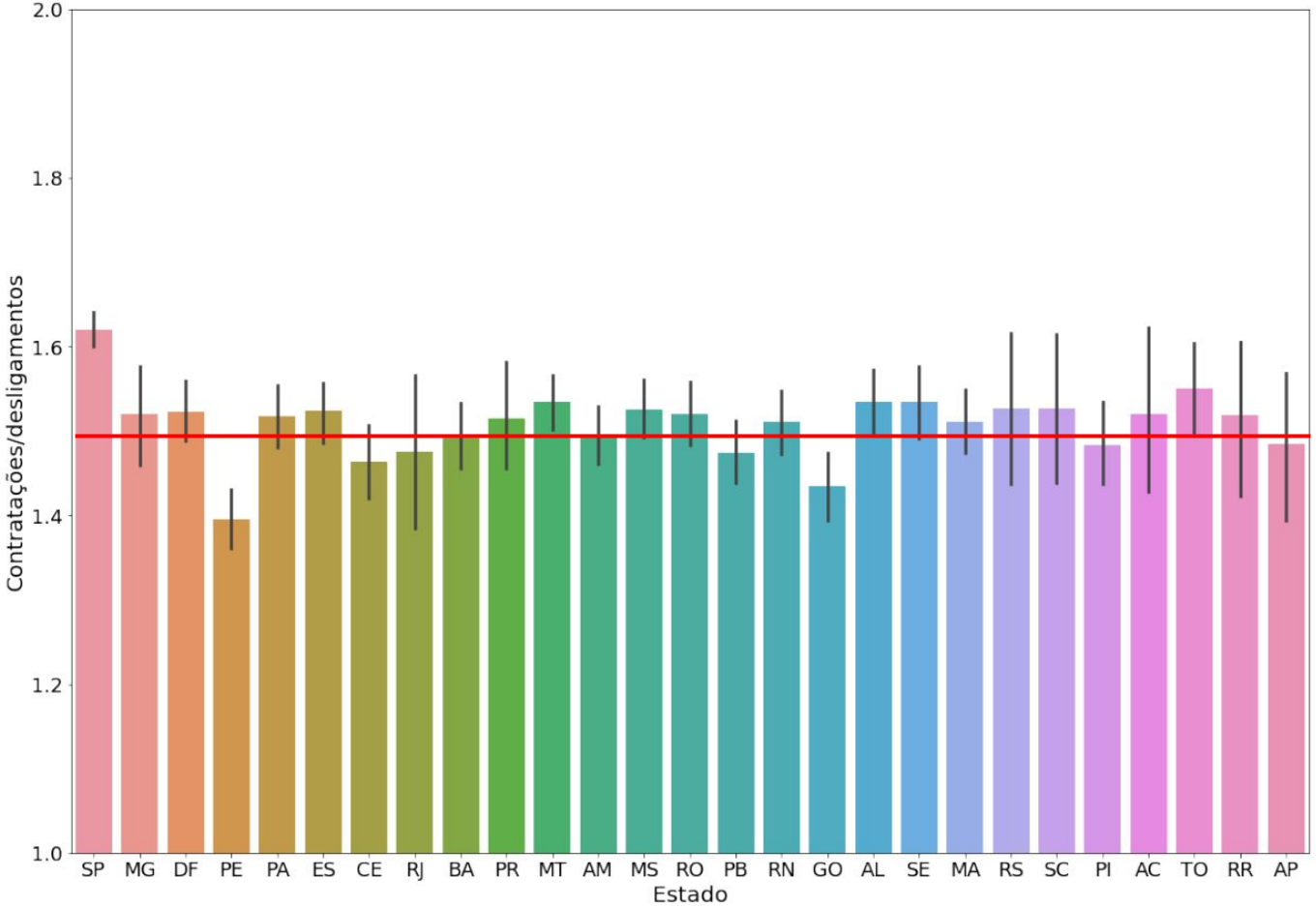


FIGURA 7 – RELAÇÃO ENTRE ADMISSÕES E DESLIGAMENTOS EM CADA ESTADO BRASILEIRO ENTRE 2007 E 2019



ANÁLISE E MÉTODOS

- *Devido a quantidade massiva de registros, foi decidido limitar a aplicação da aprendizagem de máquina a somente um intervalo bem documentado, os anos de 2010 a 2012 (3 anos), que contavam com ~198M de registros, do qual, foi retirado apenas uma amostra aleatória sem reposição de ~15.5k de registros de cada ano (2010, 2011 e 2012), totalizando ~46.5k de registros.*

- *Tomou-se a variável target como sendo o Grau de Instrução, pois seus valores estavam uniformemente distribuídos nos registros (método value_counts() do pandas), diferentemente de variáveis como salário ou tempo de trabalho. Assim, o problema em questão é o de classificação multiclasse (9 ao total).*

LIMPEZA PRÉ-PROCESSAMENTO

- *Remoção de variáveis pouco expressivas/variantes para o problema, utilizando o método de Seleção de Variáveis “GenericUnivariateSelect”.*
- *Tratamento dos dados faltantes interpretando o problema.*
- *Processo de Ordinal Encoding para as variáveis de sigla de estado*
- *Limpeza de alguns dados sujos (conversão de string para int/float).*

VARIÁVEIS UTILIZADAS PELOS ALGORITMOS (21)

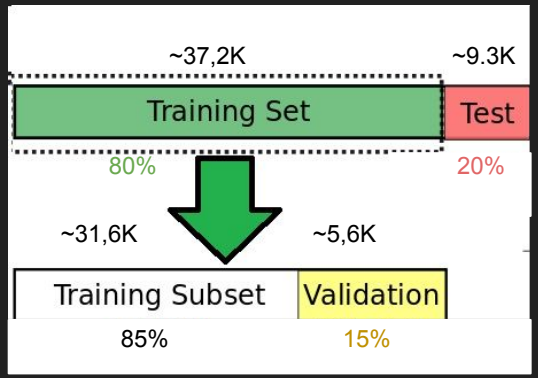
ano
mes
admitidos_desligados
tipo_estabelecimento
tipo_movimentacao_desagregado
faixa_emprego_inicio_janeiro
tempo_emprego
quantidade_horas_contratadas
salario_mensal

saldo_movimentacao
indicador_aprendiz
indicador_portador_deficiencia
tipo_deficiencia
cnae_2
grau_instrucao
idade
sexo
raca_cor
subsetor_ibge
regiao_administrativas_df
sigla_estado

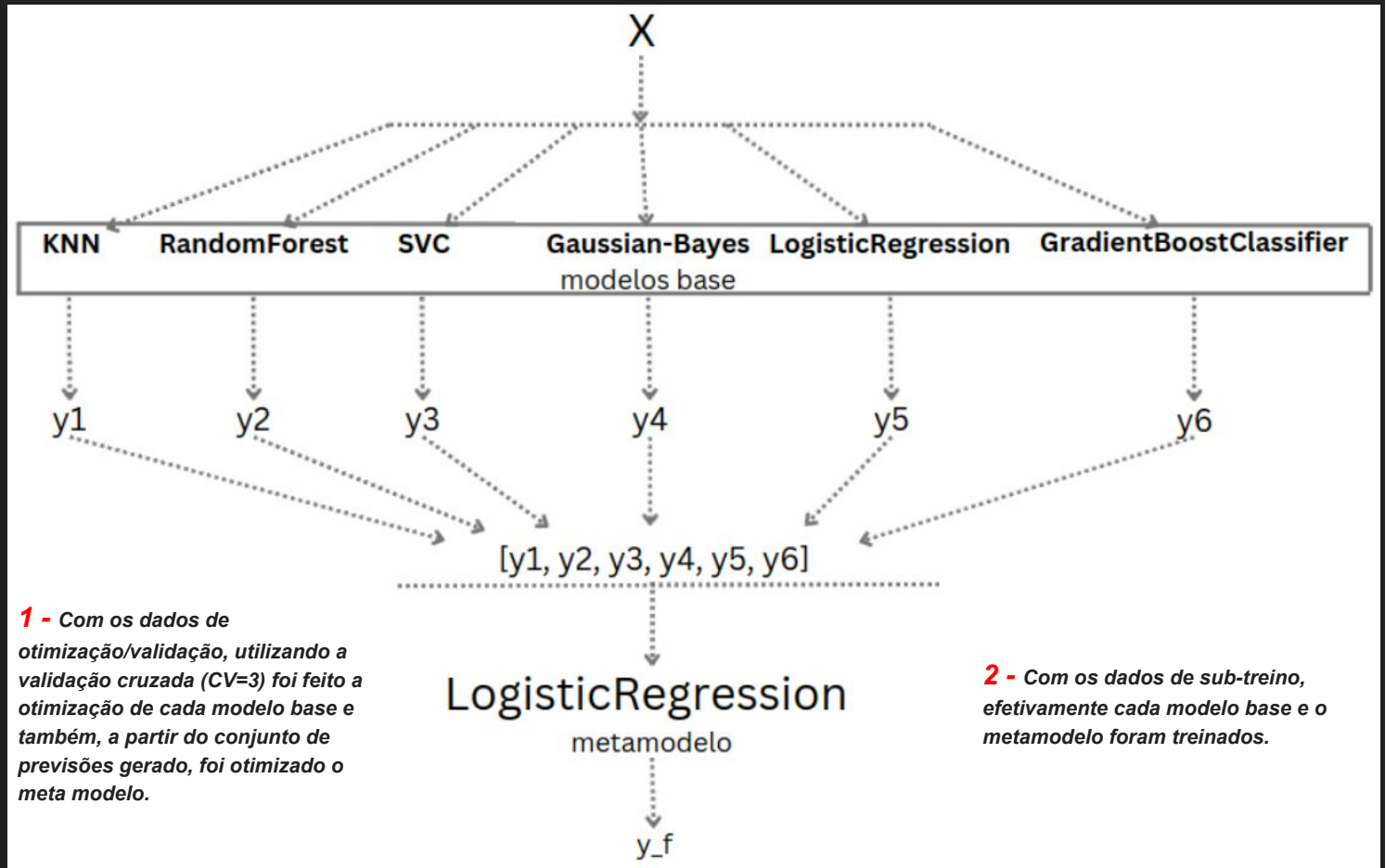
DIVISÃO EM TREINO, VALIDAÇÃO E TESTE

- *Separação dos registros (~46.5k) em treino (80% → ~37,2k) e teste (20% → ~9.3k)*

- *Sobre a amostra de treino (~37,2k) foi feita a separação em amostra de validação/otimização (15% → 5,6k) e o restante efetivamente para treino (sub-treino) (85% → 31,6k).*



STACKING DE ALGORITMOS | OTIMIZAÇÃO E TREINO



1 - ***Com os dados de otimização/validação, utilizando a validação cruzada (CV=3) foi feito a otimização de cada modelo base e também, a partir do conjunto de previsões gerado, foi otimizado o meta modelo.***

2 - ***Com os dados de sub-treino, efetivamente cada modelo base e o metamodelo foram treinados.***

FIGURA 11 – Gráfico Importância de variáveis das 20 principais variáveis do modelo base Random Forest.

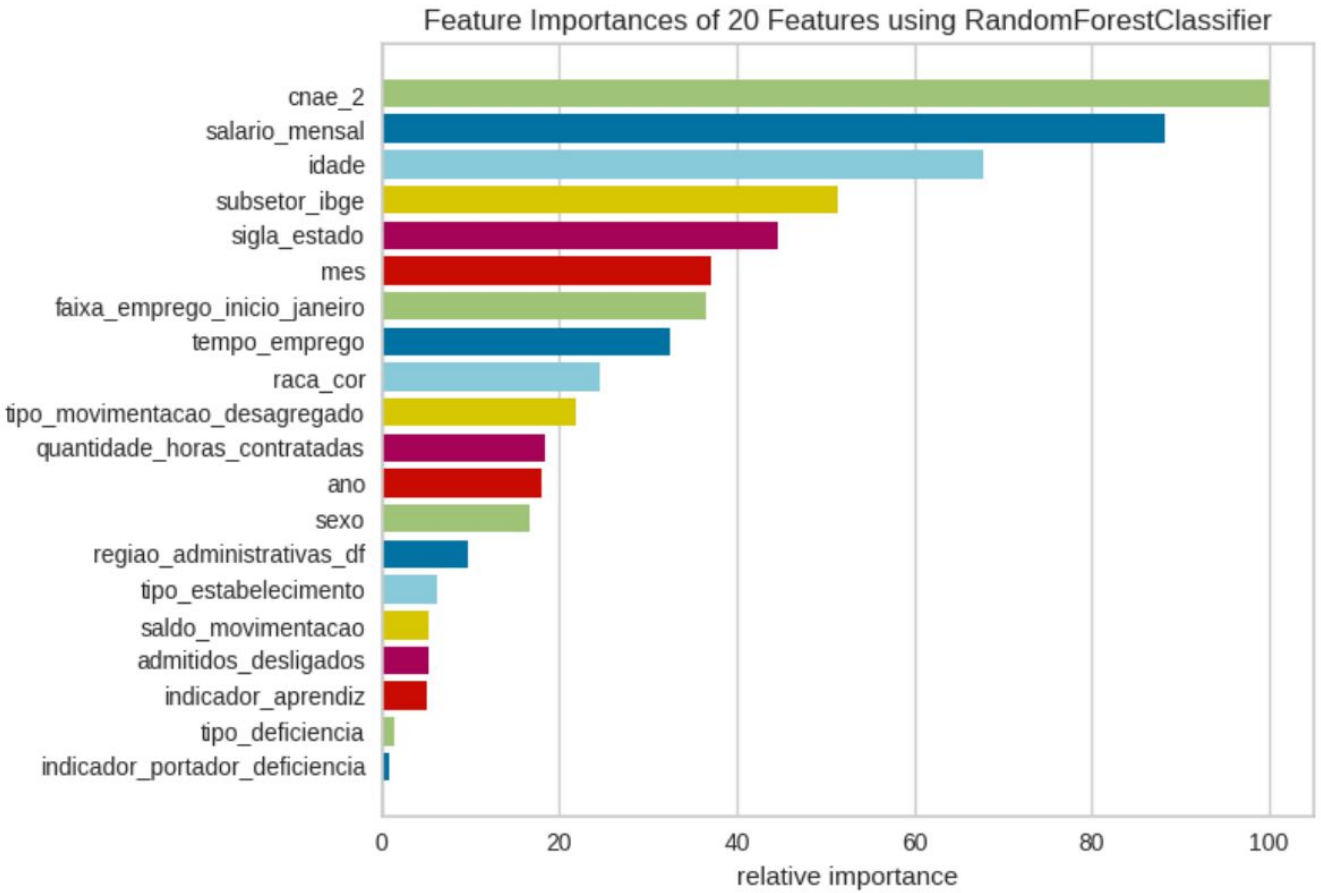


FIGURA 12 – Gráfico Importância de variáveis das 20 principais variáveis do modelo base de Regressão Logística.

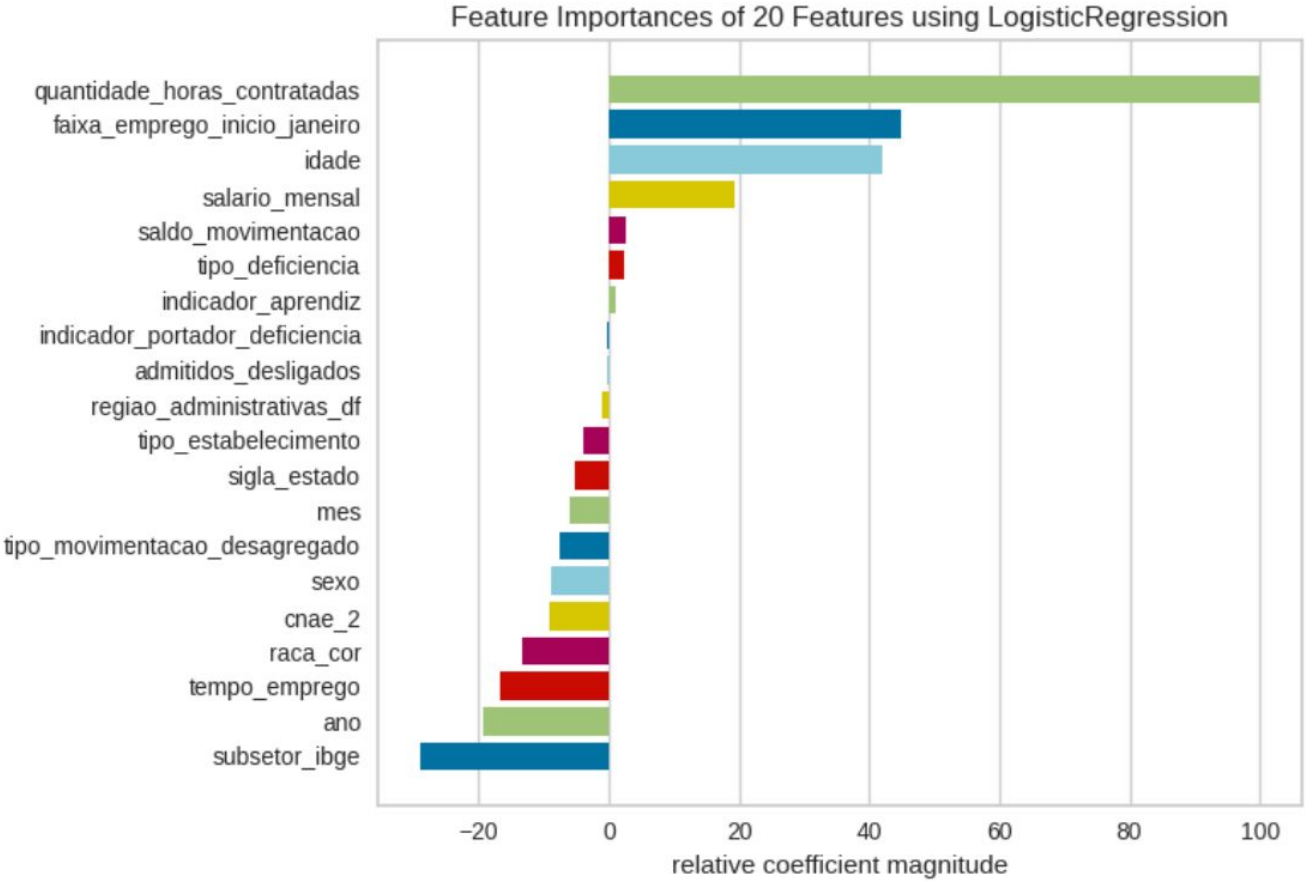


FIGURA 13 – Matriz de Confusão do meta modelo, a intensidade da cor verde indica a maior quantidade numérica de acertos, o ideal seria somente os termos da diagonal preenchidos.

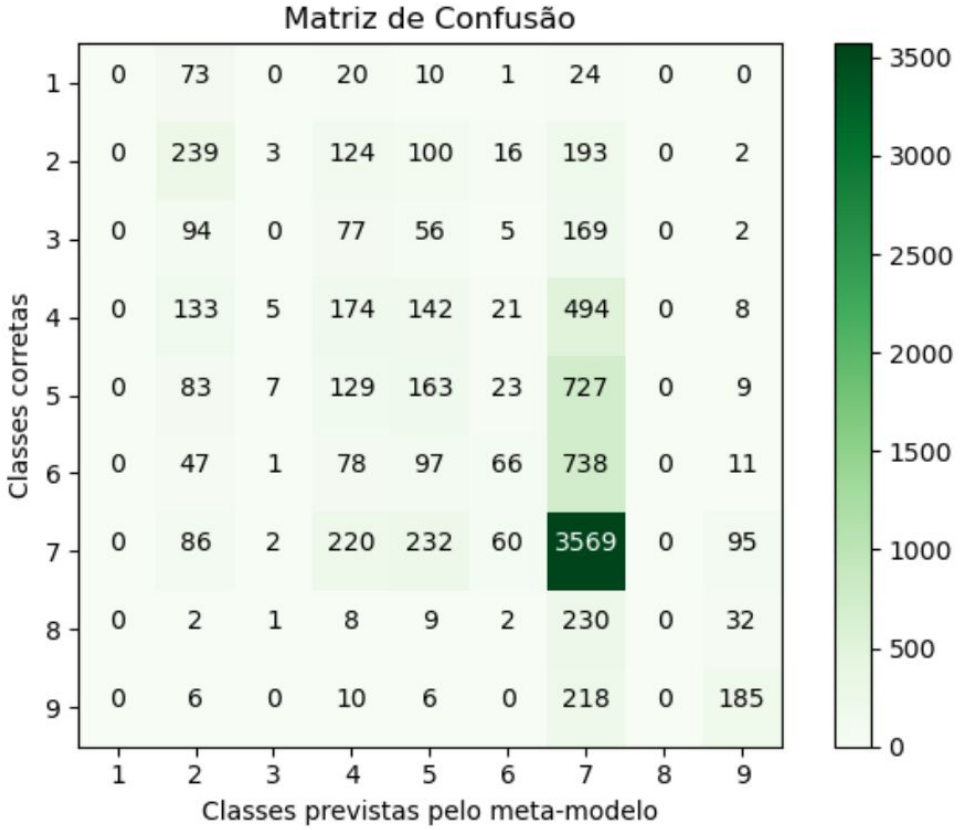


FIGURA 14 – Métricas de desempenho do modelo RandomForest o modelo base de melhor desempenho da stacking que supera o próprio meta modelo.

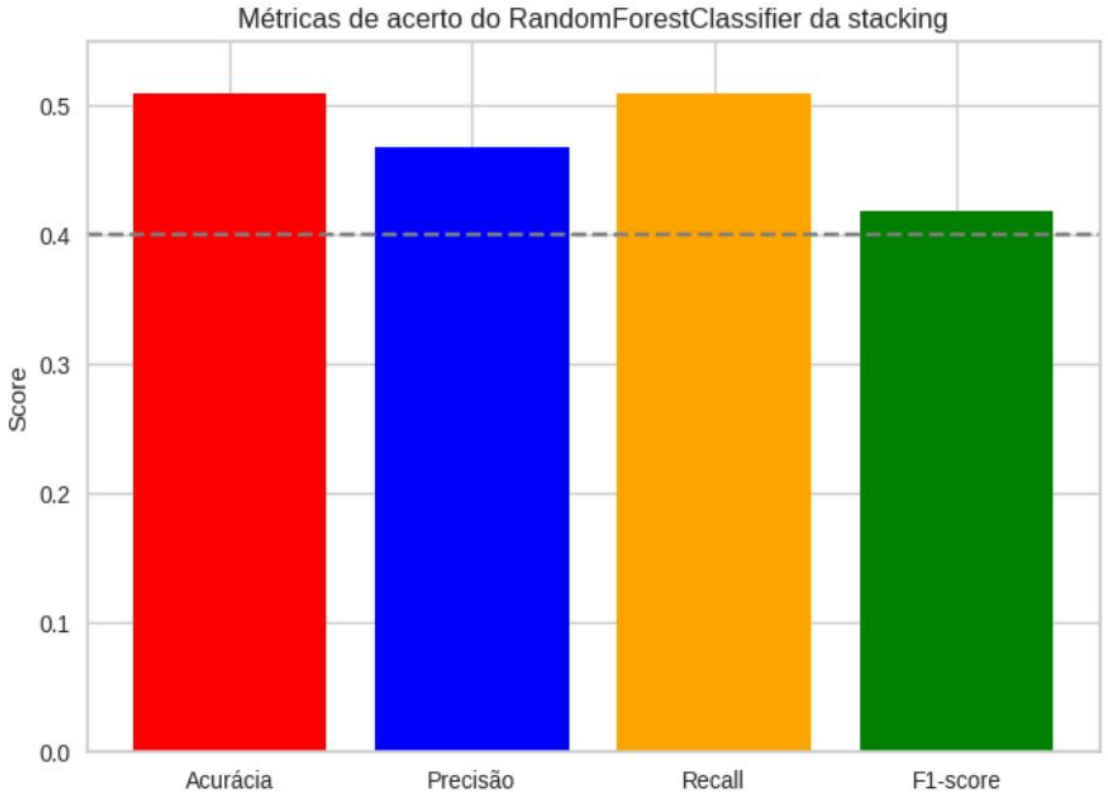
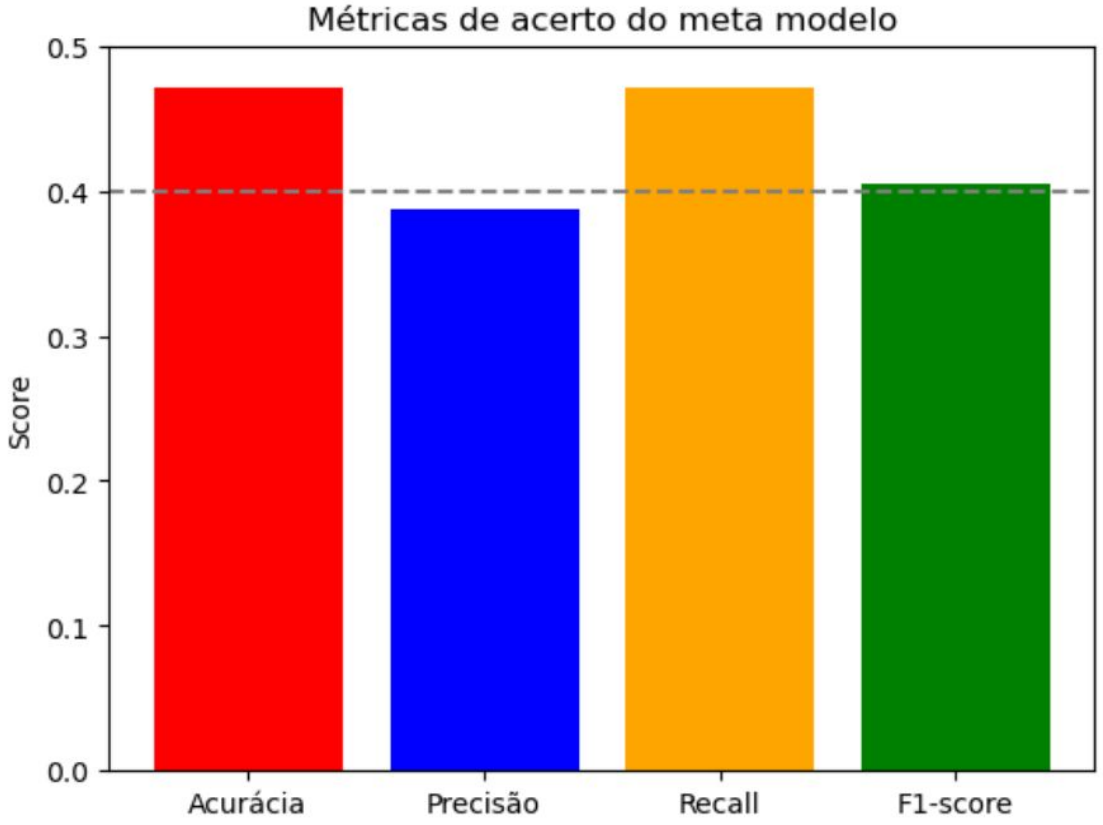


FIGURA 15 – Métricas de desempenho do modelo final (meta modelo), de Regressão Logística. Os valores obtidos são respectivamente 0.47 0.39 0.47 0.41)



ANÁLISE DE RESULTADOS

Embora as métricas obtidas sejam pouco assertivas, e a primeira vista pareçam um tanto decepcionantes, temos de levar em conta:

- *Poucas variáveis disponíveis (21) comparado ao nível de complexidade do problema.*
- *São 9 classes para se prever, e assim, valores de ~50% nas métricas ainda são razoáveis*
- *Os registros estão dispersos ao longo de 3 anos, e portanto, nossa abordagem é uma simplificação, seria necessário a aplicação de um Forecasting para melhores resultados (e ainda assim talvez não funcione tão bem dada a escassez de variáveis)*
- *Abrangemos apenas ~45.5k de registros dentre os ~198M disponíveis só na faixa de 2010-2012 (e +400M no total).*

thiagogonca / **Machine_Learning_CAGED** Public

Pin | Unwatch 1 | Fork 0

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

main | 1 branch | 0 tags | Go to file | Add file | Code

thiagogonca relatorio escrito adicionado — c663637 3 days ago — 3 commits

anos_caged_completo	trabalho final pronto	3 days ago
tabelas_outros	trabalho final pronto	3 days ago
LICENSE	Initial commit	3 days ago
README.md	trabalho final pronto	3 days ago
TRAB_FINAL_MACHINE_LEARNIN…	relatorio escrito adicionado	3 days ago
caged_machine_learning_model.ipynb	trabalho final pronto	3 days ago

README.md

CAGED_model

Modelo de Aprendizado de Máquina aplicado ao banco de dados "Cadastro Geral de Empregados e Desempregados (CAGED)", disponibilizado publicamente em: https://basedosdados.org/dataset/br-me-caged?bdm_table=microdados_antigos

About

No description, website, or topics provided.

Readme
MIT license
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Muito obrigado pela atenção :D