

Trabalho Prático Final de Machine Learning

LUIZA CASTELAR RODRIGUES PEREIRA

Identificar e detalhar o problema que será abordado.

Considerando a base presente em:

<https://dadosabertos.tse.jus.br/dataset/candidatos-2020-subtemas/resource/8187b1aa-5026-4908-a15a-0bf777ee6701>

sobre os dados dos candidatos a vereador, prefeito e vice-prefeito no ano de 2020 para o estado de Minas Gerais, quero prever quais candidatos irão se eleger para os cargos aos quais estão concorrendo nesse mesmo ano, considerando como os dados deles (gênero, cor/raça, grau de instrução, estado civil e cargo ao qual estão concorrendo) influenciam pouco ou muito na eleição.

Para isso, utilizarei aprendizado supervisionado, uma vez que tenho o label (DS_SIT_TOT_TURN0) que diz se o candidato se elegeu ou não.

Compreender os dados e como eles podem ser utilizados para resolver o problema.

A tabela possui **81.785 registros** ao todo, com **63 colunas totais**.

Para responder ao questionamento inicial, irei precisar somente de **16 colunas**.

Separei as colunas entre as colunas que uso para tratar os dados, colunas que uso para explorar mais os dados e coluna que são os atributos do meu questionamento inicial. São elas:

Colunas atributo para a resolução do problema:

- 1) **DS_CARGO** – nome do cargo ao qual o candidato concorre. – STRING
 - Prefeito, vice-prefeito, vereador.
- 2) **DS_GENERO** – gênero do candidato – STRING
 - Masculino
 - Feminino
 - Não divulgável (após as eleições, o candidato pode pedir para anonimizar seus dados)
- 3) **DS_GRAU_INSTRUCAO** – nível de estudo do – STRING
 - Não divulgável (após as eleições, o candidato pode pedir para anonimizar seus dados)
 - Lê e escreve
 - Ensino fundamental incompleto
 - Ensino fundamental completo
 - Ensino médio incompleto

- Ensino médio completo
- Superior incompleto
- Superior completo

4) DS_ESTADO_CIVIL – estado civil do candidato – STRING

- Não divulgável
- Solteiro(a)
- Casado(a)
- Viuvo(a)
- Separado(a)
- Divorciado(a)

5) DS_COR_RACA – cor/raça do candidato (autodeclarado) – STRING

- Amarela
- Branca
- Indígena
- Não divulgável
- Não informado
- Parda

6) DS_SIT_TOT_TURNO – resultado da eleição com relação ao candidato, por turno - STRING

- #nulo
- 2º turno
- Eleito
- Eleito por média
- Eleito por QP (Quociente eleitoral)
- Não eleito
- Suplente

Colunas para exploração maior de como os dados estão distribuídos:

- 1) **NM_SOCIAL_CANDIDATO** – nome social (para pessoas transgeneras e travestis) do candidato que aparece na urna. – STRING
- 2) **TP_AGREMIACAO** – forma como o candidato concorrerá na eleição. - STRING
 - Coligação
 - Partido Isolado
- 3) **DS_NACIONALIDADE** – nacionalidade do candidato. – STRING
- 4) **NR_IDADE_DATA_POSSE** – idade do candidato na data da eleição. – INTEIRO

Colunas de tratamento dos registros:

- 1) **NR_TURNO** – número do turno da eleição – INTEIRO
 - a. 1 – primeiro turno; 2 – segundo turno

- 2) **SQ_CANDIDATO** – número único para cada candidato. – INT
- 3) **DS_SITUACAO_CANDIDATURA** – situação do registro de candidatura do candidato. A opção “cadastrado” jamais vai aparecer, pois o candidato começa com essa opção e depois passa para apto ou inapto, antes da planilha ser gerada, pois todos candidatos passam por análise.
– STRING
 - a. Apto, Inapto, Cadastrado
- 4) **DS_DETALHE_SITUACAO_CAND** – motivo da situação da candidatura – STRING
 - a. Deferido
 - b. Indeferido com recurso
 - c. Renúncia
 - d. Indeferido
 - e. Cassado com recurso
- 5) **ST_CANDIDATO_INSERTIDO_URNA** – informa se o candidato foi ou não inserido na urna – STRING.
 - a. SIM
 - b. NÃO
- 6) **NM_CANDIDATO** – nome completo do candidato - STRING

Todas as variáveis acima citadas são os atributos, exceto a coluna DS_SIT_TOT_TURNO que é o label.

Transformei o arquivo em .xlsx para facilitar a importação para o pandas e para diminuir o peso do arquivo. Subi para o drive para só precisar chamar o arquivo e não fazer upload manual do mesmo.

Realizar a seleção dos dados relevantes (registros e atributos).

Há certas colunas que não são necessárias para chegarmos ao objetivo e/ou tem valores repetidos em todos os registros. Algumas dessas colunas eu explico o porquê descartei, como exemplo:

- 1) **DT_GERAÇÃO** – tem um valor único para todos os registros da base, e **HH_GERAÇÃO** tem 2 valores, porém não precisamos saber quando foi gerado o arquivo em si (sabemos que foi gerado no dia 25/05/2020, às 02:18:09 ou 02:26:12. Não são os horários que cada eleitor votou).
- 2) **ANO_ELEIÇÃO** - valor único para todos registros, de 2020, informação essa que é óbvia também.
- 3) **CD_TIPO_ELEICAO, SG_UE, CD_CARGO, CD_SITUACAO_CANDIDATURA, CD_NACIONALIDADE, CD_GENERO, CD_ESTADO_CIVIL, CD_COR_RACA** uma vez que essas colunas representam os números e as colunas seguintes

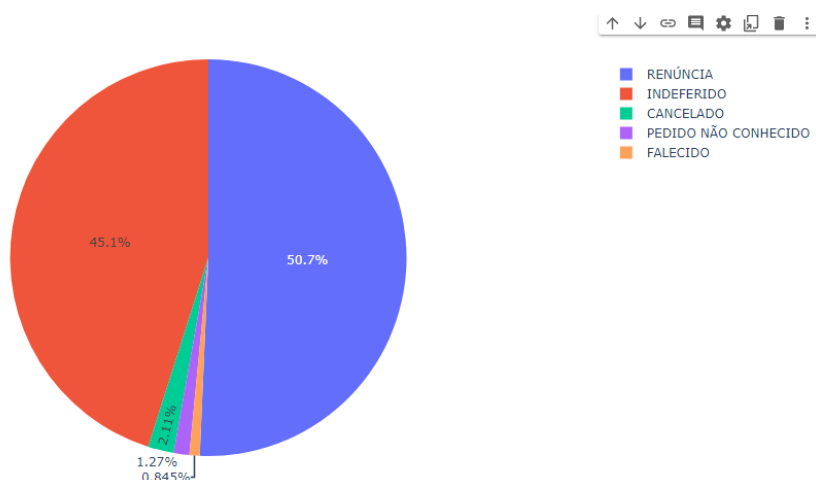
representam seus respectivos nomes. Ou seja, colunas com o mesmo tipo de informação. No caso de SG_UE, não existem 2 municípios com mesmo nome dentro do estado, somente em estados diferentes.

- 4) **TP_ABRANGENCIA**, pois só estamos tratando dados municipais (essa coluna assume valores de “municipal”, “estadual” e “federal”). Todos os registros serão sempre os mesmos.
- 5) **SG_UF**, a sigla do estado, também valor único para todos os registros. Não precisamos desse valor nessa tabela já que é um dado óbvio (MG). Somente se olhássemos a tabela de eleição do Brasil inteiro, separado por estados que faria sentido manter.
- 6) **CD_MUNICIPIO_NASCIMENTO** todos os registros são -3 (o que significa que nenhum código de município foi registrado pela urna naquele ano). Não precisamos desse dado que nada diz.
- 7) **NR_PROTOCOLO_CANDIDATURA** todos os registros são -1 (indicando valor nulo). Nenhum dos registros dessa coluna tem valor relevante para a análise.
- 8) **DT_NASCIMENTO** já que temos o campo NR_IDADE_DATA_POSSE, que representa a idade do candidato.

REDUÇÃO DE REGISTROS

Há certos registros que não são úteis na nossa análise. Eles são compostos de candidatos que por algum motivo **não foram inseridos na urna** (seja por renúncia do candidato, falecimento do mesmo, análise de candidatura indeferida ou cancelada, ou o pedido de candidatura não tenha sido reconhecido). Eles totalizam 1894 candidatos.

A grande maioria dos candidatos não inseridos na urna, renunciou ou teve sua candidatura indeferida:



A redução de registros também acontece quando os dados do candidato estão “duplicado” na base. Isso ocorre pois um mesmo candidato pode ter ido para 2º turno, em data adversa do 1º turno. Os dados são os mesmos, o que muda é o nº do turno, a data de acontecimento desse novo turno e se o candidato foi ou não eleito no 2º turno. Como os dados do 2º turno são mais relevantes, escolho esses dados ao invés dos que estão em 1º turno.

São 32 registros de duplicados.

	NM_TIPO_ELEICAO	NR_TURNO	CD_ELEICAO	DS_ELEICAO	DT_ELEICAO	NM_UE	DS_CARGO	SQ_CANDIDATO	NR_CANDIDATO	NM_CANDIDATO	...	DS_SI
72561	ELEIÇÃO ORDINÁRIA	1	426	Eleições Municipais 2020	2020-11-15	JUIZ DE FORA	VICE-PREFEITO	130000640722	40	ALEXANDRE NOCELLI	...	
61001	ELEIÇÃO ORDINÁRIA	2	427	Eleições Municipais 2020	2020-11-29	JUIZ DE FORA	VICE-PREFEITO	130000640722	40	ALEXANDRE NOCELLI	...	
42029	ELEIÇÃO ORDINÁRIA	1	426	Eleições Municipais 2020	2020-11-15	GOVERNADOR VALADARES	PREFEITO	130000883935	45	ANDRÉ LUIZ COELHO MERLO	...	
4347	ELEIÇÃO ORDINÁRIA	2	427	Eleições Municipais 2020	2020-11-29	GOVERNADOR VALADARES	PREFEITO	130000883935	45	ANDRÉ LUIZ COELHO MERLO	...	
23170	ELEIÇÃO ORDINÁRIA	1	426	Eleições Municipais 2020	2020-11-15	UBERABA	PREFEITO	130000773401	14	ANTONIO CARLOS SILVA NUNES	...	
76418	ELEIÇÃO ORDINÁRIA	2	427	Eleições Municipais 2020	2020-11-29	UBERABA	PREFEITO	130000773401	14	ANTONIO CARLOS SILVA NUNES	...	

Há candidatos que foram indeferidos na candidatura, mas foram colocados na urna (totalizam 980 registros). Isso é possível por vários motivos, dentre eles, sua candidatura estar ainda sendo julgada, sem parecer definitivo, pois tem recurso. Portanto, manterei esses candidatos na base como válidos, pois além de acrescidos nas urnas, eles podem receber votos e inclusive ganhar eleições.

Enriquecer e melhorar os dados.

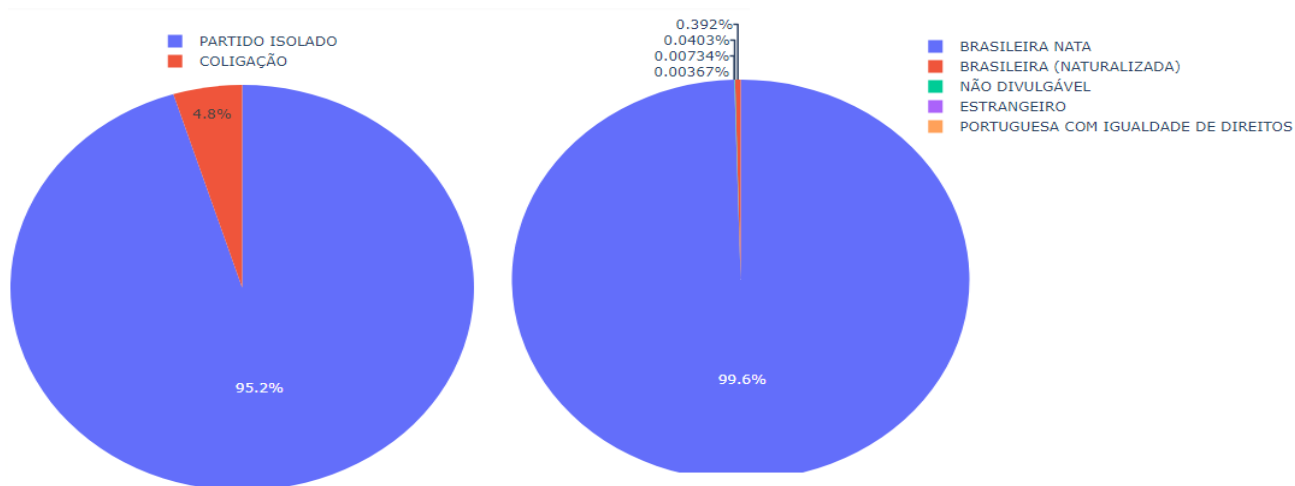
Melhorei os dados do label para poder ter 3 classes (eleito, não eleito e suplente) invés de 6 (nulo, suplente, não eleito, eleito por QP, eleito por média). Agrupei o “eleito por QP”, “eleito por média” e “eleito” em apenas “eleito”. Os dados nulos eram somente 2 ocorrências (portanto descartei).

Embora o nome da variável contenha a descrição de turno, não será considerado o turno uma vez que tratei isso anteriormente.

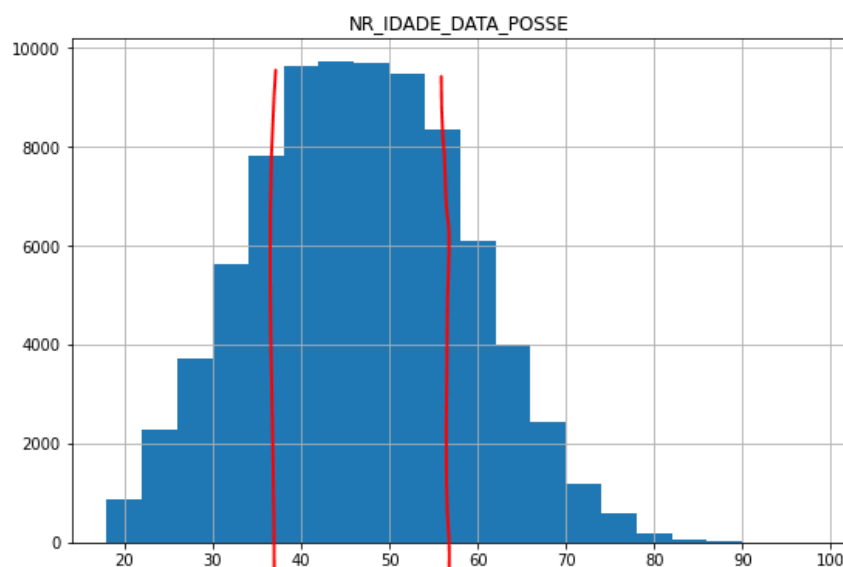
Análises dos dados graficamente ou em tabela

Os dados apresentados nessa base possibilitam uma infinidade de análises diferentes que podem ajudar a responder a questão inicial, ou tirar outros insights interessantes sobre os candidatos da eleição.

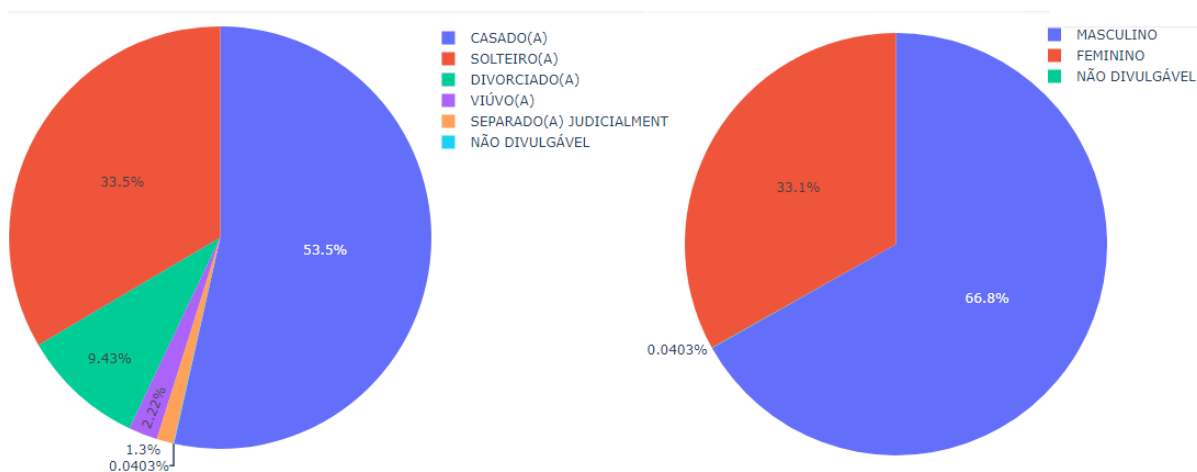
- Verificamos que 95% dos candidatos estão sem coligação e participam apenas de 1 partido;
- 99,6% dos candidatos são brasileiros, mas também temos estrangeiros concorrendo numa parcela bem pequena, totalizando 0.4% da base.



- Verificamos que a grande maioria dos candidatos tem entre 35 a 55 anos aproximadamente.



- Verificamos que tem muito mais homem concorrendo a cargos que mulheres e a maioria dos candidatos está solteira.



Com relação a se foi ou não eleito, vemos que:

- O número de homens eleitos para um cargo é 8x maior que mulheres
- O número de homens suplentes é quase 3x mais que o de mulheres.

	DS_SIT_TOT_TURNO	DS_GENERO	QUANTIDADE
0	ELEITO	FEMININO	1350
1	ELEITO	MASCULINO	8832
2	NÃO ELEITO	FEMININO	5991
3	NÃO ELEITO	MASCULINO	13053
4	SUPLENTE	FEMININO	19008
5	SUPLENTE	MASCULINO	31639

- Tanto para prefeito, vereador ou vice-prefeito, vemos que é entre 7 a 10x mais elegíveis os homens que as mulheres.

	DS_SIT_TOT_TURNO	DS_CARGO	DS_GENERO	QUANTIDADE
0	ELEITO	PREFEITO	FEMININO	61
1	ELEITO	PREFEITO	MASCULINO	790
2	ELEITO	VEREADOR	FEMININO	1182
3	ELEITO	VEREADOR	MASCULINO	7298
4	ELEITO	VICE-PREFEITO	FEMININO	107
5	ELEITO	VICE-PREFEITO	MASCULINO	744

- Dos eleitos, vemos que a maioria tem ensino médio completo ou superior completo. A surpresa vem ao saber que pessoas que só tem o básico, como saber ler e escrever também foram eleitas para vereador/vice-prefeito ou prefeito.

	DS_SIT_TOT_TURNO	DS_GRAU_INSTRUCAO	QUANTIDADE
0	ELEITO	ENSINO FUNDAMENTAL COMPLETO	1247
1	ELEITO	ENSINO FUNDAMENTAL INCOMPLETO	1602
2	ELEITO	ENSINO MÉDIO COMPLETO	3510
3	ELEITO	ENSINO MÉDIO INCOMPLETO	454
4	ELEITO	LÊ E ESCRIVE	216
5	ELEITO	SUPERIOR COMPLETO	2803
6	ELEITO	SUPERIOR INCOMPLETO	350

- O maior número de pessoas se encontra como suplentes tendo ensino médio completo.

17	SUPLENTE	ENSINO MÉDIO COMPLETO	18332
----	----------	-----------------------	-------

- Foram encontrados 20 nomes sociais (pessoas que se identificam como transgêneros/travestis concorrendo aos cargos).

```
Total dos nomes sociais: 21
['LORRAYNE COELHO',
 'RANNEY MENDES SILVA',
 'LUANA EMERENCIANO',
 'RHIelly GOMES BARCARO',
 'ANDRÉ SILVA OLIVEIRA',
 'JAMÍLY SILVÉRIO DO CARMO',
 'MALÚ ÁGATHA EDUARDA DE SOUZA',
 'AYSLA LIARAH CARVALHO',
 'EDNA LUIZA BULGARELLI IDE',
 'ANDRESSA MARINA FERREIRA',
 'ALESSANDRA PEREIRA GOMES',
 'RHAUHANNA FARIAS GONÇALVES',
 'JULIA GUIMARÃES',
 'LETICIA FERREIRA DE SOUZA',
 'ARIELLA SOUZA DUTRA',
 'MAMUSKA OLIVEIRA GOMES',
 'JULIANA CARLA DOS SANTOS',
 'RAYCA TEIXEIRA VIANA',
 'ANA LUZ FLORES SILVA',
 'LARA FABIA CANDIDA PEREIRA',
 'REBECA BORGES GONÇALVES']
```


- Dos eleitos, maioria se declara como branco. Seguido da cor parda.

	DS_SIT_TOT_TURNO	DS_COR_RACA	QUANTIDADE
0	ELEITO	AMARELA	36
1	ELEITO	BRANCA	5768
2	ELEITO	INDÍGENA	9
3	ELEITO	NÃO INFORMADO	108
4	ELEITO	PARDA	3584
5	ELEITO	PRETA	677

Ao observarmos mais de perto a tabela do describe dos valores em string, percebo que curiosamente nem todos os valores de nome são únicos, o que pode significar que uma pessoa está concorrendo a mais de um cargo em municípios diferentes:

	DS_GENERO	DS_ESTADO_CIVIL	ST_CANDIDATO_INSERIDO_URNA	DS_DETALHE_SITUACAO_CAND	DS_SITUACAO_CANDIDATURA	NM_CANDIDATO	TP_AG
count	79873	79873	79873	79873	79873	79873	79873
unique	3	6	1	9	2	78002	
top	MASCULINO	CASADO(A)	SIM	DEFERIDO	APTO	JOÃO BATISTA DA SILVA	
freq	53522	42871	79873	78833	78895	22	

ANÁLISE DE DADOS NULOS, NANS E VAZIOS

Os dados nulos, NaNs e vazios são representados com os valores numéricos ou strings nessa tabela (não com nulo).

- 1) Campos com #NE (caso sejam campo de string) ou -3 (caso seja campo numérico) indica que naquele ano, as urnas não computavam aquela informação. Não foram encontradas nenhuma coluna com campos “#NE”, porém foi encontrado o valor -3 somente na coluna de CD_MUNICIPIO_NASCIMENTO (em todos os dados dela), que é o código do município. Essa coluna foi eliminada completamente, já que não fornecia qualquer informação relevante.
- 2) Campos com “NÃO DIVULGÁVEL” são os dados que chegaram a existir, mas foram anonimizados posteriormente à eleição por pedido do candidato para privacidade dos mesmos. Na análise com os principais campos que preciso para resolver o problema, o “NÃO DIVULGÁVEL” era tão insignificante numericamente que coloquei a média, moda no local ou exclui o registro.

- 3) Campos com #NULO, se fossem poucos registros, descartava o registro ou colocava a média dos valores.
- 4) Campos com “Não informado” considere como dados importantes (não os modifiquei).

Preparar os dados de acordo com os algoritmos.

Como preparação de dados para ambos algoritmos (naive bayes e random forest), separei o label dos atributos, transformei os atributos em dicionários e binarizei eles. Já o label foi apenas binarizado, direto, sem precisar transformar em dicionário.

Após essa transformação, dividi a base em treino e teste, na proporção, 20% para teste e 80% para treino.

Aplicar algoritmos de machine learning de dados (pelo menos 2).

Inseri no Naive Bayes primeiramente o treino (X_train e Y_train) para ele treinar o algoritmo e depois fiz o.predict() do teste (X_test).

Fiz o mesmo com o Random Forest, considerando 10 classificadores, e o restante dos parâmetros default.

Comparei o resultado do .predict() de cada um dos algoritmos com o label real do teste que eu tinha (Y_test), usando um iterator para iterar e somar cada diferença. Também usei métricas como a matriz de confusão para averiguar seu acerto.

Após essa primeira análise dos resultados dos 2 algoritmos, performei mais 2 teste, com mudanças variadas a fim de melhorar as previsões e treino.

Explicar o motivo do uso de cada técnica.

Utilizarei a técnica de Ensemble de Random Forest, pois é uma técnica combinada de várias árvores que potencializa o que seria encontrado se utilizássemos apenas uma árvore, ou algum outro método que não contenha combinações. Devido às múltiplas respostas e posterior agregação, o Ensemble se torna um dos mais poderoso/confiáveis em termo de acerto das previsões. Cada árvore é independente umas das outras, e retira bem viés/outliers devido ao fato de usar a técnica de bagging com reposição (ou seja, reparte a base em pedaços de tamanho igual, repetindo elementos se preciso em cada nova base de teste).

Utilizarei o Naive Bayes como aprendizado supervisionado, pois não fizemos nenhuma prática sozinho relativa a esse algoritmo, ao contrário da árvore de decisão e cluster por exemplo, e por eu já ter feito uma atividade prática relativa à regressão linear/múltipla/logística na matéria de Modelos Estatísticos esse mês. Também justifico por ser uma abordagem mais simples do que o Ensemble, boa para uma análise inicial/resultado inicial para depois conseguirmos um resultado melhor via Random Forest.

Fazer análise dos resultados.

NAIVE BAYES

Ao treinar com 63.898 registros e usar como base real 15975 registros usando os 5 atributos (gênero, raça, estado civil, cargo e grau de instrução), verificamos que o naive bayes errou 5186 registros (aproximadamente 1/3 da base de teste), com acurácia no treino de 0.66 e acurácia no teste de 0.67. Esses resultados nos mostram que não há overfitting mas também não está prevendo os dados com acurácia satisfatória.

Pela matriz de confusão abaixo, vemos que a classe “Suplente” foi reconhecida com sucesso, enquanto a classe “Não Eleito” foi a classe que mais foi reconhecida erroneamente como “Suplente”.

	ELEITO(prev)	NÃO ELEITO(prev)	SUPLENTE(prev)
ELEITO	269	86	1681
NÃO ELEITO	449	286	2966
SUPLENTE	0	4	10234

Random Forest

Análise feita sem muitas mudanças em parâmetros para primeira visualização de resultados. Inserir 10 classificadores.

Vemos que o Random Forest já performou melhor que o Naive Bayes, pois ele errou menos (5023 do random X 5186 do naive).

Sua acurácia foi de 0.681 na base de treino e 0.685 para base de teste. Novamente não tive overfitting e o resultado da acurácia foi melhor que o naive bayes em ambas as bases de treino e teste (naive 0.66 no treino e 0.67 no teste).

Nesse caso, também acho que a acurácia pode melhorar bastante ainda, mexendo nos parâmetros.

Na matriz de confusão, temos erros maiores na classe “Suplente” que o naive, bem menos acertos de eleitos, mas uma proporção bem maior no acerto de não eleitos.

NAIVE BAYES

	ELEITO(prev)	NÃO ELEITO(prev)	SUPLENTE(prev)
ELEITO	269	86	1681
NÃO ELEITO	449	286	2966
SUPLENTE	0	4	10234

RANDOM FOREST

	ELEITO(prev)	NÃO ELEITO(prev)	SUPLENTE(prev)
ELEITO	6	350	1680
NÃO ELEITO	13	730	2958
SUPLENTE	1	21	10216

Naive Bayes e Random Forest - 2º Análise

Outras análises que poderíamos testar para ver se há melhora do algoritmo poderiam envolver:

- Adição de outros atributos (como os atributos que escolhi são nominais ou ordinais, e não numéricos, não foi possível fazer matriz de correlação para verificar quais atributos mais se correlacionam com o label);
- Mudança na porcentagem de treino/teste (10% de teste para 90% de treino? Ou 30% de teste para 70% de treino?)
- Randomização das amostras (não randomizei as amostras pois o excel certamente NÃO estava ordenado por nenhuma dos atributos que inseri no problema. Provavelmente estava por alguma data ou por nome de candidato).

Ao adicionar as colunas "TP_AGREMIACAO" (que diz respeito se o candidato está participando de coligação ou está no partido isolado) e a "Nacionalidade", juntamente com uma porcentagem de treino de 70% para 30% de teste, não vi mudanças significativas tanto pro Naive como para o Random Forest:

- Naive:
 - Acurácia no treino: 0.65
 - Acurácia no teste: 0.66
 - Quantos errou: 8127


- Random Forest:
 - Acurácia no treino: 0.68
 - Acurácia no teste: 0.68
 - Quantos errou: 7586


Ao randomizar primeiramente a base e aplicar novamente ambos algoritmos, vemos que o Random Forest piorou um pouco, mas o Naive Bayes piorou drasticamente muito:

- Naive:
 - Acurácia no treino: 0.13
 - Acurácia no teste: 0.12
 - Quantos errou: 20.896
- Random Forest:
 - Acurácia no treino: 0.63
 - Acurácia no teste: 0.63
 - Quantos errou: 8724

Se eu randomizar novamente, os valores da acurácia e a quantidade de erros muda, agora melhorando um pouco mais com relação ao último resultado.

```
treinarPreverMedir(resultado['binarizacao'], binarizadoLabel)
```

Acurácia no treino 0.23705532006224178, no teste 0.23119939904849346, diferença 18422  NAIVE BAYES

Acurácia no treino 0.6345978430004829, no teste 0.6354644854352726, diferença 8735  RANDOM FOREST

Suspeito que isso pode acontecer pois na hora de dividir a base, os registros com variabilidade podem acabar ficando na base de teste, enquanto a grande maioria repetida fica na base de treino. Portanto o treino não teria tanta variabilidade assim para poder comparar (a randomização acaba enviesando para um lado das bases de treino/teste).

Conclusão

Para uma limpeza e preparação maior dos dados (ou uma continuação desse trabalho), seria interessante ver a coluna de CPFs para analisar e aprofundar se um mesmo candidato poderia estar concorrendo a 2 posições diferentes ou a mesma posição em localidades diferentes, igual explicitado anteriormente.

Já na parte dos algoritmos e análises de previsão da base, vemos que o Random Forest continua sendo melhor que o Naive Bayes, nos 2 testes feitos (agora provado na prática nos experimentos que fiz, não somente na teoria).

No segundo teste com ambos, creio que o valor da acurácia não mudou muito uma vez que as 2 colunas inseridas não tem variabilidade significativa, (ou seja, não ajudam aos algoritmos a decidirem, uma vez que mais de 99% dos candidatos é brasileiro e mais de 95% estão em partido isolado).

O fato de eu randomizar a base antes de aplicar os algoritmos e ter piorado em ambos, tendo destaque negativo para o Naive Bayes foi uma surpresa, pois eu esperava que essas técnicas melhorassem, e não piorassem o algoritmo.

Como outra possibilidade de melhora dos algoritmos, podemos utilizar futuramente a hiperparametrização do Random Forest, mudando a quantidade de classificadores e outros parâmetros, seja usando técnicas de Grid Search ou Random Search.