Relatório Trabalho 3

Disciplina INF1771 – Inteligência Artificial, PUC-Rio 2018.2

Aluno André Mazal Krauss

Professora Renatha Capua

Organização do projeto e execução

O código do projeto está organizado em dois arquivos que acompanham este relatório. *Exploracao e Justificativa.ipynb* é um Jupyter Notebook que contém toda a parte de experimentação de parâmetros e desenho de gráficos. Para executá-lo, é necessário abri-lo com o Jupyter, porém os resultados da execução também estão em *Exploracao e Justificativa.html* em formato html e *Exploracao e Justificativa.pdf* em formato pdf. O mesmo vale para *Aplicacao.ipynb*, que contém uma simples execução dos algoritmos escolhidos sobre a base de dados, e cujo resultado está em *Aplicacao.html* e *Aplicacao.pdf.* A base de dados original está em covtype.data.

Os slides para a apresentação em sala estão em apresentacao.pdf.

Dependências para execução do código fonte:

- Python: 3.6.4
- scipy: 1.1.0
- numpy: 1.15.4
- matplotlib: 2.1.2
- pandas: 0.22.0
- sklearn: 0.20.1

O código fonte também está em *Exploracao e Justificativa.py* e *Aplicacao.py*, porém, sem a interface gráfica do Jupyter Notebook, os gráficos e figuras não serão visualizadas.

Informações Gerais e Definição do Problema

A base de dados escolhida para a realização do trabalho é a *Covertype Data Set.* A base de dados descreve características de 4 reservas florestais localizadas na *Roosevelt National Forest,* na parte Norte do Colorado, nos EUA. O terreno do parque foi repartido em lotes de 30x30 metros para a sua categorização na base de dados, cujas informações foram determinadas, pela *US Forest Service (USFS)* e *US Geological Survey (USGS).* A base possui 581012 instâncias, cada uma com 12 medições (que são representadas em 54 colunas de dados).

O problema consiste em, para um lote 30x30, prever qual tipo de cobertura vegetal nele predomina. Para isso, devem ser usados os demais atributos fornecidos, que descrevem aspectos geográficos e geológicos do lote em questão.

Para sua resolução, utilizo os algoritmos da Árvore de Decisão e o *K Nearest Neighbors Classifier*(KNN).

A base de dados está disponível abertamente no *Machine Learning Repository* da *University of California Irvine,* e foram doados por Jock A. Blackard, Denis J. Dean e Charles W. Anderson. Mais informações estão disponíveis na página web[1] da base de dados.

Descrição da Modelagem dos Exemplos de Treinamento

Atributos selecionados para descrever os exemplos

Farei aqui uma breve descrição dos atributos usados na predição, conforme estão descritos na documentação da base de dados.

1. Elevação – A elevação do terreno, medida em metros acima do nível do mar.
2. Aspecto – O aspecto do terreno, ou seja, a direção cardinal para onde sua face inclinada aponta. Medida em graus azimute entre 0 e 360.
3. Inclinação – A inclinação do terreno, em graus.
4. Distância Horizontal até corpo dágua – A distância em metros até o corpo dágua mais próximo
5. Distância Vertical até corpo d'água – A distância em metros até o corpo de água mais próximo. Pode ser negativa, indicando que o terreno está localizado abaixo do corpo d'água.
6. Distância Horizontal até Rodovias – a distância horizontal até rodovias, medida em metros.
7. Sombreamento_9am – Índice de 0 a 255 quantificando a exposição ao sol às 9:00 do solstício de verão.
8. Sombreamento_12am – Índice de 0 a 255 quantificando a exposição ao sol ao meio-dia do solstício de verão.
9. Sombreamento_3pm – Índice de 0 a 255 quantificando a exposição ao sol às 15:00 do solstício de verão.
10. Distância Horizontal até foco de incêndio – a distância horizontal até um foco inicial de incêndios florestais, medida em metros.
11. Área florestal - a Área de Preservação em que se encontra o lote em questão. No arquivo essas informação é expressa em 4 colunas binárias, que podem ser sintetizadas em uma simples categorização, em que cada área florestal tem os seguintes índices: Rawah(1), Neota(2), Comanche Peak(3) e Cache la Poudre(4).

Para além dos atributos, também há a própria classificação a ser predita. Há 7 diferentes classes de cobertura vegetais presentes na base de dados:

1. *Spruce/Fir*
2. *Lodgepole Pine*
3. *Ponderosa Pine*
4. *Cottonwood / Willow*
5. *Aspen*
6. *Douglas-fir*
7. *Krummholz*

[1] http://archive.ics.uci.edu/ml/datasets/Covertype

Esta mesma numeração é usada na base de dados e no código desenvolvido.

Justificativa para a escolha dos atributos

Da base de dados original, a única medição descartada inteiramente foi o tipo de solo presentes no terreno. Optei por descartar de imediato esta informação para diminuir o volume de dados (esta categorização está descrita em 40 colunas binárias) e para focar exclusivamente em atributos que dispensam qualquer pesquisa detalhada no terreno em questão.

Eliminada essa medição, procurei mensurar a relevância das demais para uma classificação correta. Para isso, cabe analisar certos dados estatísticos do conjunto das amostras como um todo.

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology
count	581012.000000	581012.000000	581012.000000	581012.000000	581012.000000
mean	2959.365301	155.656807	14.103704	269.428217	46.418855
std	279.984734	111.913721	7.488242	212.549356	58.295232
min	1859.000000	0.000000	0.000000	0.000000	-173.000000
25%	2809.000000	58.000000	9.000000	108.000000	7.000000
50%	2996.000000	127.000000	13.000000	218.000000	30.000000
75%	3163.000000	260.000000	18.000000	384.000000	69.000000
max	3858.000000	360.000000	66.000000	1397.000000	601.000000

Figura 1 - Sumário estatístico pt1

Horizontal_Distance_To_Roadways	Hillshade_9am	Hillshade_Noon	Hillshade_3pm	Horizontal_Distance_To_Fire_Points
581012.000000	581012.000000	581012.000000	581012.000000	581012.000000
2350.146611	212.146049	223.318716	142.528263	1980.291226
1559.254870	26.769889	19.768697	38.274529	1324.195210
0.000000	0.000000	0.000000	0.000000	0.000000
1106.000000	198.000000	213.000000	119.000000	1024.000000
1997.000000	218.000000	226.000000	143.000000	1710.000000
3328.000000	231.000000	237.000000	168.000000	2550.000000
7117.000000	254.000000	254.000000	254.000000	7173.000000

Figura 2 - Sumário Estatístico pt2

As figuras 1 e 2 mostram a média(mean), desvio padrão(std), os percentis e os valores mínimo e máximo do conjunto dos dados. Ainda é difícil inferir sobre quais atributos tem maior relação com a classificação, porém essa análise já é útil por garantir que não há lacunas na base(a contagem em todos os atributos é idêntica e igual a 581012) e que cada atributo foi corretamente importado e condiz com o esperado(o Aspecto varia somente de 0 a 360, o índice de sombreamento varia somente de 0 a 255, só há atributos negativos para a distância vertical e não há valores mínimos ou máximos que sejam impossíveis).

Para prosseguir, julguei interessante realizar a mesma análise, porém dividida pela classificação. Desta análise, selecionei mostrar alguns atributos.

Elevation

Cover_Type	count	mean	std
1	211840.0	3128.644888	157.805543
2	283301.0	2920.936061	186.579366
3	35754.0	2394.509845	196.732427
4	2747.0	2223.939934	102.524587
5	9493.0	2787.417571	96.594047
6	17367.0	2419.181897	188.762292
7	20510.0	3361.928669	106.915301

Horizontal_Distance_To_Fire_Points

Cover_Type	count	mean	std
1	211840.0	2009.253517	1234.823898
2	283301.0	2168.154849	1424.315110
3	35754.0	910.955949	527.109484
4	2747.0	859.124135	480.861801
5	9493.0	1577.719794	995.611463
6	17367.0	1055.351471	576.374422
7	20510.0	2070.031594	1087.258556

Slope

Cover_Type	count	mean	std
1	211840.0	13.127110	6.826445
2	283301.0	13.550499	7.096956
3	35754.0	20.770208	9.005553
4	2747.0	18.528941	9.347785
5	9493.0	16.641315	8.214169
6	17367.0	19.048886	7.953378
7	20510.0	14.255924	7.463676

Wilderness_Area

Cover_Type	count	unique	top	freq
1	211840	3	1	105717
2	283301	4	1	146197
3	35754	2	4	21454
4	2747	1	4	2747
5	9493	2	3	5712
6	17367	2	4	9741
7	20510	3	3	13105

Horizontal_Distance_To_Roadways

Cover_Type	count	mean	std
1	211840.0	2614.834517	1497.907501
2	283301.0	2429.530799	1618.718859
3	35754.0	943.940734	614.681991
4	2747.0	914.199490	366.290139
5	9493.0	1349.765722	1044.690988
6	17367.0	1037.169805	570.571139
7	20510.0	2738.250463	1200.589081

Hillshade_Noon

Cover_Type	count	mean	std
1	211840.0	223.430211	18.159926
2	283301.0	225.326596	18.509548
3	35754.0	215.826537	27.905613
4	2747.0	216.997088	20.917345
5	9493.0	219.035816	24.913339
6	17367.0	209.827662	24.417547
7	20510.0	221.746026	20.000063

Figura 3 - Sumário estatístico por categoria

Primeiramente, devemos notar que, dos sete tipos de cobertura possíveis, dois se destacam: *Spruce/Fir*(1) e *Lodgepole Pine*(2). Juntas, essas duas categorias representam 85.2% de todas as amostras. Sobre os atributos, vale destacar que a Elevação tem médias bem espaçadas entre as categorias, com um baixo desvio padrão; a média das duas distâncias horizontais, até corpos d'água e até focos de incêndio, são muito distintas por categoria, porém também há um alto desvio padrão; e a média do sombreamento e da inclinação são pouco distinguíveis por categoria. Além disso, no atributo categórico Área Florestal percebe-se que há áreas coberturas de vegetação exclusivas para áreas florestais específicas. Porém, essas coberturas não são muito expressivas numericamente.

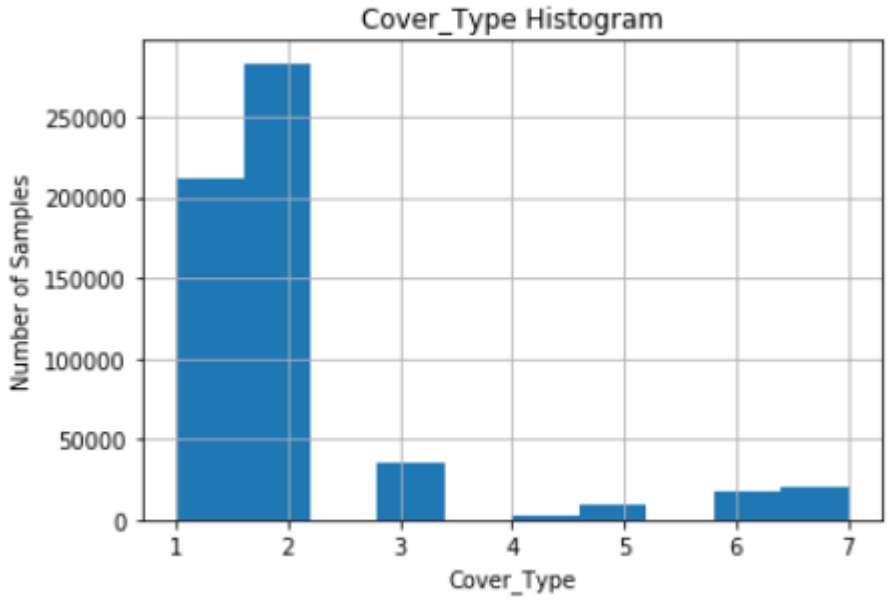


Figura 4 - O histograma por classe mostra que sua distribuição não é equilibrada

As estatísticas acima e os histogramas ajudam a entender o porque de, como veremos mais abaixo, os atributos mais relevantes para classificação serem elevação, distância até focos de incêndio e distância até estradas. Os demais atributos ajudam muito pouco individualmente, mas conjuntamente ainda acrescentam cerca de 15% de acurácia ao KNN., como veremos abaixo.

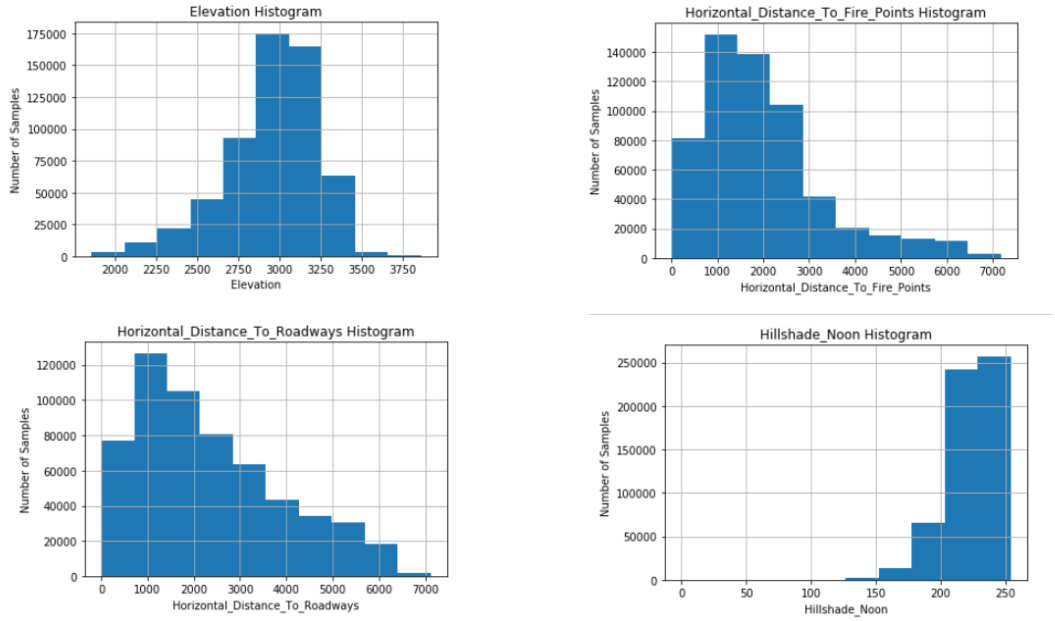
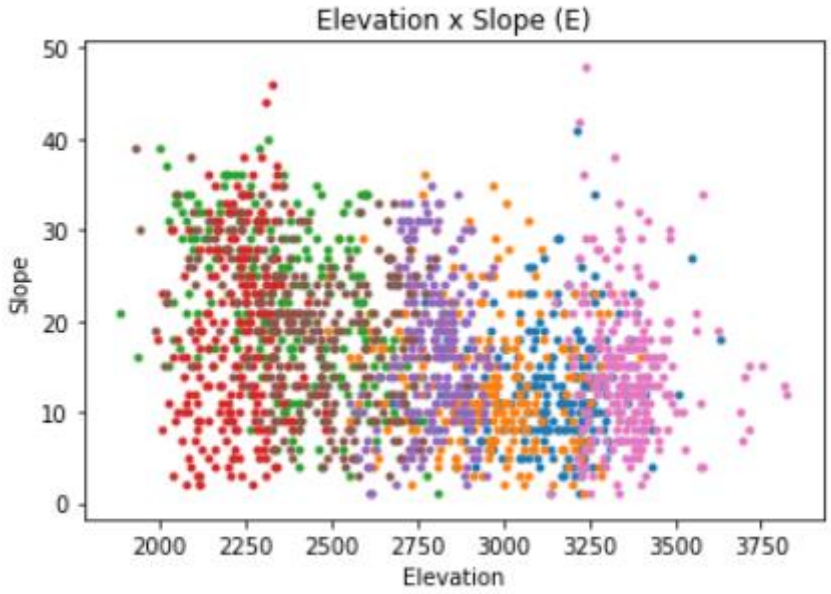
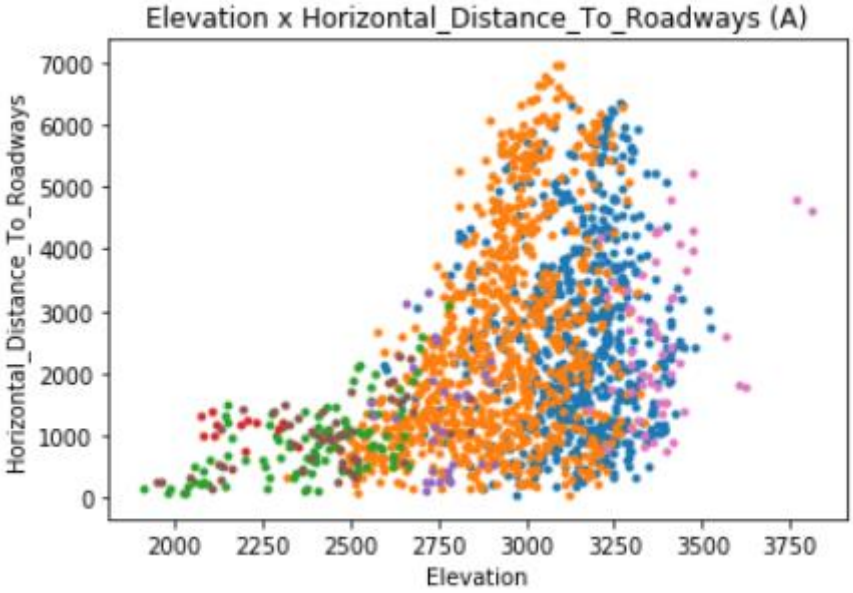


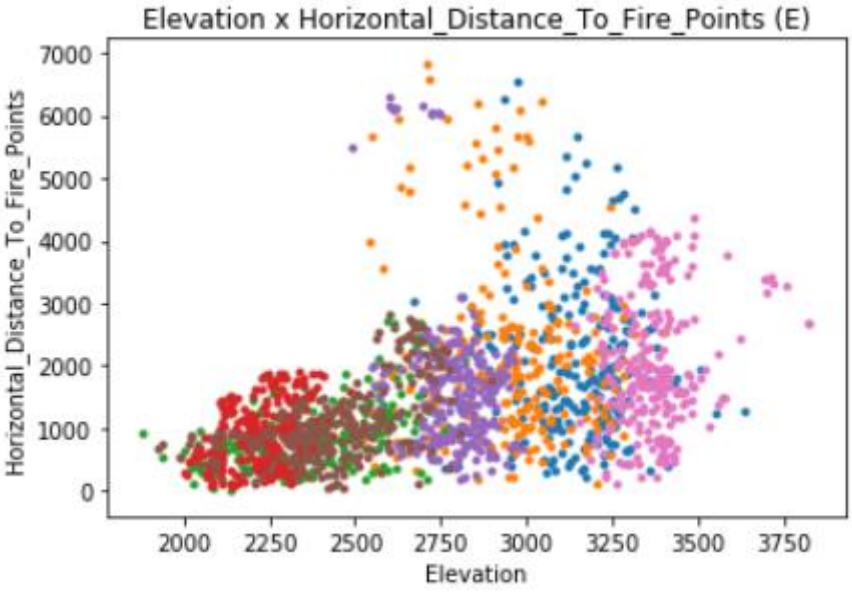
Figura 5 - Histogramas por atributo. Repare como o sombreamento é mais concentrado do que a elevação e as distancias mostradas

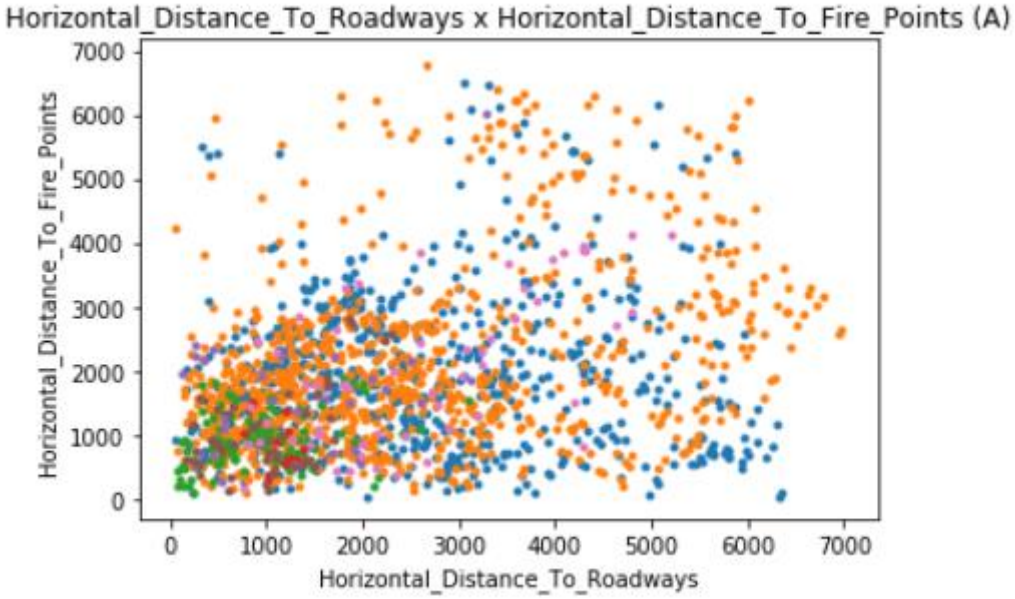
Por último, para justificar novamente a intuição de que Elevação, distância até focos de incêndio e distância até estradas são mais relevantes, mostro alguns *scatter plots* que reforçam essa ideia, mostrando que estes tem boa separabilidade. Para alguns plot foram utilizadas 250 amostras *por categoria*, ou seja, as menores categorias estão super-representadas, enquanto as maiores estão sub-representadas. Para outros, as amostras foram

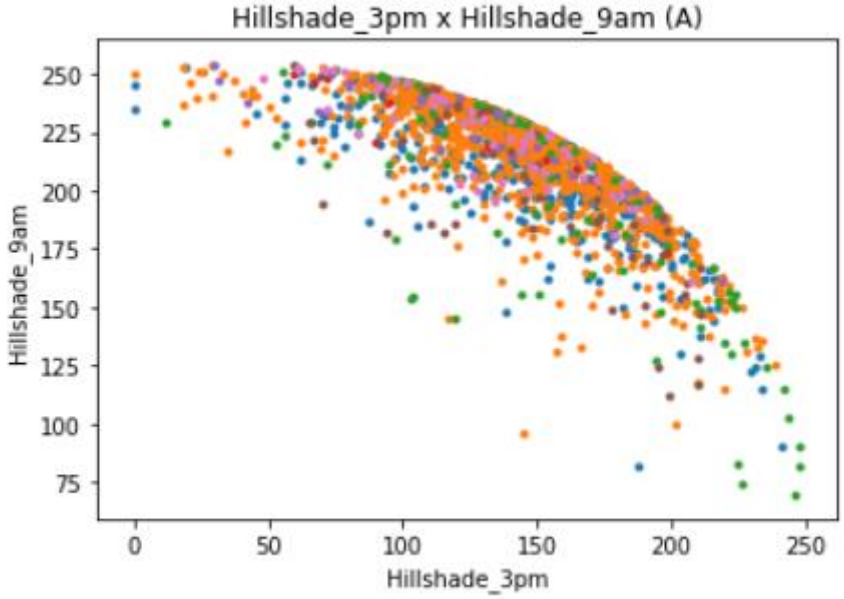
simplesmente escolhidas aleatoriamente. Isso é indicado pela letra E(de estratificação por classe) ou A(de aleatório) no título do gráfico. O meu único critério para escolha entre um e outro foi qual fornecia a melhor visualização.

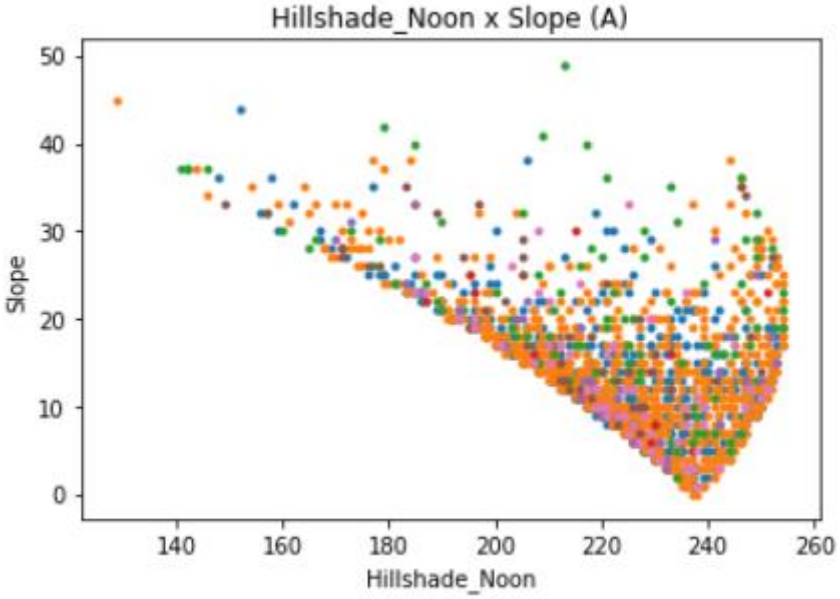


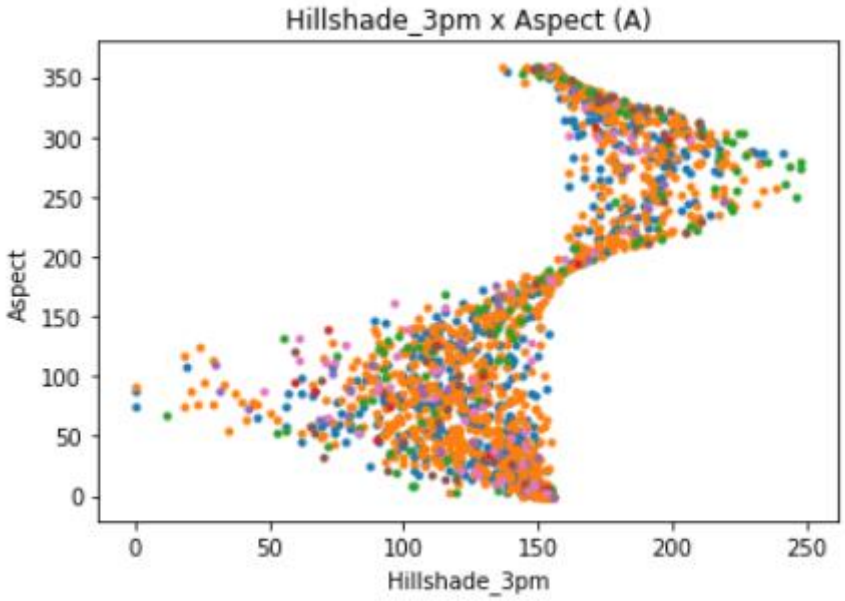



Elevation x Horizontal_Distance_To_Fire_Points (E)
Horizontal_Distance_To_Fire_Points
7000
6000
5000
4000
3000
2000
1000
0
2000
2250
2500
2750
3000
3250
3500
3750
Elevation


Horizontal_Distance_To_Roadways x Horizontal_Distance_To_Fire_Points (A)
Horizontal_Distance_To_Fire_Points
7000
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
7000
Horizontal_Distance_To_Roadways


Hillshade_3pm x Hillshade_9am (A)
Hillshade_9am
Hillshade_3pm
250
225
200
175
150
125
100
75
0
50
100
150
200
250


Hillshade_Noon x Slope (A)
Slope
Hillshade_Noon
50
40
30
20
10
0
140
160
180
200
220
240
260



Visualizando esses gráficos, poderíamos intuir que KNN seria uma abordagem adequada ao nosso problema, se usarmos os atributos corretos. Para além disso, alguns comentários:

- Nos gráficos envolvendo elevação, podemos facilmente perceber que ela divide bem as classes, com “colunas” de cor sendo facilmente visíveis.
- A distância até focos de incêndio e a distância até estradas se saem bem quando pareadas com a elevação. Porém, elas aparentemente não bastam em separado, principalmente por falhar em dividir as duas maiores classes (em cor laranja e azul).
- Como é de se esperar, existem relações interessantes entre os sombreamentos em diferentes horas do dia e a inclinação do terreno. Porém, infelizmente, essa informação aparentemente não ajuda na classificação da cobertura vegetal.

Descrição dos Experimentos Realizados

Para realizar os experimentos e escolher os parâmetros, comecei dividindo a base de dados numa relação 80/20 entre uma base de experimentação e outra de validação. Realizei então repetidos k-folds na base de experimentação buscando estimar a acurácia do método para diversos parâmetros. Tendo por fim escolhido os métodos e parâmetros adequados, os aplico sobre o grupo de validação para avaliar se a acurácia esperada se confirmou.

Realizado o K-Fold, cada treinamento é realizado com 418320 amostras.

Árvore de Decisão limitada a profundidade N

Limite de profundidade N	**Acurácia média dos 10-folds**	**Tempo de treino médio(segundos)**	**Número médio de nós na árvore**
1	63.36%	1.22	3.0
3	67.45%	1.81	15.0
5	69.24%	2.39	63.0
10	76.31%	3.72	1534.8
20	90.48%	5.76	34653.0
30	92.25%	6.09	59640.2
Sem limites (profundidade máxima: 41)	92.26%	6.07	52896.0

Escolhi treinar primeiro a árvore de decisão porque, ao analisarmos a árvore, nos é possível avaliar a importância relativa dos atributos utilizados. A biblioteca Sklearn oferece para isso a funcionalidade dos *feature importances,* que calcula, para cada atributo, sua importância relativa (de 0.0 a 1.0), baseado em quantas amostras ele separa e em qual grau de pureza[2]. Me baseei nesses valores para realizar mais testes sobre a árvore e o KNN.

```
Elevation: 0.346055
Aspect: 0.031844
Slope: 0.022997
Horizontal_Distance_To_Hydrology: 0.070997
Vertical_Distance_To_Hydrology: 0.056597
Horizontal_Distance_To_Roadways: 0.164803
Hillshade_9am: 0.034835
Hillshade_Noon: 0.039353
Hillshade_3pm: 0.027806
Horizontal_Distance_To_Fire_Points: 0.165863
Wilderness_Area: 0.038850
```

Os valores se alinham com a análise anterior, destacando a Elevação como o atributo mais importante, seguido da distância até focos de incêndio e da distância até estradas.

Árvore de decisão limitada a profundidade N e 3 atributos

Como discutido acima, os três atributos são Elevação, Distância até foco de incêndio e distância até estrada. Decidi, portanto, experimentar com uma árvore de decisão que só considerasse esses três. Porém, como visto na tabela abaixo, a árvore apresentou acurácia muito inferior com uma estrutura mais complexa que anteriormente (possui maior quantidade

[2] A documentação do *Sklearn* não é muito específica sobre esse cálculo, e somente afirma que usa a *Gini Importance*.

de nós e maior profundidade). Este pior resultado é esperado, dado que há menos informações disponíveis para o treinamento. O tempo de treinamento sim sofreu melhora, mas não creio que compensa o ônus geral.

Limite de profundidade N	Acurácia média dos 10-folds	Tempo de treino médio	Número médio de nós na árvore
Sem limites (profundidade: 47)	81.55%	2.61	124847.2

KNN

Apresento na tabela abaixo os resultados da experimentação com a quantidade de vizinhos a ser usada no KNN. Vê-se que 3 vizinhos foi a quantidade com maior acurácia. Além disso, vale notar que o tempo de treinamento não varia de acordo com a quantidade de vizinhos considerada, mas o tempo de avaliação sim, aumentando consideravelmente.

N vizinhos	Acurácia média dos 10-folds	Tempo de treino médio(segundos)	Tempo de avaliação médio (por amostra, em ms)
1	96.40%	2.94	0.029
3	96.76%	2.28	0.032
5	96.63%	2.34	0.041
7	96.39%	2.75	0.057
9	96.03%	2.35	0.055
11	95.67%	2.32	0.064
15	94.96%	2.44	0.077
25	93.38%	2.27	0.099
51	90.06%	2.33	0.151
101	85.83%	2.26	0.230

KNN limitado a três atributos

Novamente, decidi experimentar utilizar somente os três atributos mais relevantes (Elevação, Distância até focos de incêndio e Distância até rodovias), pretendendo obter uma acurácia satisfatória com tempos de execução menores. De fato, houve uma expressiva redução no tempo de treinamento e avaliação, mas com um grande custo na acurácia. Além disso, o numero ótimo de vizinhos se elevou para 9, apesar da faixa de 3 a 15 estar muito próxima.

N vizinhos	Acurácia média dos 10-folds	Tempo de treino médio(segundos)	Tempo de avaliação médio (por amostra, em ms)
1	78.08%	1.06	0.009
3	81.28%	0.96	0.012
5	82.00%	0.95	0.014
7	82.70%	0.97	0.016
9	82.87%	0.96	0.017
11	82.85%	0.95	0.018
15	82.55%	0.98	0.023

KNN considerando Aspecto como medida modular

Logo de início, pensei que deveria considerar o Aspecto do terreno diferentemente das demais medidas. Pensei nisso porque, diferentemente das demais, o aspecto tem uma característica modular, por ser uma medição de ângulo em um círculo de 360 graus. Isso significa que um valor de 359 deveria ser mais próximo do 1 do que o valor de 90, por exemplo, porque poderíamos "dar a volta completa". Para isso, implementei uma pequena função, que passei como métrica customizada ao KNN. Porém, o uso da minha função aumentou enormemente o tempo de processamento sem melhorar significativamente a predição, então abandonei a ideia. De fato, creio que usar uma função própria atrapalha as otimizações nativas do Sklearn, e o aspecto, como visto anteriormente, não seria um bom divisor para as predições.

Resultados Finais

Com toda a análise de alternativa feita acima, decidi pelo Árvore de Decisão com 12 atributos e sem limite de profundidade e pelo KNN com 12 atributos e 3 vizinhos. Apresento abaixo os resultados dos testes, com set de treinamento de 80% da base de dados (464809 amostras) e set de validação de 20% (116203 amostras).

Nas imagens abaixo, os seguintes termos significam:

- *Precision:* é a razão tp / (tp + fp), onde tp é o número de verdadeiros positivos para aquela classe, e fp o número de falsos negativos. Representa a capacidade do preditor de não classificar erroneamente um objeto como sendo pertencente à classe em questão.
- *Recall:* É a razão tp / (tp + fn), onde tp é o número de verdadeiros positivos para a classe e fn o número de falsos negativos. Ou seja, representa a capacidade do preditor de corretamente identificar um objeto como pertencente à classe em questão.
- *F1* – score: É o cálculo (2*precision * recall) / precision + recall. É essencialmente uma média entre *precision* e *recall.*
- *Support:* quantidade de amostras daquela classe entre as 116203 amostras de validação.

De forma geral, esses valores calculados por classe revelam que obtivemos melhores resultados para aquelas que tem mais amostras na base de dados, o que é esperado.

Árvore de Decisão

Acurácia obtida: 92.65%

Tempo de treinamento: 14.15 segundos, para 464809 amostras

Tempo médio para uma predição: 0.1 ms

	precision	recall	f1-score	support
1	0.93	0.93	0.93	42524
2	0.94	0.94	0.94	56386
3	0.91	0.91	0.91	7185
4	0.81	0.80	0.80	600
5	0.81	0.80	0.81	1905
6	0.85	0.85	0.85	3456
7	0.94	0.93	0.93	4147

KNN

Acurácia obtida: 96.96%

Tempo de treinamento: 3.96 segundos, para 464809 amostras

Tempo médio para uma predição: 1.2 ms

	precision	recall	f1-score	support
1	0.97	0.97	0.97	42524
2	0.97	0.98	0.97	56386
3	0.96	0.97	0.97	7185
4	0.92	0.81	0.86	600
5	0.91	0.90	0.91	1905
6	0.94	0.94	0.94	3456
7	0.97	0.97	0.97	4147

Considerações Finais

Apesar de escolher dois métodos bem simples de aprendizado supervisionado, considero que obtive resultados com bons níveis de acurácia e tempo. A plataforma python + Sklearn se provou muito adequada para obter resultados de maneira rápida e prática, oferecendo também acesso a outras ferramentas, como a biblioteca matplotlib e numpy, para obter gráficos estatísticas a partir dos dados. Esse conjunto me permitiu focar na análise dos dados e dos resultados, e não na implementação dos algoritmos. Entretanto, descobri uma desvantagem na perda de desempenho sofrida ao tentarmos usar métricas customizadas.