# Relatório Trabalho 3

Disciplina INF1771 – Inteligência Artificial, PUC-Rio 2018.2

Aluno André Mazal Krauss

Professora Renatha Capua

## Informações Gerais 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]](#footnote-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*

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.

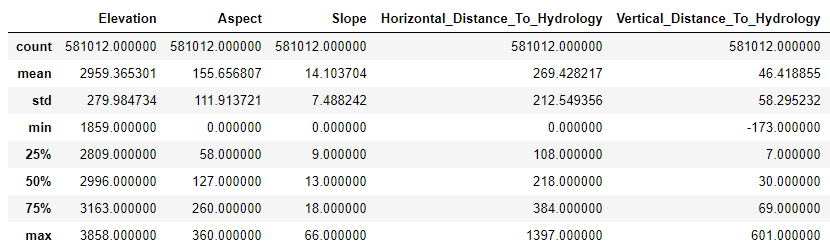


Figura 1 - Sumário estatístico pt1

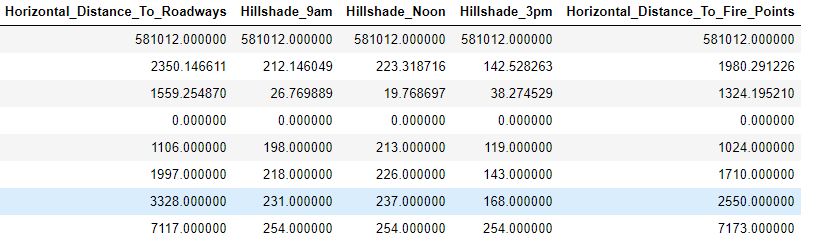
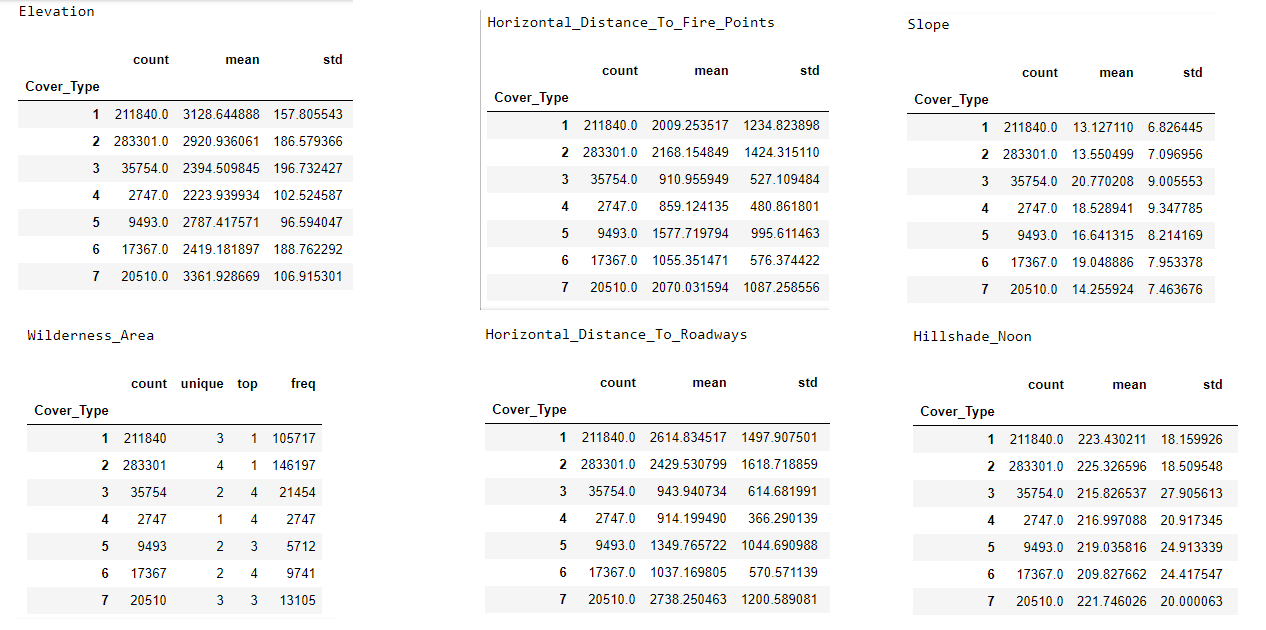


Figura 2 - Sumário Estatístico pt2

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.

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.

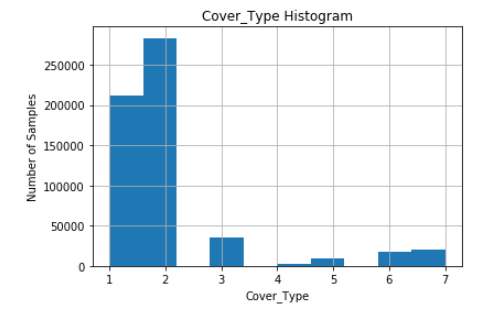
As estatísticas acima e os histogramas ajudam a entender o porque de, como veremos mais abaixo, os atributos mais relevantes para classificação serem elevação, distância até focos de incêndio e distância até estradas. Os demais atributos ajudam muito pouco individualmente, mas conjuntamente ainda acrescentam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

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

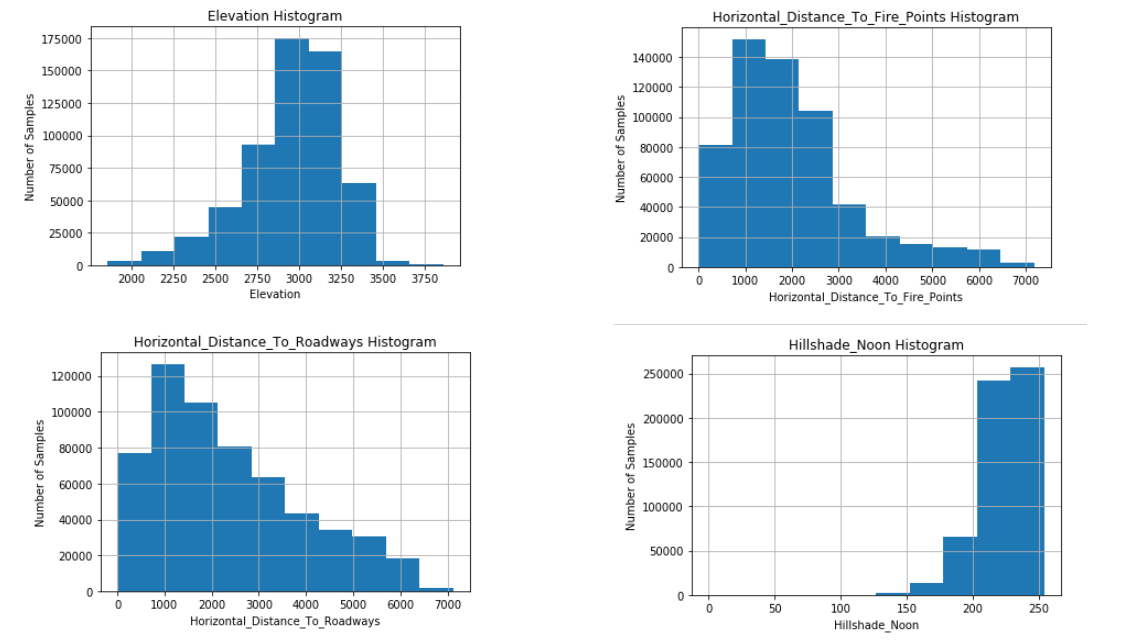
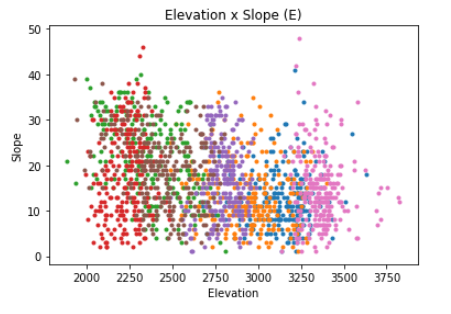
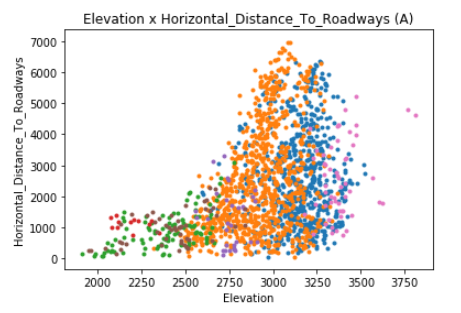
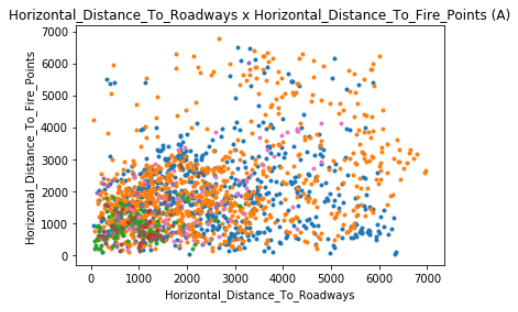
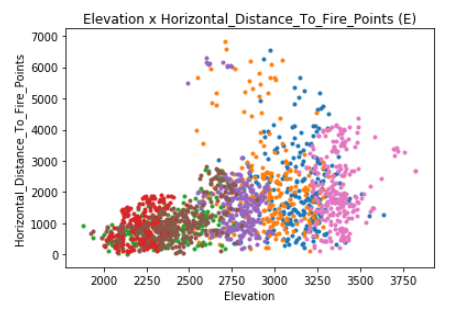
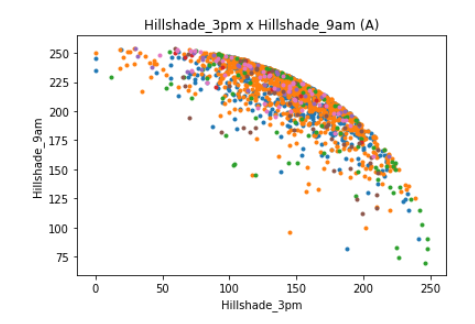
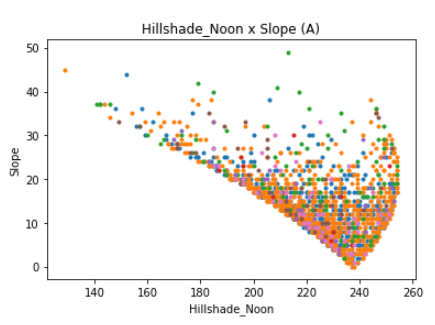
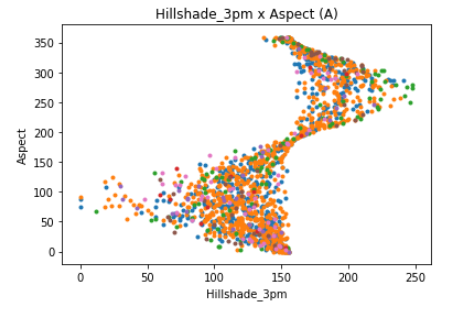


Figura 5 - Histogramas por atributo. Repare como o sombreamento é mais concentrado do que a elevação e as distancias mostradas

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.



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

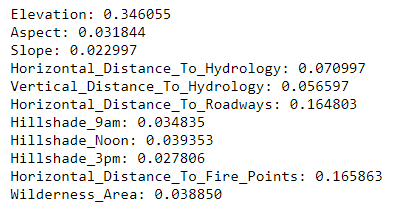
Para realizar os experimentos e escolher os parâmetros, comecei dividindo a base de dados numa relação 80/20 entre uma base de experimentação e outra de validação. Realizei então repetidos k-folds na base de experimentação buscando estimar a acurácia do método para diversos parâmetros. Tendo por fim 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]](#footnote-2). Me baseei nesses valores para realizar mais testes sobre a árvore e o KNN.



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 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

### KNN com ajuste de pesos

### KNN limitado a três atributos

## Comparação dos algoritmos realizados

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 pesos?). Apresento os resultados abaixo.

### Descrição de como os algoritmos escolhidos foram implementados

### Taxa de Reconhecimento

### Tempo gasto no processo de treinamento

### Tempo gasto no processo de classificação de um evento desconhecido

## Resultados Finais

### (redundante)Escolha dos classificadores e conjunto de atributos

### (redundante)Resultados alcançados pelo algoritmo que foi implementado(tempo gasto no processo de treinamento, tempo gasto no processo de classificação de um exemplo desconhecido, taxa de reconhecimento)

### Conclusões finais

1. http://archive.ics.uci.edu/ml/datasets/Covertype [↑](#footnote-ref-1)
2. A documentação do *Sklearn* não é muito específica sobre esse cálculo, e somente afirma que usa a *Gini Importance*. [↑](#footnote-ref-2)