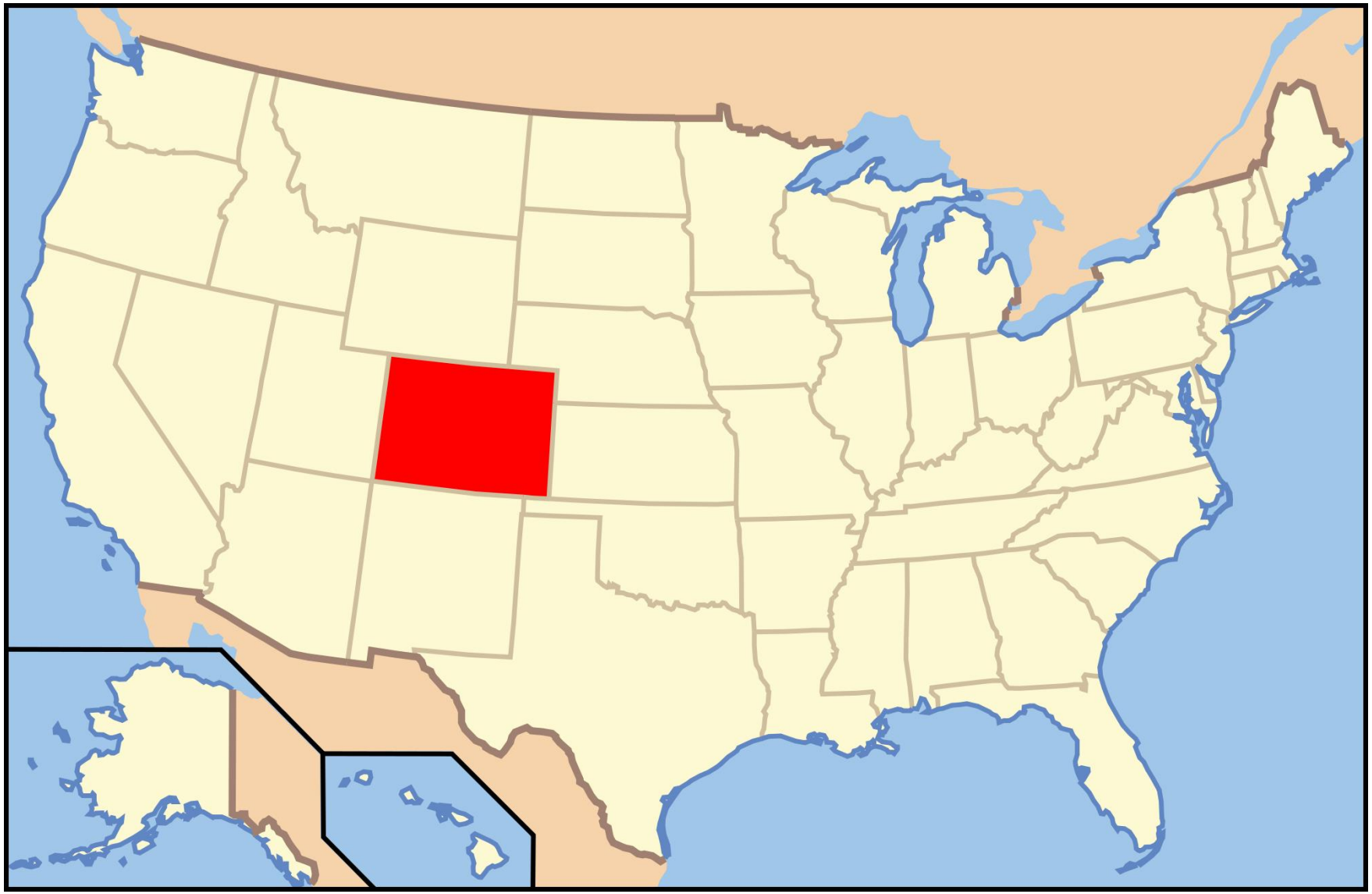
# INF1771 – Inteligência Artificial

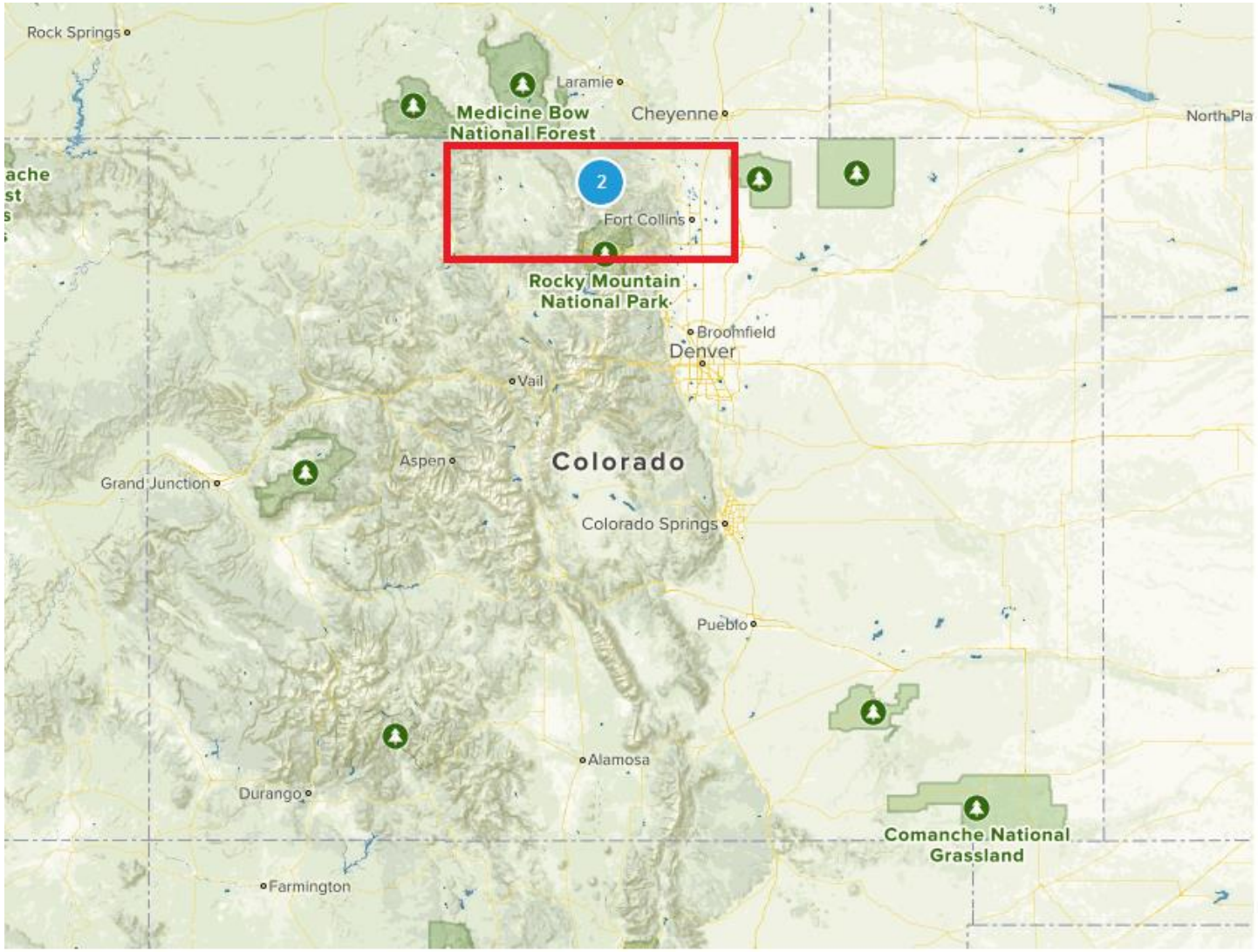
Trabalho 3 – Machine Learning

Aluno: André Mazal Krauss

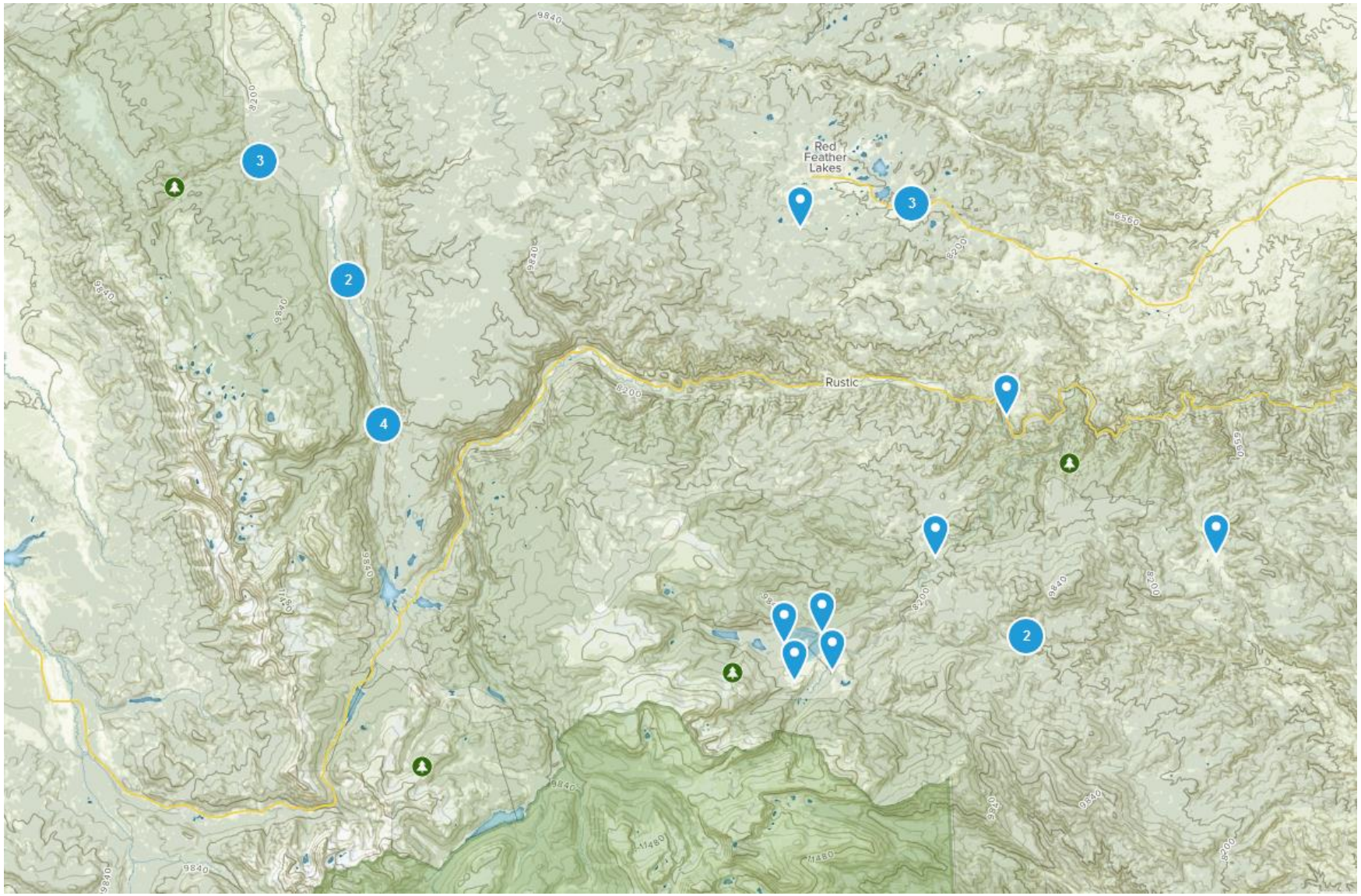
# Introdução



# Introdução


Rock Springs
Laramie
Medicine Bow
National Forest
Cheyenne
North Pla
2
Fort Collins
Rocky Mountain
National Park
Broomfield
Denver
Vail
Aspen
Colorado
Grand Junction
Colorado Springs
Pueblo
Alamosa
Durango
Comanche National
Grassland
Farmington

# Introdução



# Introdução

# Definição do Problema

Usando os dados disponíveis na página da UCI, prever se:

*Lodgepole Pine*

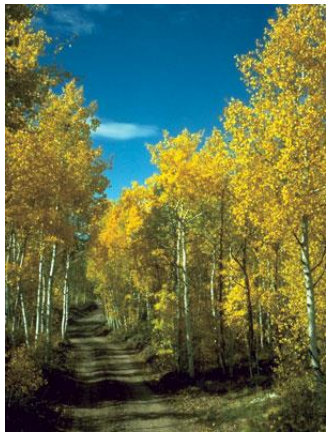
*Cottonwood*

*Krummholz*

*Ponderosa Pine*

*Douglas-Fir*

*Spruce*



*Aspen*

# Definição do Problema

Usando os dados disponíveis na página da UCI, prever se:

***Lodgepole Pine***

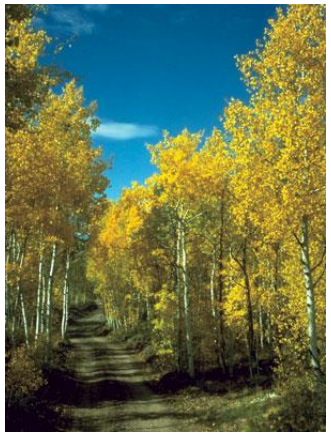
*Cottonwood*

*Krummholz*

*Douglas-Fir*

*Ponderosa Pine*

***Spruce***



*Aspen*

# Definição do Problema

Dados os atributos de um terreno 30x30 metros, baseado no aprendizado sobre a base de dados (581012 instâncias), prever qual tipo de cobertura vegetal prevalece naquele terreno.

**ATRIBUTOS:**

- Elevação
- Aspecto
- Inclinação
- Distância horizontal até corpo d'água
- Distância vertical até corpo d'água
- Distância horizontal até rodovias
- Sombreamento às 9:00
- Sombreamento às 12:00
- Sombreamento às 15:00
- Distância horizontal até focos de incêndio
- Área Florestal


**CLASSIFICAÇÃO:**

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

# Definição do Problema

Dados os atributos de um terreno 30x30 metros, baseado no aprendizado sobre a base de dados (581012 instâncias), prever qual tipo de cobertura vegetal prevalece naquele terreno.

**ATRIBUTOS:**

- **Elevação**
- Aspecto
- Inclinação
- Distância horizontal até corpo d'água
- Distância vertical até corpo d'água
- **Distância horizontal até rodovias**
- Sombreamento às 9:00
- Sombreamento às 12:00
- Sombreamento às 15:00
- **Distância horizontal até focos de incêndio**
- Área Florestal


**CLASSIFICAÇÃO:**

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

# Experimentação e Escolha dos Algoritmos

Implementação em Python 3.6, usando o ambiente Jupyter Notebook e as bibliotecas:

- scipy: 1.1.0: computação numérica
- numpy: 1.15.4: computação numérica, estruturas de dados
- matplotlib: 2.1.2: plotagem dos gráficos
- pandas: 0.22.0: importação e tratamento dos dados
- sklearn: 0.20.1: machine learning

# Experimentação e Escolha dos Algoritmos

| | 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 |

| | Horizontal_Distance_To_Roadways | Hillshade_9am | Hillshade_Noon | Hillshade_3pm | Horizontal_Distance_To_Fire_Points |
|---|---|---|---|---|---|
| count | 581012.000000 | 581012.000000 | 581012.000000 | 581012.000000 | 581012.000000 |
| mean | 2350.146611 | 212.146049 | 223.318716 | 142.528263 | 1980.291226 |
| std | 1559.254870 | 26.769889 | 19.768697 | 38.274529 | 1324.195210 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 1106.000000 | 198.000000 | 213.000000 | 119.000000 | 1024.000000 |
| 50% | 1997.000000 | 218.000000 | 226.000000 | 143.000000 | 1710.000000 |
| 75% | 3328.000000 | 231.000000 | 237.000000 | 168.000000 | 2550.000000 |
| max | 7117.000000 | 254.000000 | 254.000000 | 254.000000 | 7173.000000 |

# Experimentação e Escolha dos Algoritmos

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 |

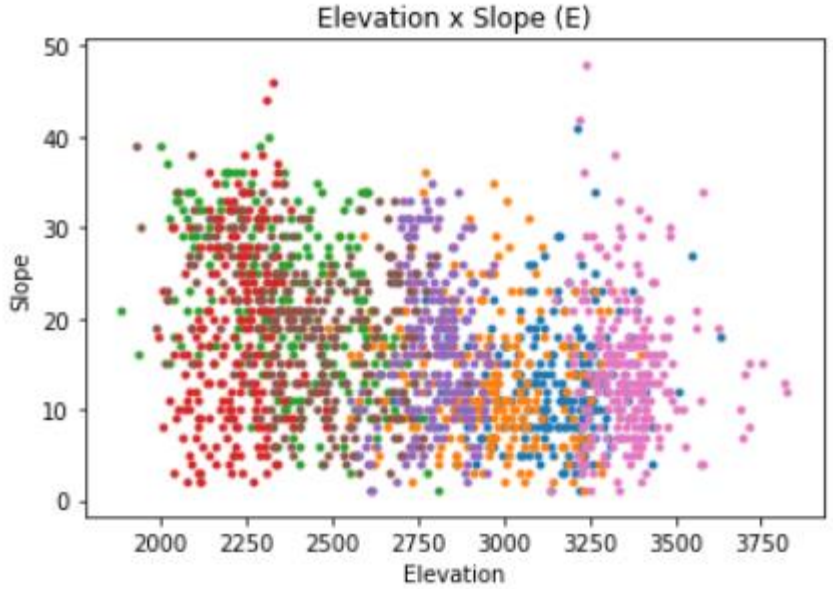
Horizontal_Distance_To_Roadways

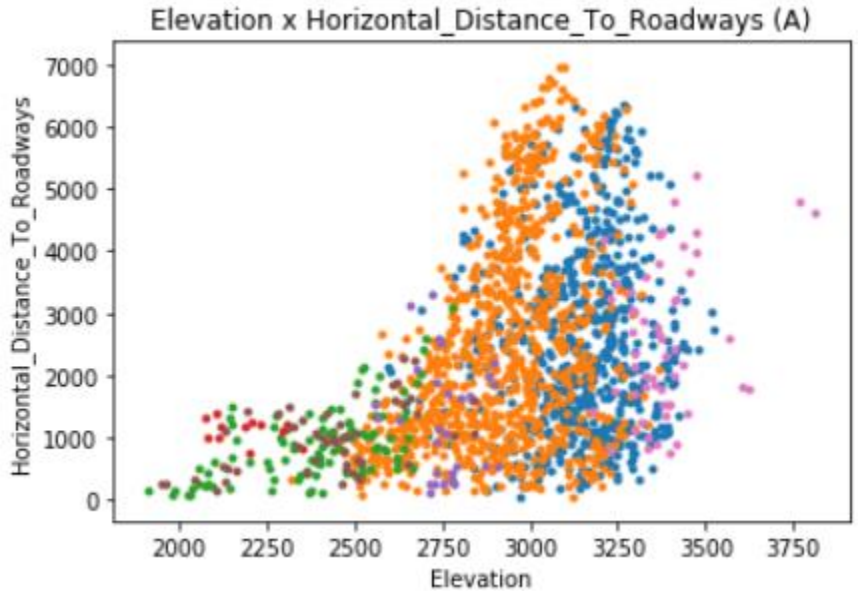
| Cover_Type | count | mean | std |
|---|---|---|---|
| 1 | 211840.0 | 2614.834517 | 1497.907501 |
| 2 | 283301.0 | 2429.530799 | 1618.718859 |
| 3 | 35754.0 | 943.940734 | 614.681991 |
| 4 | 2747.0 | 914.199490 | 366.290139 |
| 5 | 9493.0 | 1349.765722 | 1044.690988 |
| 6 | 17367.0 | 1037.169805 | 570.571139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

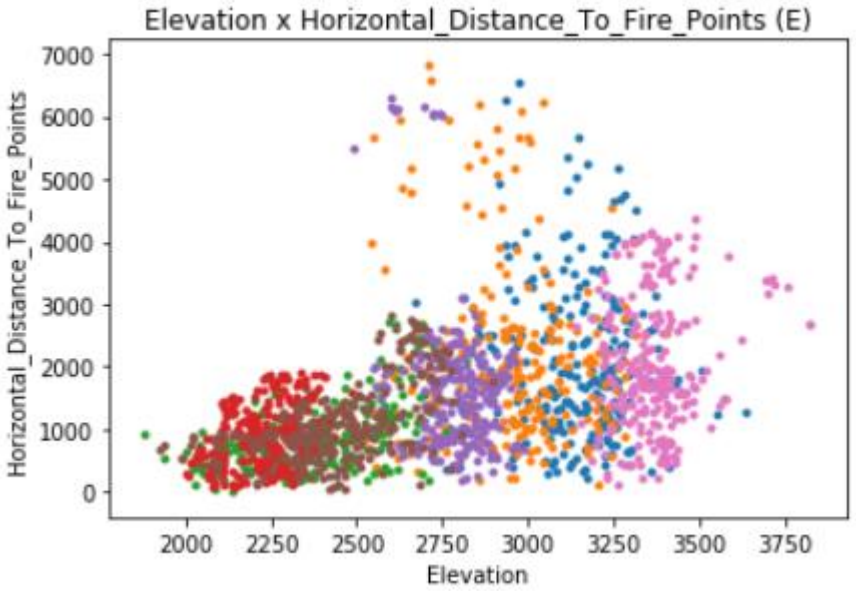
Hillshade_Noon

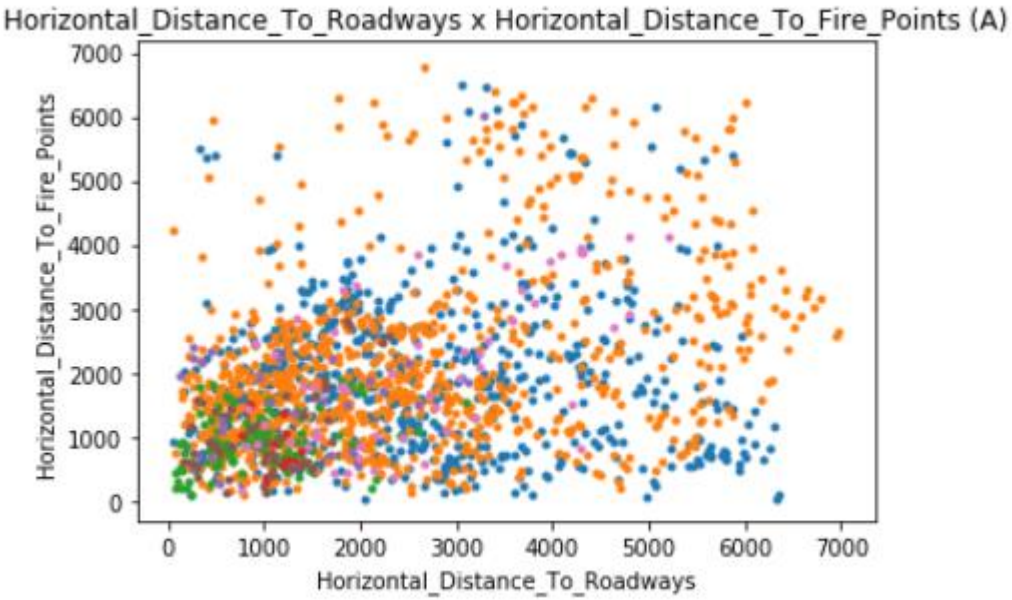
| Cover_Type | count | mean | std |
|---|---|---|---|
| 1 | 211840.0 | 223.430211 | 18.159926 |
| 2 | 283301.0 | 225.326596 | 18.509548 |
| 3 | 35754.0 | 215.826537 | 27.905613 |
| 4 | 2747.0 | 216.997088 | 20.917345 |
| 5 | 9493.0 | 219.035816 | 24.913339 |
| 6 | 17367.0 | 209.827662 | 24.417547 |
|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

# Experimentação e Escolha dos Algoritmos

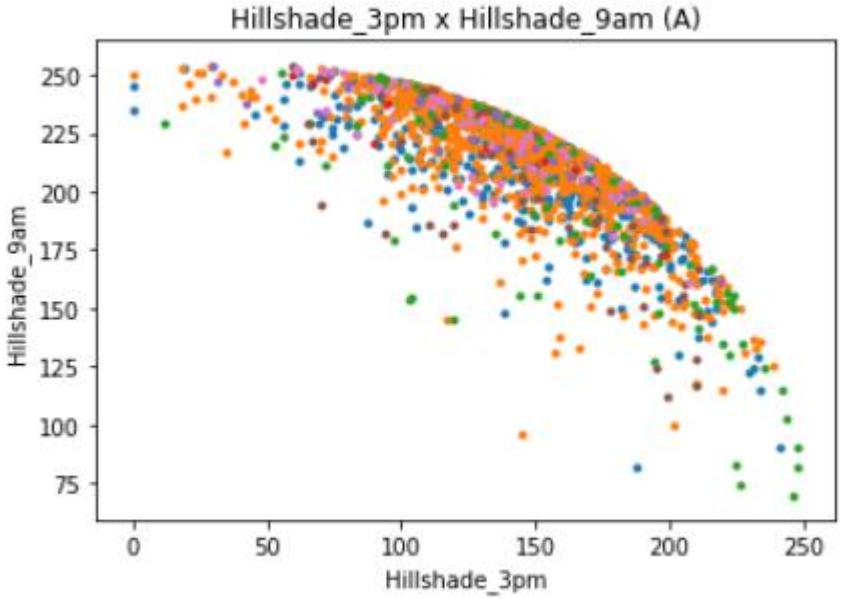


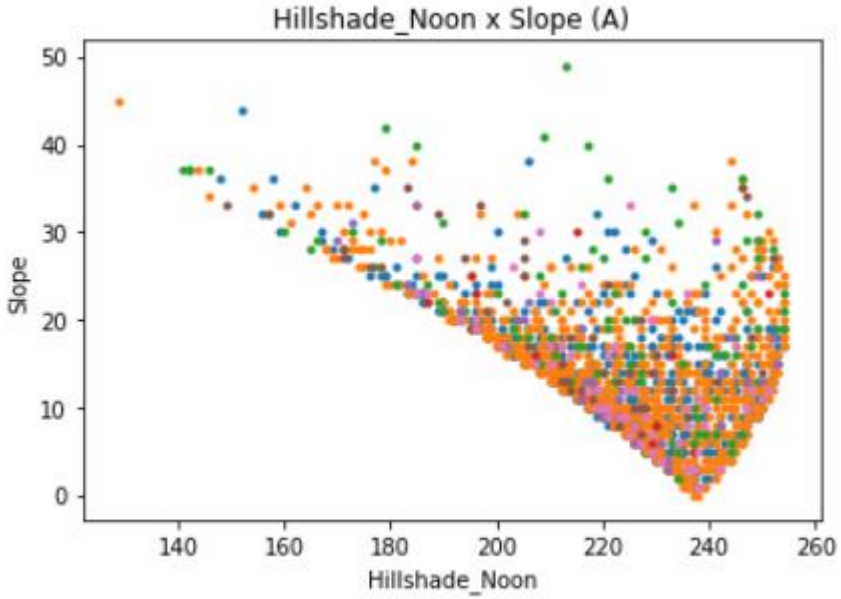



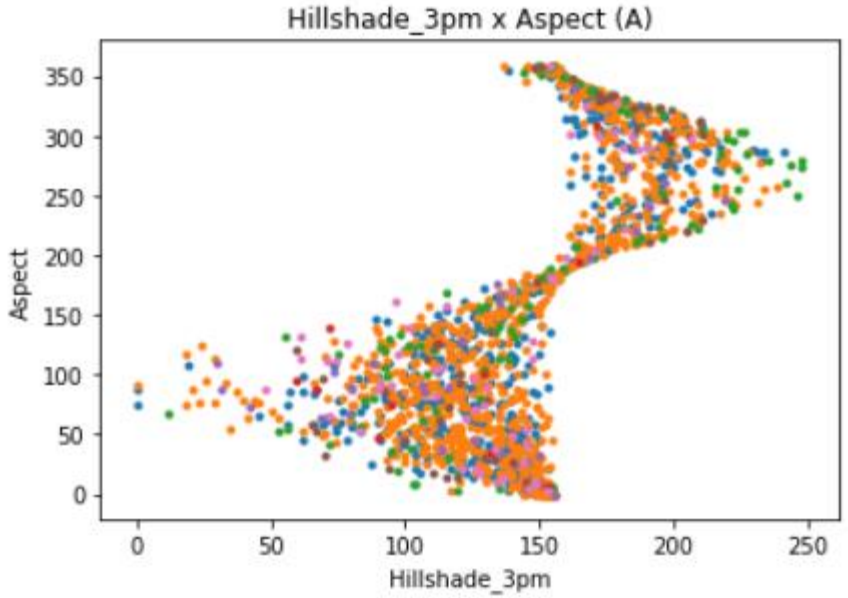



# Experimentação e Escolha dos Algoritmos



# Experimentação e Escolha dos Algoritmos

Decidi pela **árvore de decisão** e pelo **KNN.** Para experimentar com seus parâmetros, dividi a base em 80/20. Realizei então vários K-Fold sobre os 80% para tentar descobrir os melhores parâmetros e atributos para os algoritmos. Por último, com os atributos e parâmetros escolhidos, meço a acurácia obtida treinando sobre os 80 e avaliando contra os 20.

# Experimentação: K-Fold sobre 80% dos dados

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 |

# Experimentação: K-Fold sobre 80% dos dados

```
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
```

# Experimentação: K-Fold sobre 80% dos dados

Árvore de decisão limitada a profundidade N e 3 atributos

| 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 |

# Experimentação: K-Fold sobre 80% dos dados

KNN com 12 atributos

| 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 |

# Experimentação: K-Fold sobre 80% dos dados

KNN com somente 3 atributos

| 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 |

# Validação: testando sobre os 20%

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 |

Obs:
Precision = tp / (tp + fp)
Recall = tp / (tp + fn)
f1-score = media dos dois

# Validação: testando sobre os 20%

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 |

# Fontes das imagens

- Colorado State Forest Service - https://csfs.colostate.edu/colorado-trees/colorados-major-tree-species/#1466527937174-cd5c5e60-5efc
- All Trails - https://www.alltrails.com/explore/us/colorado/buena-vista?ar[]=10119078&ar[]=10151087&ar[]=10151359