Incêndios Florestais

RESUMO:

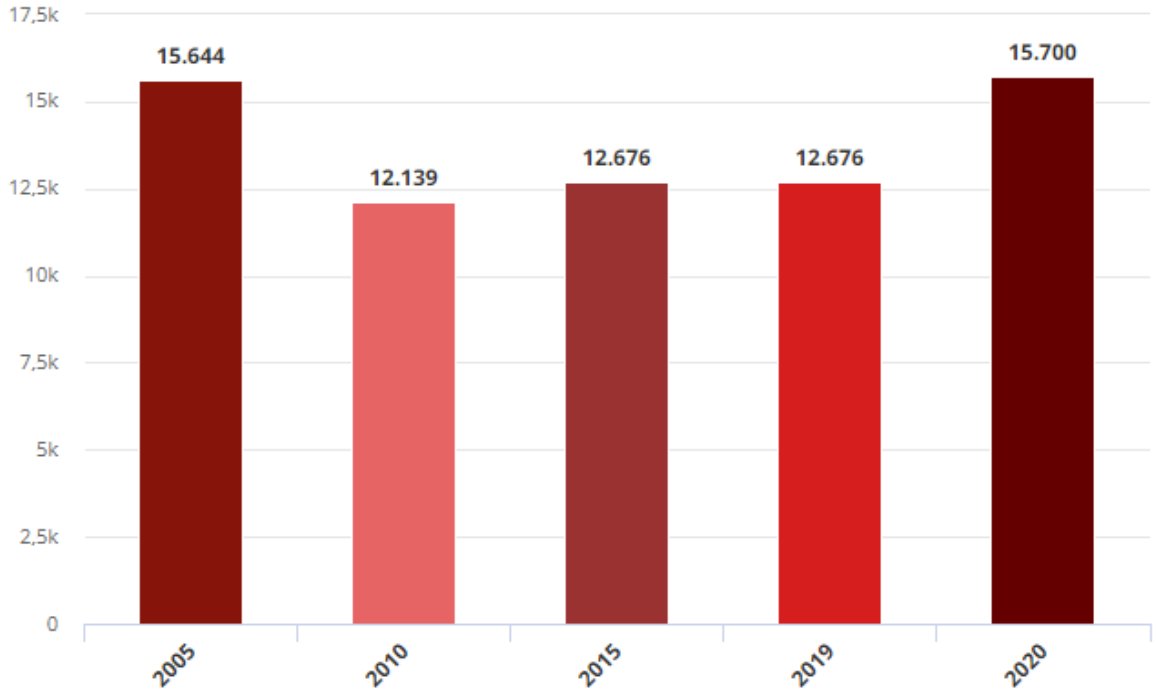
Este artigo resume o Trabalho de Conclusão de Curso de Data Science & Machine Learning da Tera. O nosso trabalho tem como objetivo fazer predições de probabilidade de incêndio florestal (risco de fogo) de das áreas na região amazônica a partir da bases de dado do INPE e INMET, de forma que as entidades públicas ou órgãos responsáveis pela área possam, de forma prévia, agir evitando um desastre ambiental.

INTRODUÇÃO:

Diante do aumento de queimadas observado diariamente em anúncios de jornais e organizações governamentais, a busca para prevenir incêndios e mapeá-los tem se tornado um esforço constante e necessário de órgãos como o INPE.

Os dados revelam uma intensificação no aumento dos focos de incêndio na Floresta Amazônica a partir do início do século XXI, com mais de 100 mil focos em 2002 e com o dobro da quantidade dois anos depois, em 2004. No ano de 2005, houve o pico do número de incêndios, mais de 160 mil focos na região. Em 2010, os números também assustam, com mais de 130 mil focos na região.

A partir de 2011, houve um equilíbrio nos números, os focos de incêndio na Amazônia Legal encontram-se numa média de 50 mil por ano. Em 2020 os alarmes soaram mais forte novamente, pois os números voltaram a patamares que não eram atingidos em muito tempo, como pode ser observado no gráfico abaixo:

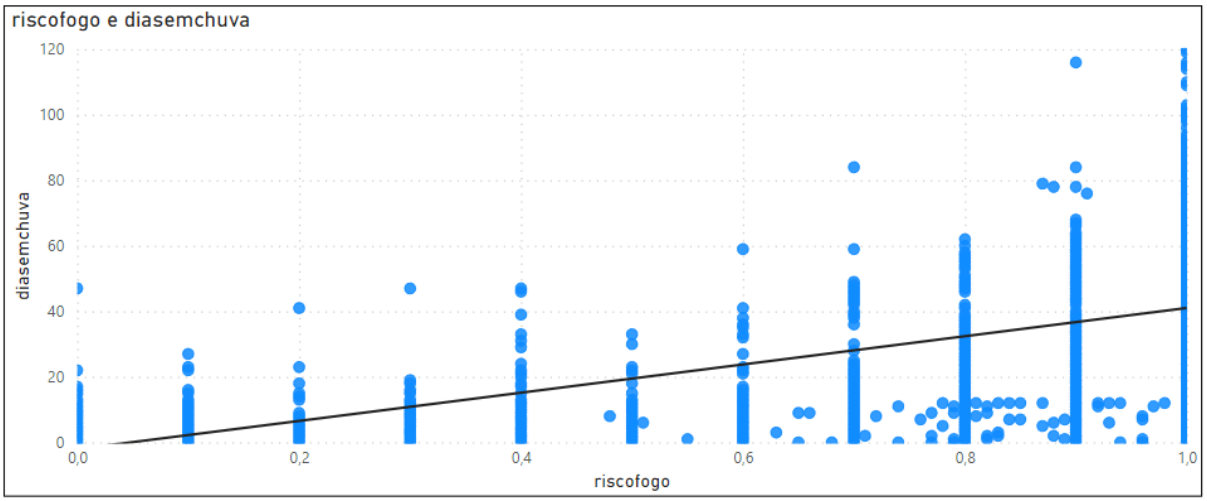


Fonte: Instituto Nacional de Pesquisas Espaciais (Inpe)

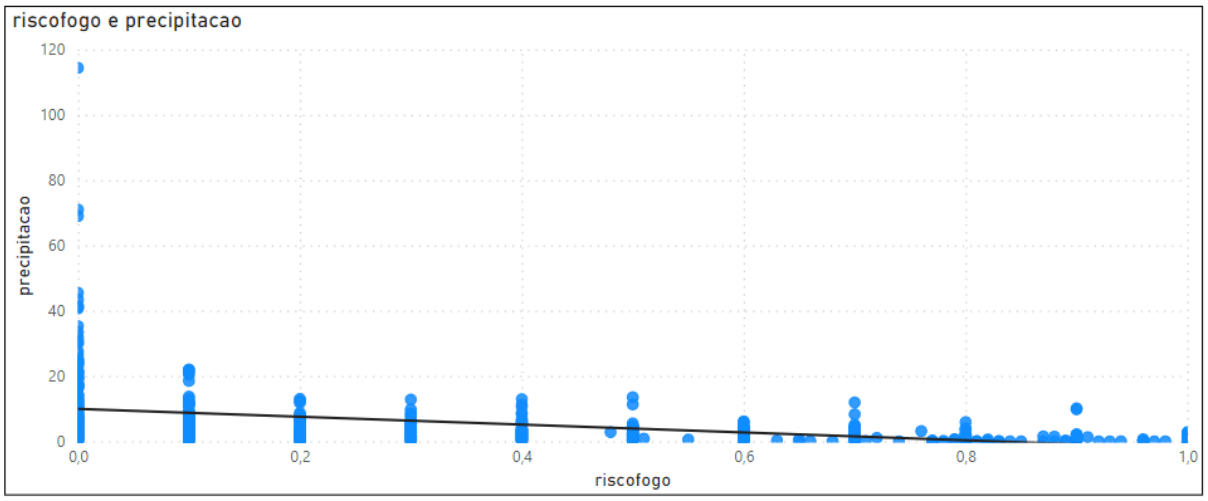
MATERIAIS E MÉTODOS:

Para iniciar nossa busca em entender como as variáveis obtidas afetariam o que o INPE denomina como risco de fogo (“riscofogo”). No artigo destacamos as variáveis com maior impacto no risco de fogo.

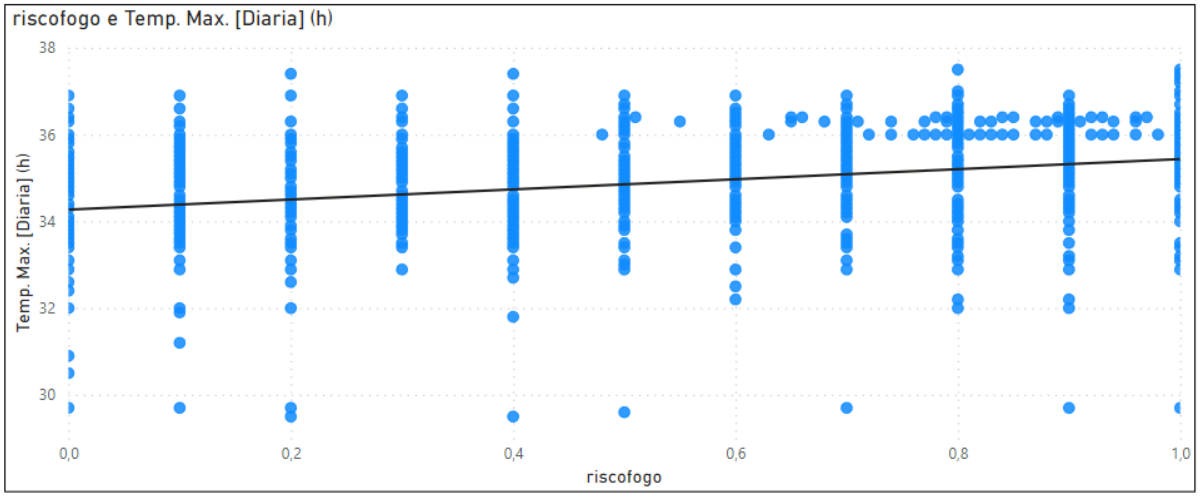
Primeiramente observamos a variável *“diasemchuva”*, que é autoexplicativa e a variável que se destaca em alto impacto no aumento do risco de fogo.



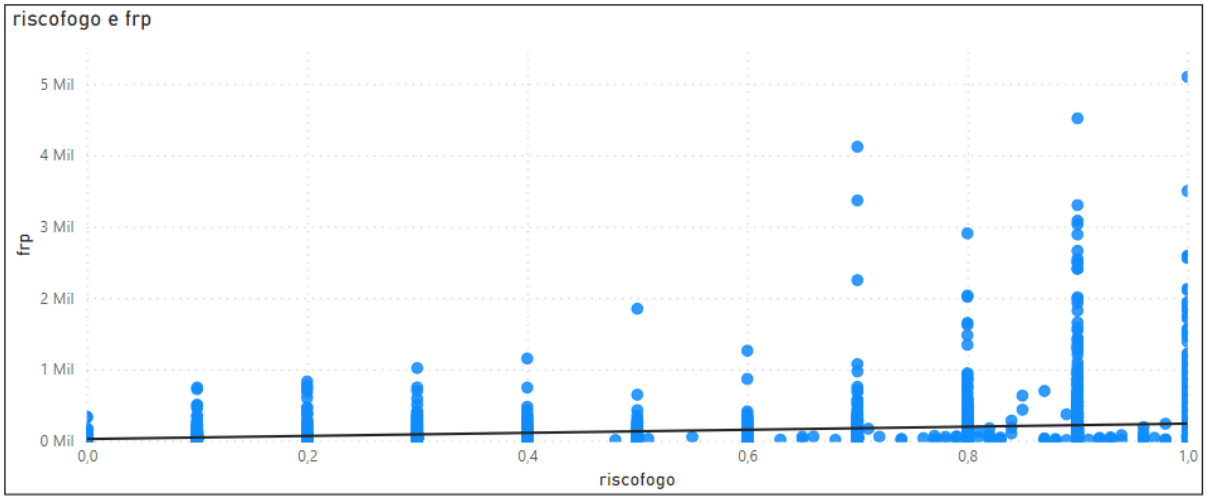
Em sequência destacamos o impacto da precipitação no risco de fogo, que o reduz de maneira considerável quanto maior for o nível de precipitação.



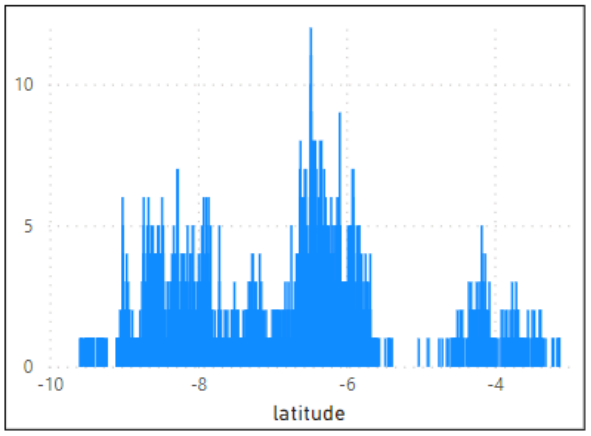
Como já é sabido a temperatura na região da amazônia é constantemente alta e nos dados que obtivemos ela é bastante concentrada em 34°, apesar disso vemos que o aumento desta variável gera um alto impacto no risco de fogo.

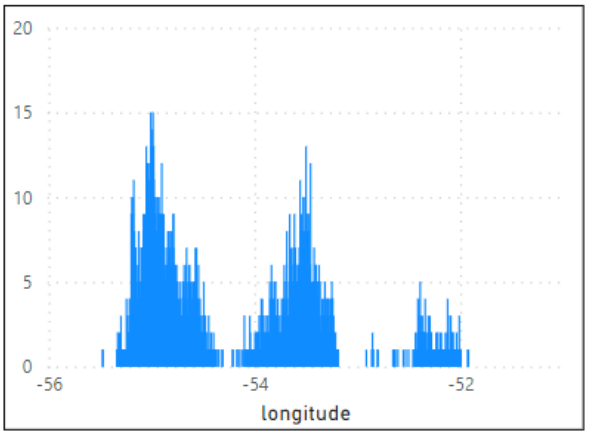


Ao observar a variável *“frp”, que* é um índice de radioatividade medido pelo satélite do INPE, vemos que o aumento dele está ligado ao aumento de risco de fogo, ainda que de maneira sutil.



Ao analisar os dados de latitude e longitude vemos que não temos as informações dispostas da melhor maneira, pois o melhor caso seria ter a mesma quantidade de dados para cada localização.





RESULTADOS E DISCUSSÃO:

Após observarmos estas informações, nosso grupo pensou em maneiras de prever o foco de incêndio e reportá-los de maneira interativa. A princípio pensamos que o melhor caminho para isto seria o uso de séries temporais, uma vez que existe uma sazonalidade conhecida no aumento dos focos de incêndio. Inicialmente separamos apenas a região de Altamira para realizarmos uma análise mais específica, uma vez que esta era a região que possui maior quantidade de informações na base de dados. Porém nos deparamos com dados pouco consistentes em relação a datas, dificultando uma análise temporal.

Sendo assim, partimos para o uso de regressão linear, a fim de prever o risco de fogo dos dados inseridos por um usuário do nosso modelo, baseado no cálculo de uma equação feita pelo modelo de regressão.

Para isso utilizamos o modelo Random Forest Regressor disponibilizado pela biblioteca do sklearn. Este foi o método escolhido para que fosse garantida a possibilidade de observar as principais variáveis que impactam o modelo de

regressão e também para que fosse possível observar o real impacto das variáveis no risco de fogo.

Iniciamos separando 5% dos dados para um teste de “realidade”, tentando diminuir as chances destes dados se adaptarem à média/mediana dos dados. Em seguida separamos os dados restantes em 70 % para treino, 15% para teste e 15% para validação.

```
[ ]  X, X_reality, y, y_reality = train_test_split(X, y, test_size=0.05, random_state=42)

[ ]  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
     X_test, X_valid, y_test, y_valid = train_test_split(X_test, y_test, test_size=0.5, random_state=42)
```

Após o treinamento do modelo com os dados de treino obtivemos a seguinte tabela:

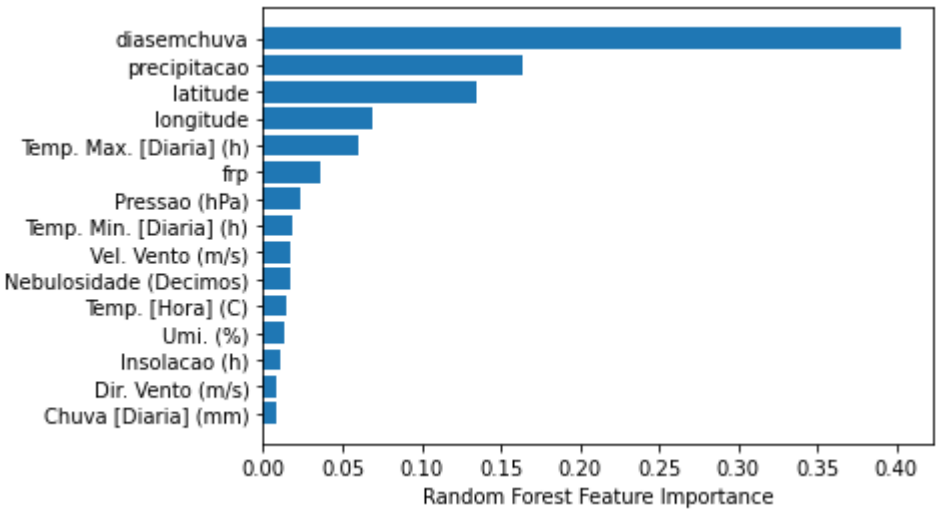
Métrica	Treino	Teste	Validação
RMSE	0.04	0.04	0.04
MAE	0.02	0.02	0.02
R^2	0.98	0.98	0.98
MAPE	0.08	0.08	0.08
Avg. Target	0.75	0.76	0.76
Avg. Prediction	0.75	0.76	0.76

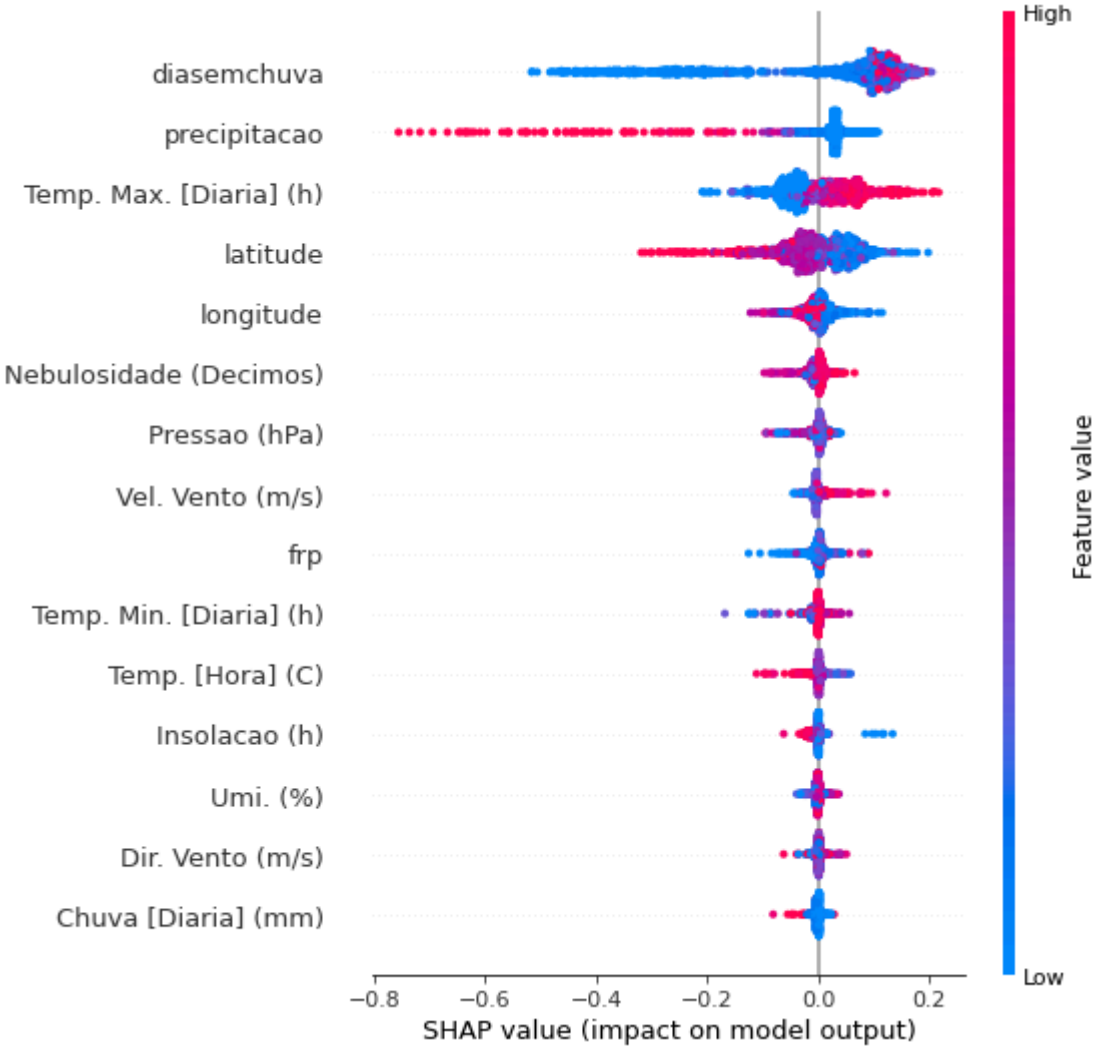
Na tabela acima vemos diversas medidas para avaliar o modelo de regressão. Vemos que o modelo obteve nos dados de treino, teste e validação um R^2 de 0.98 enquanto no teste de “realidade”, 0.85. Poderíamos ter sido mais ambiciosos e buscar aproximar esses valores por meio de “*ensemble*”, mas consideramos um resultado “real” de 85% um bom valor.

Além disso observamos também os valores de RMSE (Raiz quadrada do erro-médio) e MAE (Mean Absolute Error ou Média do Erro Absoluto), que são métricas estatísticas importantíssimas ao medir a performance de um modelo. E como obtivemos erros menores ou próximos a 0.1, decidimos que os resultados eram consistentes e poderíamos seguir com o projeto sem maiores aprimoramentos.

Em seguida partimos para o compreendimento de como as variáveis afetaram o modelo por meio da “feature_importances” disponibilizada pela própria biblioteca do modelo do Random Forest Regressor e em sequência utilizamos o SHAP (SHapley Additive exPlanations) para entender ainda mais o impacto das variáveis no modelo.

Abaixo temos a imagem que representa as variáveis em seu nível de importância no modelo e posteriormente o gráfico de SHAP que se baseia na teoria dos jogos:





O interessante do SHAP é que demonstra de forma visual o impacto da variável na predição do modelo (riscofogo). Lembrando que a coloração se refere ao valor alto ou baixo da variável em si e que o posicionamento dela se refere ao target (riscofogo).

CONCLUSÃO:

Por fim, encaminhamos nosso modelo para o “*deploy”,* pois temos um modelo com boas métricas e temos maneiras de entender e explicar o funcionamento e as métricas dele. Sendo assim, adaptamos nosso modelo e seus resultados para uma execução em um site simples feito por meio do Streamlit, a fim de que qualquer pessoa possa inserir dados para uma predição do risco de fogo. O retorno dessa interação ocorre de maneira muito rápida, o que se faz necessário já que o nosso objetivo é prevenir incêndios e o seu espalhamento.

Algumas melhorias que podem ser realizadas para que o modelo tenha melhores resultados e sejam mais confiáveis seria uniformizar os dados em relação a latitude e longitude, de tal modo que a localização de Altamira não impacte diretamente nos resultados. E também, poderia ser aplicado Ensemble, para alcançar um maior equilíbrio entre treino/teste e validação/realidade.

REFERÊNCIAS:

- https://shap.readthedocs.io/en/latest/