

**CRASH**   **IA**

# Acidentes de Trânsito e IA

Será que existe uma forma de podermos mitigar problemas dos acidentes de trânsito, baseado em dados?

# INTEGRANTES



**Felipe Bevilacqua**



**Filipe Aguiar**



**Thiago Kido**



**Victor Pereira**



Segundo o Ministério da Saúde, a cada **15 MINUTOS** uma **MORTE** é registrada por **ACIDENTE** no Brasil, são **34 mil** pessoas **MORTAS** por ano.


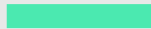
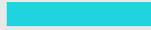
*Dados do  
Ministério da  
Saúde*



Esse problema **CUSTA** para a **SOCIEDADE BRASILEIRA** cerca de **50 BILHÕES** de reais por ano. São cerca de **40 BILHÕES** com acidentes em **RODOVIAS** e **10 BILHÕES** em **ÁREAS URBANAS**.

# CONTEXTO DA SOLUÇÃO

**Uma possível solução poderia interessar a diversos grupos, como por exemplo: motoristas, órgãos públicos, concessionárias de rodovias e empresas do segmento logístico. A partir disso, algumas perguntas aparecem dentro deste contexto:**

-  Onde podemos alocar investimentos e recursos de forma a aumentar a segurança nas estradas?
-  Neste momento, quais os pontos mais prováveis de ocorrer um acidente?
-  Dado um acidente qualquer, podemos prever a gravidade do acidente ?

## Valores Nulos

### Valores Numéricos

- Outliers
- Nulos



**Substituídos  
pela média da  
coluna**

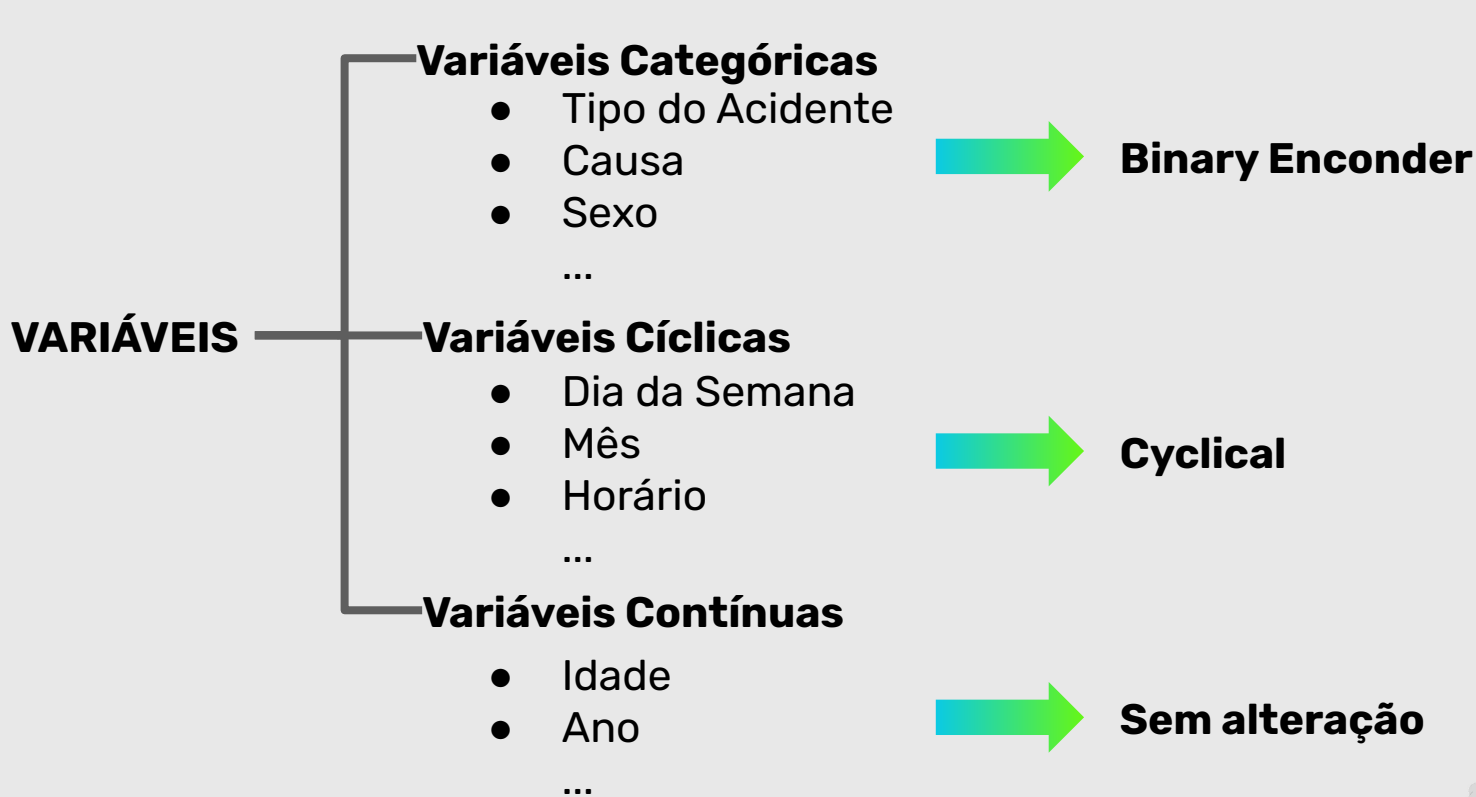
### Valores Categóricos

- Não Informados
- Ignorados
- Nulos



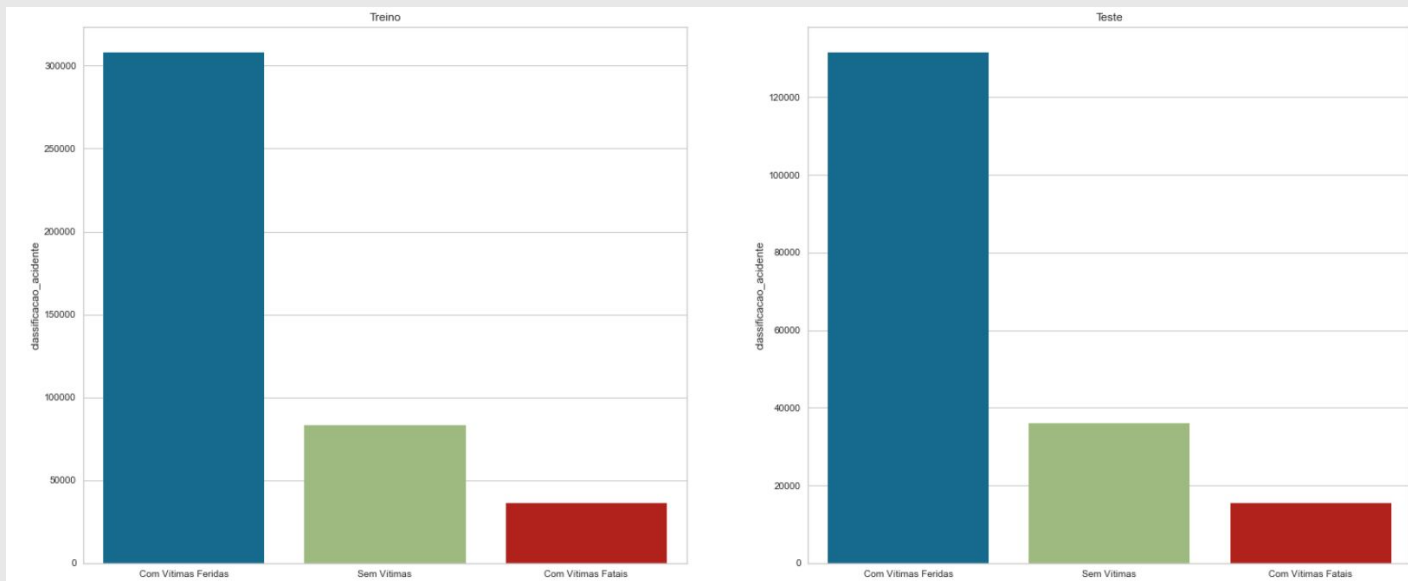
**Substituídos  
de forma  
aleatória,  
respeitando a  
distribuição**

# TRANSFORMAÇÕES DAS VARIÁVEIS





# TARGET



# MODELO CLASSIFICATÓRIO

## Escolhendo o Modelo

Observando os valores de F1, o RandomForestClassifier apresentou o melhor resultado.

Modelo a ser utilizado



	Acurácia %	Classificação	Precision	Recall	f1	AUC
XGBClassifier	71.26	Sem vítimas	0.40	0.08	0.14	0.58
		Com vítimas	0.73	0.96	0.83	0.55
		Fatais	0.32	0.04	0.07	0.63
DecisionTreeClassifier	52.14	Sem vítimas	0.21	0.26	0.23	0.51
		Com vítimas	0.72	0.63	0.67	0.50
		Fatais	0.27	0.19	0.15	0.53
RandomForestClassifier	72.06	Sem vítimas	0.43	0.02	0.03	0.58
		Com vítimas	0.72	1.00	0.84	0.55
		Fatais	0.27	0.00	0.00	0.66
GaussianNB	68.07	Sem vítimas	0.23	0.01	0.03	0.54
		Com vítimas	0.72	0.93	0.81	0.49
		Fatais	0.13	0.09	0.11	0.51
KNeighborsClassifier	68.30	Sem vítimas	0.27	0.09	0.13	0.54
		Com vítimas	0.73	0.91	0.81	0.54
		Fatais	0.21	0.09	0.12	0.58

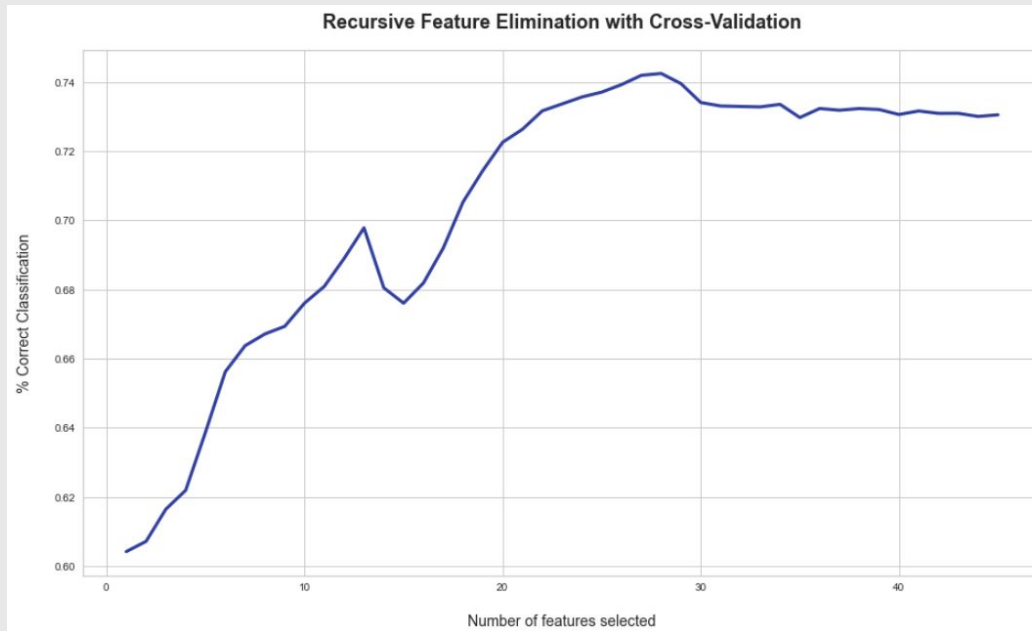
# APLICAÇÃO DO RFECV

## RFECV

Aplicação do RFECV, analisando a performance do F1 ponderado através de diferentes números de features utilizadas, a fim de definir o número ótimo de features para o RandomForestClassifier.

### Número ótimo de features

RFECV resultou em 27 features:



# BALANCEAMENTO

## Escolhendo o Modelo

Realizado o balanceamento e utilizando o Random Forest, escolhemos o modelo utilizando o mesmo critério anterior, ou seja, o F1.

**Modelo a ser utilizado**



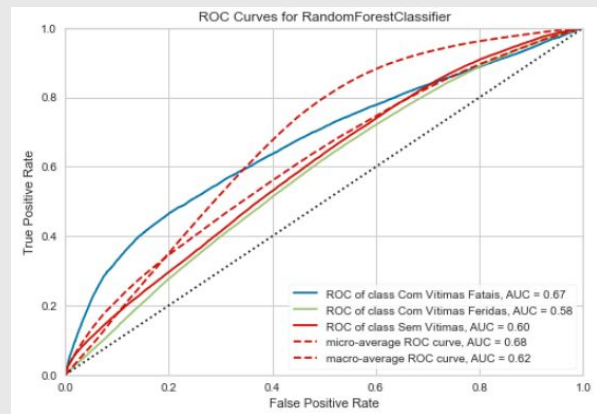
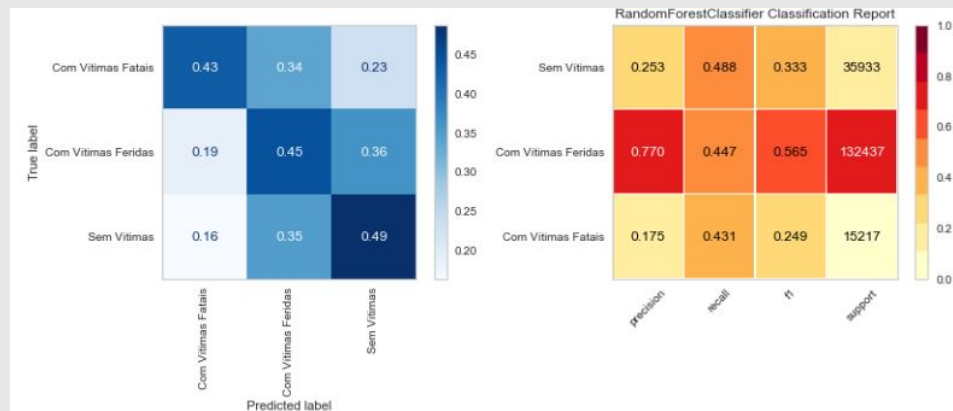
	Acurácia %	Classificação	Precision	Recall	F1	AUC
Normal	72.06	Sem vítimas	0.43	0.02	0.03	0.58
		Com vítimas	0.72	1.00	0.84	0.55
		Fatais	0.27	0.00	0.00	0.66
Undersampling	46.02	Sem vítimas	0.25	0.45	0.32	0.60
		Com vítimas	0.76	0.47	0.58	0.57
		Fatais	0.17	0.43	0.24	0.66
Oversampling	70.96	Sem vítimas	0.39	0.09	0.15	0.65
		Com vítimas	0.73	0.95	0.83	0.57
		Fatais	0.27	0.04	0.07	0.61
SMOT	60.30	Sem vítimas	0.22	0.16	0.19	0.55
		Com vítimas	0.72	0.78	0.75	0.52
		Fatais	0.09	0.08	0.09	0.50
ADASYN	60.60	Sem vítimas	0.23	0.17	0.19	0.55
		Com vítimas	0.72	0.79	0.75	0.51
		Fatais	0.12	0.08	0.09	0.50
Undersampling + Class Weights	53.49	Sem vítimas	0.26	0.42	0.32	0.59
		Com vítimas	0.75	0.60	0.67	0.56
		Fatais	0.20	0.24	0.22	0.65
<b>Undersampling + Class Weights + Grid Search</b>	<b>45.35</b>	<b>Sem vítimas</b>	<b>0.25</b>	<b>0.49</b>	<b>0.33</b>	<b>0.60</b>
		<b>Com vítimas</b>	<b>0.77</b>	<b>0.45</b>	<b>0.57</b>	<b>0.58</b>
		<b>Fatais</b>	<b>0.18</b>	<b>0.43</b>	<b>0.25</b>	<b>0.67</b>

## Performance

Modelo performou de forma mais balanceada, porém os acertos para cada, permanecem baixos, com uma tendência de “chutar” com vítimas feridas

## Métricas

Recall balanceado e Precision baixo e desbalanceado consequentemente F1 apresentou valores baixos e desequilibrados.



# AVALIAÇÃO DOS RESULTADOS

Consideramos satisfatórios os resultados até aqui alcançados?

# REVISÃO DOS DADOS FILTRADOS

## Retorno de todos os dados iniciais

No primeiro modelo havíamos retirado algumas colunas que não pareciam necessárias para o projeto

## Correção dos dados de latitude e longitude

Dados corrigidos através da biblioteca **geopy**

	Acurácia %	Classificação	Precision	Recall	F1	AUC
XGBClassifier	67.79	Sem vítimas	0.26	0.13	0.18	0.56
		Com vítimas	0.73	0.91	0.81	0.55
		Fatais	0.46	0.02	0.03	0.66
DecisionTreeClassifier	56.59	Sem vítimas	0.23	0.24	0.24	0.52
		Com vítimas	0.73	0.71	0.72	0.52
		Fatais	0.11	0.13	0.12	0.52
RandomForestClassifier	70.39	Sem vítimas	0.25	0.04	0.07	0.54
		Com vítimas	0.72	0.97	0.83	0.52
		Fatais	0.27	0.00	0.01	0.57
GaussianNB	62.93	Sem vítimas	0.14	0.07	0.09	0.55
		Com vítimas	0.71	0.85	0.77	0.48
		Fatais	0.10	0.05	0.06	0.49
KNeighborsClassifier	68.32	Sem vítimas	0.30	0.11	0.16	0.56
		Com vítimas	0.73	0.91	0.81	0.56
		Fatais	0.24	0.10	0.14	0.60

# APLICAÇÃO DO RFECV

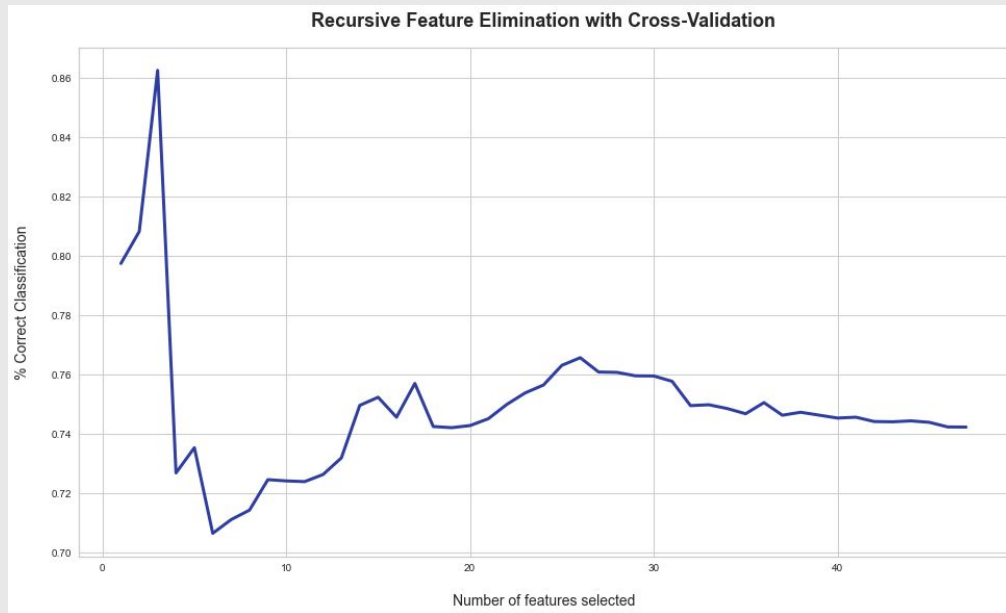
## RFECV

Aplicação do RFECV, analisando a performance do F1 ponderado através de diferentes números de features utilizadas, a fim de definir o número ótimo de features para o RandomForestClassifier.

### Número ótimo de features

RFECV resultou em 3 features:

- horario\_sin
- latitude
- longitude





# MODELO FINAL

## 3 principais features

Modelo baseado nas 3 features destacadas pelo RFECV

## Prioridades

Acertos em mortos e feridos

## Melhor modelo

RandomForestClassifier - Normal

	Acurácia %	Classificação	Precision	Recall	F1	AUC
Normal	89.56	Sem vítimas	0.86	0.66	0.74	0.92
		Com vítimas	0.90	0.97	0.93	0.93
		Fatais	0.91	0.86	0.88	0.96
Undersampling	60.63	Sem vítimas	0.36	0.63	0.46	0.75
		Com vítimas	0.88	0.57	0.69	0.78
		Fatais	0.39	0.90	0.54	0.94
Oversampling	88.87	Sem vítimas	0.79	0.70	0.74	0.92
		Com vítimas	0.91	0.94	0.93	0.94
		Fatais	0.88	0.86	0.87	0.96
SMOT	87.33	Sem vítimas	0.76	0.72	0.74	0.92
		Com vítimas	0.93	0.91	0.92	0.94
		Fatais	0.72	0.88	0.79	0.97
ADASYN	87.23	Sem vítimas	0.76	0.72	0.74	0.92
		Com vítimas	0.93	0.91	0.92	0.94
		Fatais	0.70	0.88	0.78	0.97

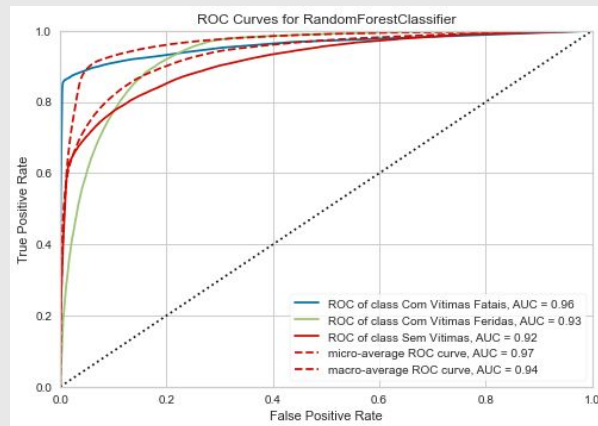
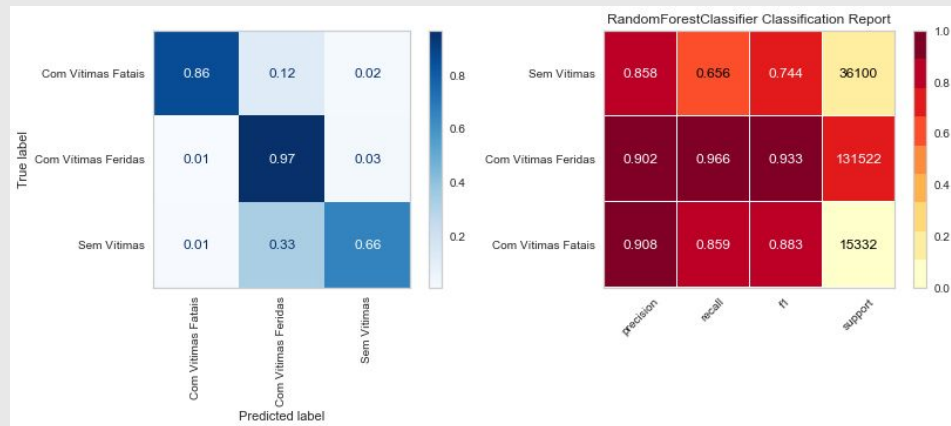
# RFC - RESULTADO

## Performance

Modelo performou melhor com 3 variáveis do que com 27 da primeira versão

## Métricas

Recall, Precision e consequentemente F1 apresentaram melhores valores para todos os labels, bem como AUC também foram mais relevantes.



# APLICAÇÃO

Características do acidente

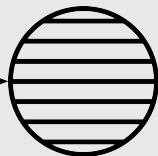


Localização do acidente

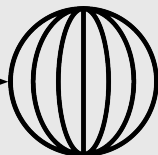


Horário do acidente

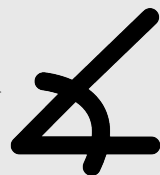
GeoPy



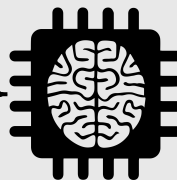
Latitude



Longitude



Horário seno



Modelo de Machine Learning

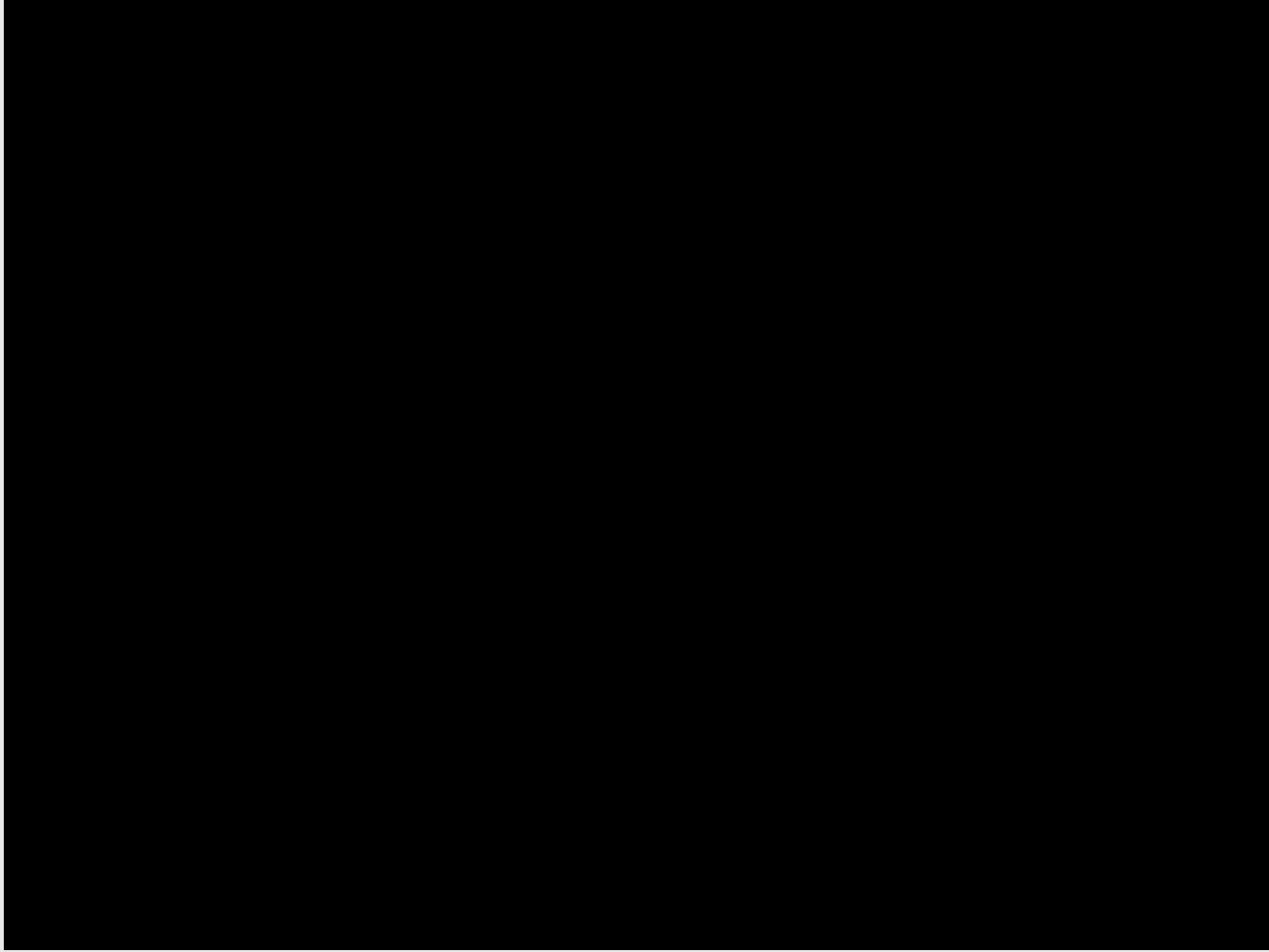
Resposta do modelo

COM VÍTIMAS  
FATAIS

COM VÍTIMAS  
FERIDAS

SEM VÍTIMAS  
FERIDAS

# APLICAÇÃO



# CONCLUSÃO

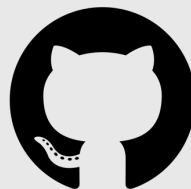
- Modelo bem acima ao puro acaso que seria de 33% de acerto para cada uma das classes.
- Foco no f1-weighted, acurácia não se mostrou eficiente como métrica.
- Deixar com que os dados nos digam por si, evitando vieses de seleção.
- Sem uma variável determinante apesar dos esforços pode não ser possível chegar a um modelo satisfatório
- Modelo é feito para o domínio de acidentes rodoviários.

# PRÓXIMOS PASSOS

- Adição de modelo em servidor;
- Redução do tamanho do modelo;
- Testar o modelo em campo, de forma a testar sua real aplicabilidade e eficiência;
- Entender mais profundamente quais são os vieses de seleção dos acidentes que hoje são registrados

**Obrigado !**

**Acesse:**



**CRASH**   **IA**