TECH CHALLENGE

Kalil Gadben de Souza - **Pós Tech - 7IADT - Grupo 113**

Vídeo apresentação: https://youtu.be/xAX7XaGqEUc

Desafio

Construir uma solução inicial com foco em IA para processamento de exames médicos e documentos clínicos, aplicando fundamentos essenciais de IA, Machine Learning e Visão Computacional.

Tema Escolhido

Identificar possíveis pessoas com Diabetes baseados em dados de seus exames.

Dados e Modelos

Nesse projeto, utilizei como fonte de dados, o dataset sugerido de diagnóstico de diabetes:
https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data

Exploração de dados

Ao visualizar os dados de forma superficial, pude notar um certo padrão onde os pacientes com diagnóstico de diabetes possuíam a Glicose acima da média:

```
dados.head()
```

	gestacoes	glicose	pressao_arterial	espessura_pele	insulina	imc	predisposicao_genetica_diabetes	idade	diabetes
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
dados.describe()
```

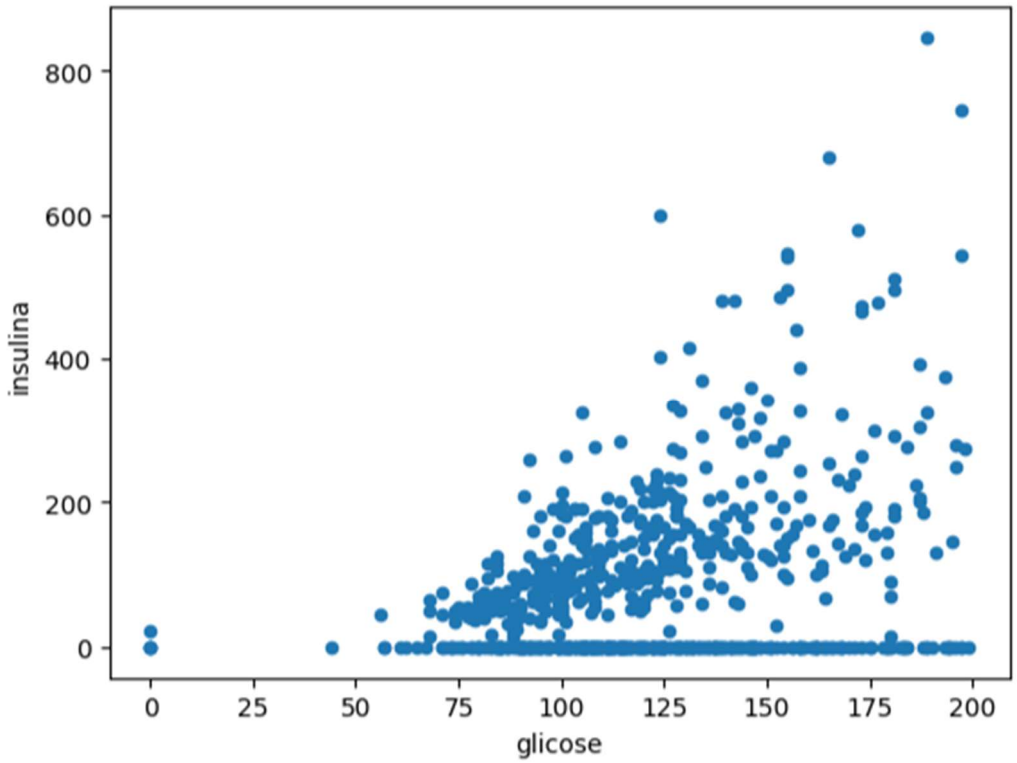
	gestacoes	glicose	pressao_arterial	espessura_pele	insulina	imc	predisposicao_genetica_diabetes	idade	diabetes
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Explorei a ideia de que poderia existir uma relação entre Glicose e Insulina, e pelo gráfico foi possível verificar que de fato elas possuem alguma relação:

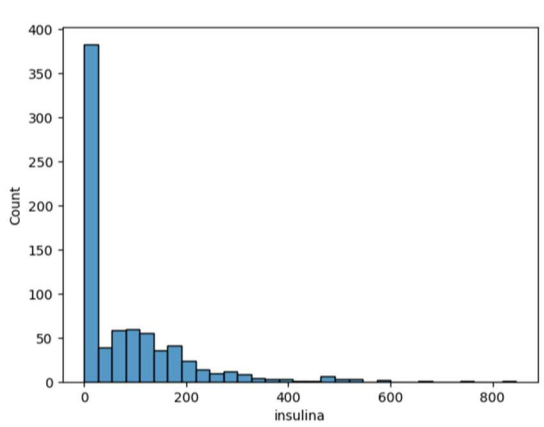
diabetes	glicose	pressao_arterial	espessura_pele	insulina	imc	idade
0	109.980000	68.184000	19.664000	68.792000	30.304200	31.190000
1	141.257463	70.824627	22.164179	100.335821	35.142537	37.067164

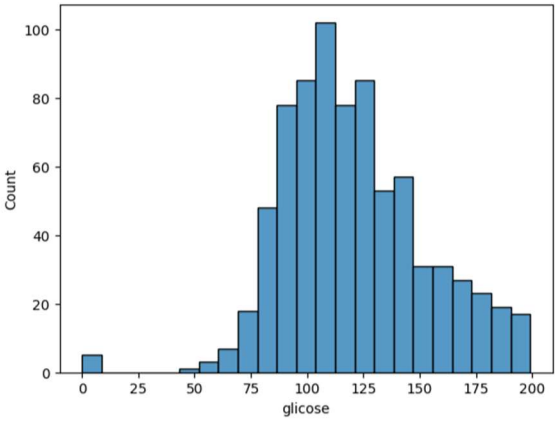
```
dados.plot.scatter(x="glicose", y="insulina")
```

```
<Axes: xlabel='glicose', ylabel='insulina'>
```



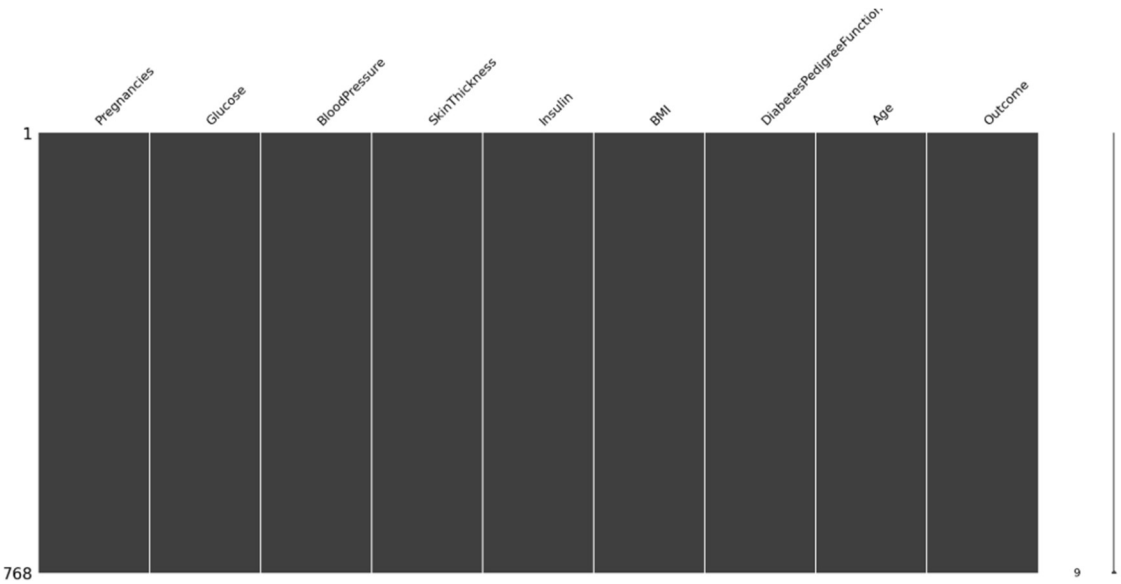
No entanto, foi possível verificar através de novos gráficos que a incidência de baixa insulina era maior na base de dados, enquanto da Glicose apresentava curva mais relevante:





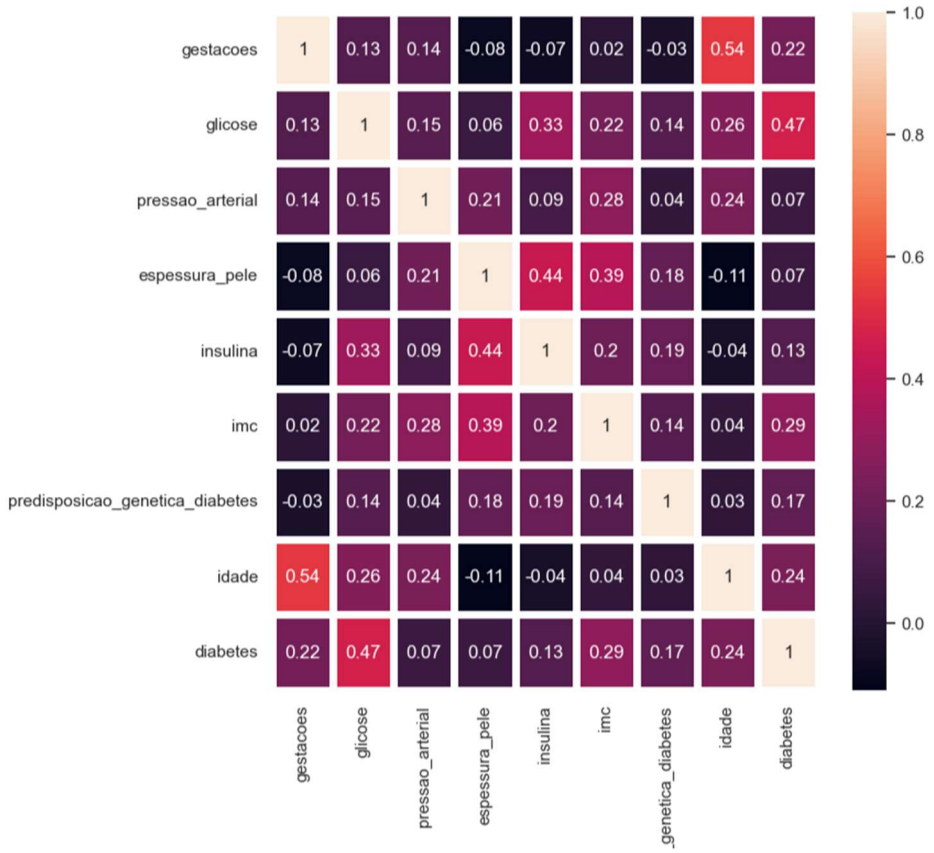
Estratégias de pré-processamento

Foi verificado a estrutura de dados existentes no dataset, a fim de encontrar campos nulos e campos com valor em texto (categóricas), no entanto não foram necessários ajustes, pois todas as colunas estavam preenchidas com números.



Os dados não foram escalonados para que eu pudesse interagir com o modelo de forma semelhante a consultar dados de um paciente específico após o treino dos modelos.

Foi feita a análise de correlação, e identificadas as colunas de maior relevância:



Modelagem

Os modelos escolhidos foram o KNN e o SVM.

Após efetuar a análise de correlação, reduzi o número de colunas para as que achei que seriam mais relevantes:

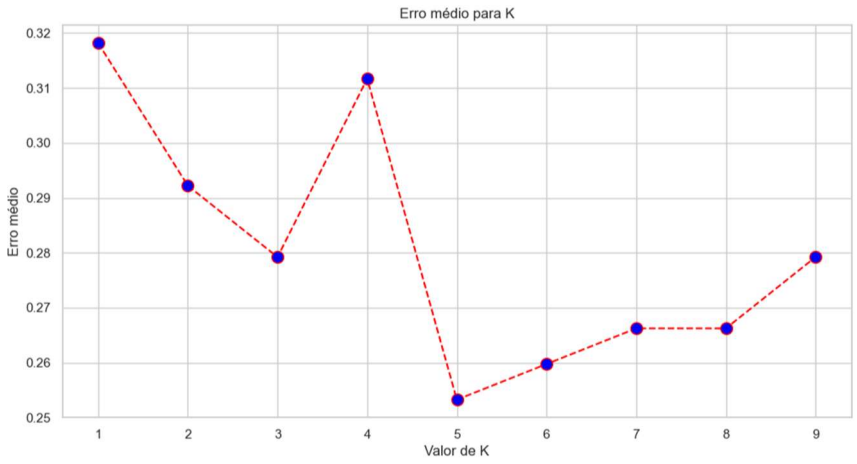
```
novo_x = dados[["gestacoes", "glicose", "insulina", "imc", "predisposicao_genetica_diabetes","idade"]]
novo_y = dados["diabetes"]
```

Para ambos os modelos, foram utilizados 20% dos dados para teste e 80% para treino.

```
nx_treino, nx_teste, ny_treino, ny_teste = train_test_split(novo_x, novo_y, test_size = 0.2, stratify = novo_y, random_state = 13)
```

Treinamento e avaliação do modelo

Para o modelo KNN, efetuei a busca para obter o melhor número para o K, e consegui encontrar o valor de k = 5:

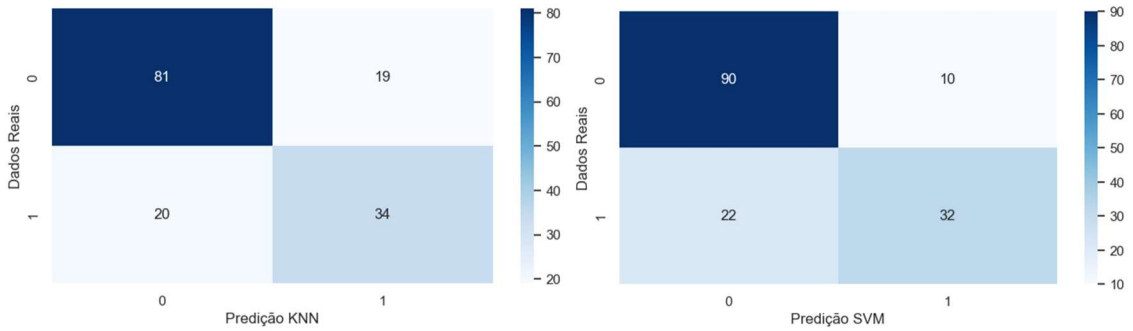


E após rodar os modelos KNN (com K=5) e o SVM (padrão), obtive a melhor acurácia através do modelo SVM, sendo:

- KNN: 0.7467532467532467

- SVM: 0.7922077922077922

Para garantir meu entendimento de que o SVM seria o melhor modelo para meu projeto, gerei os gráficos de matriz de confusão para ambos:



E com isso foi possível notar que o SVM foi mais assertivo que o KNN para os casos de pacientes sem Diabetes, e o KNN foi mais assertivo para os casos onde o paciente possui Diabetes.

Mas para ter uma visão mais especifica, gerei o ClassificationReport, onde ficou mais evidente a assertividade do modelo SVM:

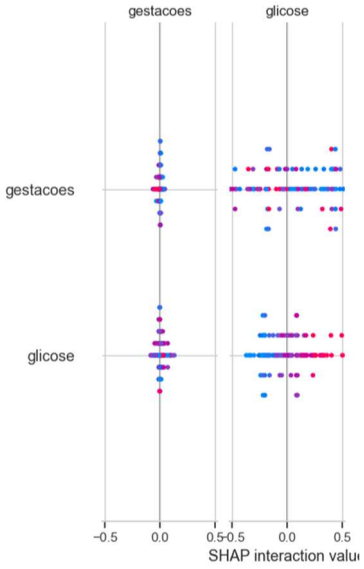
Resultado KNN

	precision	recall	f1-score	support
0	0.80	0.81	0.81	100
1	0.64	0.63	0.64	54
accuracy			0.75	154
macro avg	0.72	0.72	0.72	154
weighted avg	0.75	0.75	0.75	154

Resultado SVM

	precision	recall	f1-score	support
0	0.80	0.90	0.85	100
1	0.76	0.59	0.67	54
accuracy			0.79	154
macro avg	0.78	0.75	0.76	154
weighted avg	0.79	0.79	0.79	154

E por fim, gerei uma análise com a técnica SHAP para verificar quais foram as colunas com maior relevância para o resultado, e fiquei surpreso de ver que a quantidade de gestações foi mais relevante do que a insulina e a pré-disposição genética:



Conclusão

Acredito que o modelo poderia ter maior número de casos positivos para que o modelo pudesse melhorar sua acurácia, mas ainda assim, essa solução poderia ser utilizada para ajudar médicos a triar pacientes e até mesmo antecipar precauções em casos de internação, onde já se tem as informações necessárias para avaliar se é possível ter diabetes.