

|||||

Atividade prática 01

Machine Learning

Professor: Aydano Machado

Alunos: Ascânio Sávio, Elyssana Oliveira, Eirene Fireman

Resumo

01

Nossa solução

O que fizemos e o porquê.

02

Outras tentativas

Outras implementações que fizemos, mas não melhorou a acurácia.

03

Indo além

Soluções que não conseguimos implementar e pontos de melhoria que identificamos.

04

Dúvidas

Dúvidas que tivemos ao longo do caminho.





01

Nossa solução

O que fizemos e o porquê.

Nosso percurso...

Problema

Tentamos entender o problema de acordo com os dados fornecidos

1



2

Testes

Fomos testando soluções diferentes



Solução

Escolhemos a solução com maior acurácia

3



Discovery

Pregnancies	Gestações múltiplas e diabetes gestacional podem influenciar para que uma mulher tenha diabetes (?).
Glucose	Uma taxa de glicose alta pode ser indicativo de diabetes. Acima de 100 mg/l/d foge da normalidade.
BloodPressure	A pressão arterial elevada é um fator de risco para o desenvolvimento de diabetes, além disso, muitas pessoas com diabetes tipo 2 também têm hipertensão.
SkinThickness	Medida da espessura da dobra de pele no tríceps. Em alguns casos uma medida alta pode estar acompanhada à obesidade ou ao risco de doenças metabólicas, como diabetes tipo 2.
BMI	Medida que indica se uma pessoa está dentro do seu peso ideal de acordo com a sua altura.
DiabetesPedigreeFunction	Medida que avalia a predisposição genética de uma pessoa para desenvolver diabetes tipo 2. Quanto mais perto de 1.0, maior o risco.
Age	Pode ter relação com a diabetes. A forma mais comum da doença, é geralmente diagnosticada em adultos mais velhos. E com a idade a produção de insulina pode ser prejudicada.
Insulin	A diabetes é uma doença em que o corpo não produz insulina suficiente ou não consegue usar efetivamente a insulina que produz.



Title

```
1 import pandas as pd
2 import requests
3 from sklearn.decomposition import PCA
4 from sklearn.feature_selection import SelectKBest, f_classif
5 from sklearn.impute import SimpleImputer
6 from sklearn.neighbors import KNeighborsClassifier
7 from sklearn.preprocessing import MinMaxScaler, StandardScaler
```

Bibliotecas utilizadas.



Title

```
1 print('\n - Lendo o arquivo com o dataset sobre diabetes')
2 data = pd.read_csv('diabetes_dataset.csv')
3
4 data = data.drop('Pregnancies', axis=1) # -> Drop pregnancies from dataset
5 data = data.drop('SkinThickness', axis=1) # -> Drop SkinThickness
```

Dropamos as colunas: Skin Thickness e Pregnancies. Pois chegamos a conclusão que elas tinham menos relevância do que as outras colunas no diagnóstico de diabetes.




|||||



Title

```
1 columns_to_normalize = X.columns
2
3 scaler = MinMaxScaler()
4 X[columns_to_normalize] = scaler.fit_transform(X[columns_to_normalize])
```

Por fim, normalizamos os dados, pois os mesmos tinham diferenças grandes. Assim conseguimos melhorar a acurácia usando o MinMax.



02 Outras tentativas

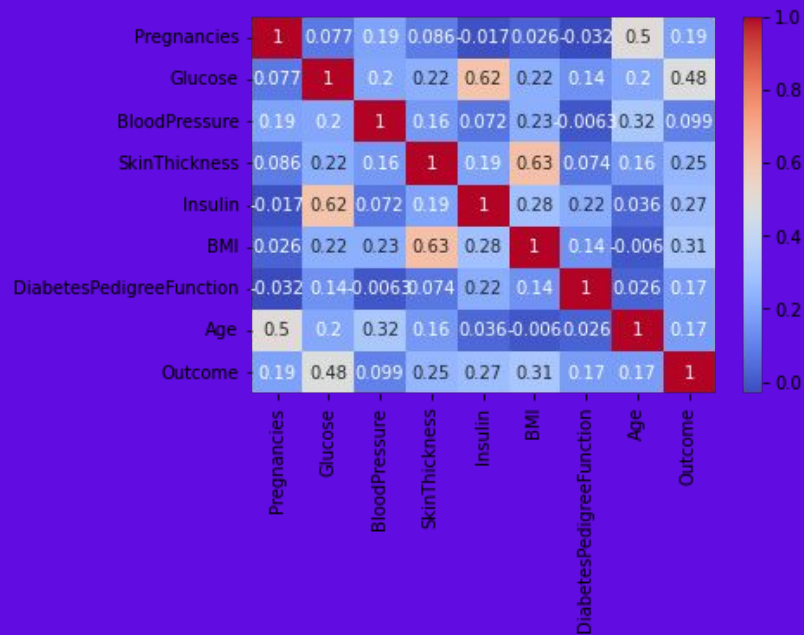
Outras implementações que fizemos, mas não melhorou a acurácia.



Correlação entre os dados

Tentamos utilizar a matriz de correlação para verificar a correlação entre os dados.

- threshold de 0.5,
- Será que variar o threshold nos ajudaria a melhorar nosso modelo?



Mais tentativas...



Algoritmo PCA

Visando a redução de dimensionalidade. Porém, mesmo variando o número de componentes não conseguimos aumentar a acurácia do modelo.



Agrupamento

Agrupamento dos dados por faixa etária, pelas taxas de insulina, glicose e pressão. Mas a acurácia caiu.



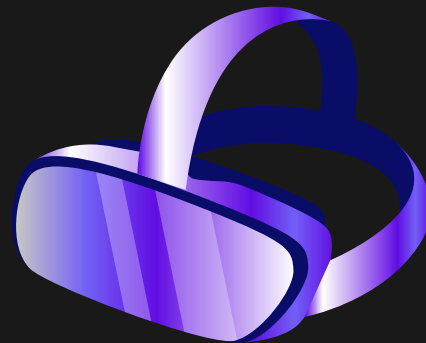
03 Indo além

Soluções que não conseguimos
implementar e pontos de melhoria que
identificamos.

Adicionar ruídos no dataset

A fim de obter uma possível melhora na acurácia, pois pensamos que nosso modelo estava sofrendo overfitting.

Porém, não conseguimos implementar a tempo, estávamos tendo bastante erros.





04 Dúvidas

Dúvidas que tivemos ao longo do caminho.

Dúvidas que tivemos ao longo do caminho...

- Como poderíamos ter implementado os agrupamentos de uma maneira eficiente?
- Existe uma alternativa mais assertiva para inferir os dados sem ser por moda, média e mediana? (que foram os modos que testamos)

Agradecemos a atenção!

CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#), infographics and images by [Freepik](#)

Please keep this slide for attribution

