Prova 1 — IA (Respostas)

Aluno: Marcílio de Oliveira Silva Júnior.
Matrícula: 11413589.
Professora: Thaís Gaudencio do Rêgo.

Links das bases de dados

Datasets para questão 2

- Classificação: https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction
- Regressão: https://www.kaggle.com/rikdifos/credit-card-approval-prediction
- Classificação / Regressão: https://www.kaggle.com/rahulgupta21/cardiac-risk-prediction

Datasets para questão 3

- Supervisionado: https://www.kaggle.com/andrewmvd/autism-screening-on-adults
- Não-Supervisionado: https://www.kaggle.com/ashwinik/consumer-complaints-financial-products
- Supervisionado / Não-Supervisionado: https://www.kaggle.com/adhyanmaji31/breast-cancer-prediction

---

Sumário

Questão 1

Um desafio de um museu de arte tem como objetivo prever os autores de obras de arte do impressionismo. Defina o ambiente e suas categorias, como na Aula de Agentes Inteligentes, os atuadores, os sensores e a(s) medida(s) de desempenho do sistema com inteligência artificial proposto.
Explique demonstrando cinco exemplos de entrada e da saída (distintos). Especifiquem se a entrada será a imagem e seus píxeis ou outras características da imagem (descreva-as). Além disso, detalhe de onde virão os dados, se já existe alguma base ou vocês irão sugerir a construção de uma nova e como seria.

Resposta

Para este desafio, precisamos primeiramente entender um pouco mais sobre o universo onde o sistema com inteligência artificial irá atuar, de modo que possamos modelar todo o projeto de forma coerente. Sendo assim, precisamos estar a par das características do impressionismo.
Algumas características do impressionismo são listadas abaixo.

- As artes geralmente são feitas em ambiente aberto, com intuito de captar as diferentes tonalidades da cena ao refletir a luz solar em determinado momento.
- As obras não possuem contornos nítidos em seus elementos.
- Sombras devem ser retratadas com cores luminosas e coloridas. Evita-se preto em obras impressionistas.
- Contrastes de luz e sombra devem ser obtidos de acordo com a lei das cores complementares.
- As cores devem ser puras e dissociadas.

Agora que conhecemos melhor a arte impressionista, é importante que conheçamos, também, as características de alguns artistas que são impressionistas. Isto é necessário para que o objetivo "*… prever os autores de obras de arte do impressionismo.*". Abaixo temos algumas informações concisas sobre alguns autores.

- **Claude Monet**: pinturas ao ar livre; natureza e luzes do céu como foco principal.
- **Pierre-Auguste Renoir**: pinturas otimistas, alegres; intensa movimentação; preferência por nus ao ar livre.
- **Edgar Degas**: pinturas de interiores e luz artificial; captação de um instante da vida das pessoas; paixão por teatro de bailados.
- **Jacob Abraham Camille Pissarro**: uso de paleta de cores quentes; firmeza na captação da atmosfera.
- **Alfred Sisley**: captações sutis das diversas matizes de luz; homogeneização da água, terra e céu.

Outra parte importante é a definição do ambiente, e para este caso temos um ambiente:

- **Totalmente observável**, já que os sensores do agente terão acesso a todas as informações sobre o estado da obra.
- **Estocástico**, uma vez que o ambiente passará por mudanças a cada nova pintura.
- **Episódico**, pois a tarefa de previsão irá se repetir de forma independente entre as pinturas.
- **Dinâmico**, pois há mudança de ambiente.
- **Discreto**, conjunto finito de estados.
- **Único agente**, que terá os sensores para aquisição de dados das obras.

Uma vez que estamos a par destas informações, podemos pensar na abordagem que poderá ser utilizada na identificação dos autores.

Pensando num modelo de classificação, precisaremos fornecer imagens (através de um dataset pré-existente) das pinturas com seus respectivos rótulos para que a máquina possa aprender. Visando fornecer uma quantidade de dados significativa e representativa para treino, podemos usar um conjunto de dados compostos por três versões de cada uma das imagens, uma redimensionada, uma rotacionada e uma cortada.

Pode ser interessante que a IA tenha alguma habilidade de reconhecimento de objetos em cena, pois pode ser de grande ajuda entender quais as particularidades que estão pintadas e, assim, definir seu criador.

Uma vez que tenhamos um sistema com um modelo treinado, podemos usar um sensor ótico que poderá realizar novas capturas e analisar a captação buscando identificar e rotular o autor da pintura que foi captada.

Questão 2

Encontre no Kaggle uma base de dados de 2019 ou 2020 que você resolva usando métodos de regressão, outra base do mesmo ano que resolve utilizando métodos de classificação e uma terceira base, também de 2019 ou 2020, onde você possa aplicar ambos os métodos. Explique quem são as entradas, quem é a saída e a motivação para previsão da saída!

Respostas

Regressão

Como já sabemos, a inteligência artificial pode ser de grande valia para as mais diversas áreas e o sistema financeiro não seria uma opção a ficar de fora. É fato que em todos os lugares, pelas mais diferentes razões, há pessoas que estão (ou estarão, em algum momento) inadimplentes com algum sistema financeiro.

Com a aplicação do **método de regressão**, pode-se ter a informação acerca dos clientes que não estão em dia com a instituição. A inteligência artificial pode ajudar, por exemplo, que as empresas descubram a porcentagem de pessoas em dívida com a instituição, podendo assim, pensar estratégias de como gerir a condição destes clientes. Pode averiguar também a relação Idade x Rendimento anual, de modo a observar algumas taxas relacionadas. Não só isto, dependendo da implementação, pode-se ainda estudar fatores que levam a inadimplência, faixas de idade onde o evento tem maior recorrência, entre outros.

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL	OCCUPATION_TYPE	CNT_FAM_MEMBERS
0	5008804	M	Y	Y	0	427500.0	Working	Higher education	Civil marriage	Rented apartment	-12005	-4542	1	1	0	0	NaN	2.0
1	5008805	M	Y	Y	0	427500.0	Working	Higher education	Civil marriage	Rented apartment	-12005	-4542	1	1	0	0	NaN	2.0
2	5008806	M	Y	Y	0	112500.0	Working	Secondary / secondary special	Married	House / apartment	-21474	-1134	1	0	0	0	Security staff	2.0
3	5008808	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1.0
4	5008809	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1.0
5	5008810	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1.0
6	5008811	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1.0
7	5008812	F	N	Y	0	283500.0	Pensioner	Higher education	Separated	House / apartment	-22464	365243	1	0	0	0	NaN	1.0
8	5008813	F	N	Y	0	283500.0	Pensioner	Higher education	Separated	House / apartment	-22464	365243	1	0	0	0	NaN	1.0
9	5008814	F	N	Y	0	283500.0	Pensioner	Higher education	Separated	House / apartment	-22464	365243	1	0	0	0	NaN	1.0
10	5008815	M	Y	Y	0	270000.0	Working	Higher education	Married	House / apartment	-16872	-769	1	1	1	1	Accountants	2.0
11	5112956	M	Y	Y	0	270000.0	Working	Higher education	Married	House / apartment	-16872	-769	1	1	1	1	Accountants	2.0
12	6153651	M	Y	Y	0	270000.0	Working	Higher education	Married	House / apartment	-16872	-769	1	1	1	1	Accountants	2.0
13	5008819	M	Y	Y	0	135000.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2.0
14	5008820	M	Y	Y	0	135000.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2.0
15	5008821	M	Y	Y	0	135000.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2.0
16	5008822	M	Y	Y	0	135000.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2.0
17	5008823	M	Y	Y	0	135000.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2.0
18	5008824	M	Y	Y	0	135000.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2.0
19	5008825	F	Y	N	0	130500.0	Working	Incomplete higher	Married	House / apartment	-10669	-1103	1	0	0	0	Accountants	2.0

Classificação

A aplicação da inteligência artificial ajudar empresas a expandir seus negócios, oferecendo novos serviços e produtos. Podemos citar como exemplo, uma determinada empresa de seguros que está com intenção de lançar um novo serviço e tem interesse em saber quais dos seus atuais clientes teriam interesse em adquirir tal serviço.
Havendo informações suficientes para que a máquina possa aprender, isto é, **existindo os parâmetros relevantes sobre os clientes e as respostas dos mesmos em relação à adesão ao novo serviço**, podemos separar uma parte deste dataset, para treinar a máquina de modo que possa prever a intenção dos clientes, ou seja, se o cliente estaria ou não interessado.

```
1 df_train.head()
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	1
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	0
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0	26.0	27	1
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0	152.0	203	0
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0	152.0	39	0

```
1 df_test.head()
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage
0	381110	Male	25	1	11.0	1	< 1 Year	No	35786.0	152.0	53
1	381111	Male	40	1	28.0	0	1-2 Year	Yes	33762.0	7.0	111
2	381112	Male	47	1	28.0	0	1-2 Year	Yes	40050.0	124.0	199
3	381113	Male	24	1	27.0	1	< 1 Year	Yes	37356.0	152.0	187
4	381114	Male	27	1	28.0	1	< 1 Year	No	59097.0	152.0	297

```
1 df_sample.head()
```

	id	Response
0	381110	0
1	381111	0
2	381112	0
3	381113	0
4	381114	0

Classificação / Regressão

Muitas vezes podemos ter bases de dados que nos permite analisar as informações nele contidas de diversas maneiras, o que possibilidade uma maior flexibilidade na hora de extrair material relevante dessas bases.
A exemplo, podemos pegar uma base contendo informações de pacientes sobre risco cardíaco. Com dados relevantes, podemos treinar a máquina para predizer se o paciente corre risco ou não de ter ataque cardíaco.
Também podemos a taxa de pessoas que reúnem uma determinada quantidade de características e possuem risco de ataque cardíaco, fazendo assim, que possamos ter uma maior clareza sobre quais características podem estar relacionadas ao risco cardíaco.

```
1 df_train.head()
```

	Gender	Chain_smoker	Consumes_other_tobacco_products	HighBP	Obese	Diabetes	Metabolic_syndrome	Use_of_stimulant_drugs	Family_history	History_of_preeclampsia	CABG_history	Respiratory_illness
0	1	0	1	1	1	0	0	0	1	0	0	0
1	1	0	1	0	1	0	0	0	1	0	0	0
2	1	1	1	0	1	0	0	0	1	0	0	0
3	2	0	0	0	1	0	0	0	1	0	0	0
4	1	0	1	0	0	0	0	1	1	0	0	0

```
1 df_test.head()
```

	Gender	Chain_smoker	Consumes_other_tobacco_products	HighBP	Obese	Diabetes	Metabolic_syndrome	Use_of_stimulant_drugs	Family_history	History_of_preeclampsia	CABG_history	Respiratory_illness	UnderRisk
0	1	1	1	0	1	0	0	0	1	0	0	0	no
1	1	0	1	0	1	0	0	0	1	0	0	0	no
2	1	0	1	0	1	0	0	0	1	0	0	0	no
3	1	0	1	0	1	0	0	0	1	0	0	0	no
4	1	0	0	0	0	0	1	1	0	0	0	0	no

```
1 df_sample.head()
```

	no	yes
0	0.471908	0.528092
1	0.720848	0.279152
2	0.933400	0.066600
3	0.859708	0.140292
4	0.712411	0.287589

```
1 df_train.describe()
```

	Gender	Chain_smoker	Consumes_other_tobacco_products	HighBP	Obese	Diabetes	Metabolic_syndrome	Use_of_stimulant_drugs	Family_history	History_of_preeclampsia	CABG_history	Respiratory_illness
count	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000	382.000000
mean	1.295812	0.120419	0.793194	0.054974	0.892670	0.052356	0.060209	0.107330	0.924084	0.005236	0.020942	0.031414
std	0.511221	0.325878	0.405547	0.228228	0.309938	0.223036	0.238186	0.309938	0.265211	0.072262	0.143379	0.174662
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
50%	1.000000	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
75%	2.000000	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
max	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Questão 3

Encontre no Kaggle uma base de dados de 2019 ou 2020 que você resolva usando métodos supervisionados, outra base do mesmo ano que resolve utilizando métodos não supervisionados e uma terceira base, também de 2019 ou 2020, onde você possa aplicar ambos os métodos. Explique que são as entradas, quem é a saída e a motivação para previsão da saída!

Resposta

Métodos supervisionados

Podemos usar a aprendizagem de máquina para prever doenças e condições de forma assertiva e automática, ajudando assim, que medidas sejam tomadas quanto antes. Tal como dados referentes a uma pesquisa com adultos para detectar a probabilidade de autismo, permitindo que os profissionais da saúde possar direcionar seus esforços.

```
1 df_autism.head(15)
```

	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	age	gender	ethnicity	jundice	austim	contry_of_res	used_app_before	result	age_desc	relation	Class/ASD
0	1	1	1	1	0	0	1	1	0	0	26.0	f	White-European	no	no	United States	no	6.0	18 and more	Self	NO
1	1	1	0	1	0	0	0	1	0	1	24.0	m	Latino	no	yes	Brazil	no	5.0	18 and more	Self	NO
2	1	1	0	1	1	0	1	1	1	1	27.0	m	Latino	yes	yes	Spain	no	8.0	18 and more	Parent	YES
3	1	1	0	1	0	0	1	1	0	1	35.0	f	White-European	no	yes	United States	no	6.0	18 and more	Self	NO
4	1	0	0	0	0	0	0	1	0	0	40.0	f	?	no	no	Egypt	no	2.0	18 and more	?	NO
5	1	1	1	1	1	0	1	1	1	1	36.0	m	Others	yes	no	United States	no	9.0	18 and more	Self	YES
6	0	1	0	0	0	0	0	1	0	0	17.0	f	Black	no	no	United States	no	2.0	18 and more	Self	NO
7	1	1	1	1	0	0	0	0	1	0	64.0	m	White-European	no	no	New Zealand	no	5.0	18 and more	Parent	NO
8	1	1	0	0	1	0	0	1	1	1	29.0	m	White-European	no	no	United States	no	6.0	18 and more	Self	NO
9	1	1	1	1	0	1	1	1	1	0	17.0	m	Asian	yes	yes	Bahamas	no	8.0	18 and more	Health care professional	YES
10	1	1	1	1	1	1	1	1	1	1	33.0	m	White-European	no	no	United States	no	10.0	18 and more	Relative	YES
11	0	1	0	1	1	1	1	0	0	1	18.0	f	Middle Eastern	no	no	Burundi	no	6.0	18 and more	Parent	NO
12	0	1	1	1	1	1	0	0	1	0	17.0	f	?	no	no	Bahamas	no	6.0	18 and more	?	NO
13	1	0	0	0	0	0	1	1	0	1	17.0	m	?	no	no	Austria	no	4.0	18 and more	?	NO
14	1	0	0	0	0	0	1	1	0	1	17.0	f	?	no	no	Argentina	no	4.0	18 and more	?	NO

Métodos não supervisionados

Havendo uma base de dados que contenham dados de reclamações de clientes, pode-se entender melhor quais categorias de reclamações são mais recorrentes.

	Date received	Product	Sub-product	Issue	Sub-issue	Consumer complaint narrative	Company public response	Company	State	ZIP code	Tags	Consumer consent provided?	Submitted via	Date sent to company	Company response to consumer	Timely response?	Consumer disputed?	Complaint ID
0	07/29/2013	Consumer Loan	Vehicle loan	Managing the loan or lease	NaN	NaN	NaN	Wells Fargo & Company	VA	24540	NaN	NaN	Phone	07/30/2013	Closed with explanation	Yes	No	468882
1	07/29/2013	Bank account or service	Checking account	Using a debit or ATM card	NaN	NaN	NaN	Wells Fargo & Company	CA	95992	Older American	NaN	Web	07/31/2013	Closed with explanation	Yes	No	468889
2	07/29/2013	Bank account or service	Checking account	Account opening, closing, or management	NaN	NaN	NaN	Santander Bank US	NY	10065	NaN	NaN	Fax	07/31/2013	Closed	Yes	No	468879
3	07/29/2013	Bank account or service	Checking account	Deposits and withdrawals	NaN	NaN	NaN	Wells Fargo & Company	GA	30084	NaN	NaN	Web	07/30/2013	Closed with explanation	Yes	No	468949
4	07/29/2013	Mortgage	Conventional fixed mortgage	Loan servicing, payments, escrow account	NaN	NaN	NaN	Franklin Credit Management	CT	06106	NaN	NaN	Web	07/30/2013	Closed with explanation	Yes	No	475823
5	07/29/2013	Bank account or service	Checking account	Deposits and withdrawals	NaN	NaN	NaN	Bank of America	TX	75025	NaN	NaN	Web	07/30/2013	Closed with explanation	Yes	No	468981
6	07/29/2013	Debt collection	Other (i.e. phone, health club, etc.)	Cont'd attempts collect debt not owed	Debt is not mine	NaN	NaN	NRA Group, LLC	VA	20147	NaN	NaN	Web	08/07/2013	Closed with non-monetary relief	Yes	No	467801
7	07/29/2013	Debt collection	I do not know	Cont'd attempts collect debt not owed	Debt was paid	NaN	NaN	SunTrust Banks, Inc.	FL	32818	NaN	NaN	Referral	08/01/2013	Closed with explanation	Yes	Yes	475728
8	07/29/2013	Credit card	NaN	Billing statement	NaN	NaN	NaN	Citibank	OH	45247	NaN	NaN	Referral	07/30/2013	Closed with explanation	Yes	Yes	469026
9	07/29/2013	Mortgage	Other mortgage	Loan servicing, payments, escrow account	NaN	NaN	NaN	Wells Fargo & Company	NV	89511	NaN	NaN	Referral	07/30/2013	Closed with explanation	Yes	Yes	469035
10	07/29/2013	Mortgage	Other mortgage	Loan modification,collection,foreclosure	NaN	NaN	NaN	Bank of America	NC	27949	NaN	NaN	Referral	07/30/2013	Closed with non-monetary relief	Yes	No	469037

Podemos usar métodos não supervisionados de modo a normalizar nomes de empresas e agrupa-las. Por exemplo, ao analisar dos dados, uma empresa que esteja com o nome de CAP ONE, deverá, no fim, estar agrupada com as empresas que tenham o nome CAPITAL ONE. Outra opção é o agrupamento das reclamações que tenham características consideradas importantes. Isto pode permitir um novo olhar sobre o estado geral das reclamações.

Métodos supervisionados / não supervisionados

Dados médicos são uma inestimável fonte de informação para os mais diversos estudos e não poderia diferir com o uso da inteligência artificial.
Com os dados sobre câncer de mama colhidos através dos anos, podemos usar métodos não supervisionados para agrupar pacientes que possuem um conjunto de características importantes de modo a descobrir alguma particularidade.

```
1 df_breast_cancer_pred.head(15)
```

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
0	1000025	5	1	1	1	2	1	3	1	1	2
1	1002945	5	4	4	5	7	10	3	2	1	2
2	1015425	3	1	1	1	2	2	3	1	1	2
3	1016277	6	8	8	1	3	4	3	7	1	2
4	1017023	4	1	1	3	2	1	3	1	1	2
5	1017122	8	10	10	8	7	10	9	7	1	4
6	1018099	1	1	1	1	2	10	3	1	1	2
7	1018561	2	1	2	1	2	1	3	1	1	2
8	1033078	2	1	1	1	2	1	1	1	5	2
9	1033078	4	2	1	1	2	1	2	1	1	2
10	1035283	1	1	1	1	1	1	3	1	1	2
11	1036172	2	1	1	1	2	1	2	1	1	2
12	1041801	5	3	3	3	2	3	4	4	1	4
13	1043999	1	1	1	1	2	3	3	1	1	2
14	1044572	8	7	5	10	7	9	5	5	4	4

Com os dados acima, podemos usar a clusterização para saber quais perfis em quais condições possuem tumores malignos. Podemos, também, treinar a máquina para predizer informações sobre tumores com base nos dados já existentes. Isto é possível, pois a base de dados possui as informações de saída, que na imagem acima é a coluna “Class”.

Questão 4

Em um Jupyter Notebook (link do colab ou arquivo), utilize o método K vizinhos mais próximos no problema supervisionado escolhido na Questão 2 ou 3 e:

a. *Aplique e explique os processos de pré-processamento necessários para execução do método K-NN.*
b. *Escolha 5 valores de K e mostre os resultados (escolha uma métrica)?*
c. *Escolha 2 métricas de similaridade (diferentes distâncias ou outras métricas) para definir o(s) vizinho(s) mais próximos.*

Respostas

a. https://colab.research.google.com/drive/1mf66O4I_C9VfVEMRTmyC3AY5CcuUa6CF?usp=sharing

   https://github.com/marciliojrr/IA/blob/main/Prova_IA_Quest%C3%A3o_4.ipynb

b. As imagens abaixo foram capturadas com valores de k igual a 1, 3, 5, 7 e 9 respectivamente. Todas utilizaram a métrica “minkowski”.

```
 1 from sklearn.metrics import classification_report, confusion_matrix
 2 print(confusion_matrix(y_test, y_pred))
 3 print(classification_report(y_test, y_pred))
```

```
[[92  3]
 [ 3 39]]
              precision    recall  f1-score   support

           2       0.97      0.97      0.97        95
           4       0.93      0.93      0.93        42

    accuracy                           0.96       137
   macro avg       0.95      0.95      0.95       137
weighted avg       0.96      0.96      0.96       137
```

```
 1 from sklearn.metrics import classification_report, confusion_matrix
 2 print(confusion_matrix(y_test, y_pred))
 3 print(classification_report(y_test, y_pred))
```

```
[[91  4]
 [ 2 40]]
              precision    recall  f1-score   support

           2       0.98      0.96      0.97        95
           4       0.91      0.95      0.93        42

    accuracy                           0.96       137
   macro avg       0.94      0.96      0.95       137
weighted avg       0.96      0.96      0.96       137
```

```
 1 from sklearn.metrics import classification_report, confusion_matrix
 2 print(confusion_matrix(y_test, y_pred))
 3 print(classification_report(y_test, y_pred))
```

```
[[86  3]
 [ 1 47]]
              precision    recall  f1-score   support

           2       0.99      0.97      0.98        89
           4       0.94      0.98      0.96        48

    accuracy                           0.97       137
   macro avg       0.96      0.97      0.97       137
weighted avg       0.97      0.97      0.97       137
```

```
1 from sklearn.metrics import classification_report, confusion_matrix
2 print(confusion_matrix(y_test, y_pred))
3 print(classification_report(y_test, y_pred))
```

```
[[83  2]
 [ 0 52]]
              precision    recall  f1-score   support

           2       1.00      0.98      0.99        85
           4       0.96      1.00      0.98        52

    accuracy                           0.99       137
   macro avg       0.98      0.99      0.98       137
weighted avg       0.99      0.99      0.99       137
```

```
1 from sklearn.metrics import classification_report, confusion_matrix
2 print(confusion_matrix(y_test, y_pred))
3 print(classification_report(y_test, y_pred))
```

```
[[87  2]
 [ 3 45]]
              precision    recall  f1-score   support

           2       0.97      0.98      0.97        89
           4       0.96      0.94      0.95        48

    accuracy                           0.96       137
   macro avg       0.96      0.96      0.96       137
weighted avg       0.96      0.96      0.96       137
```

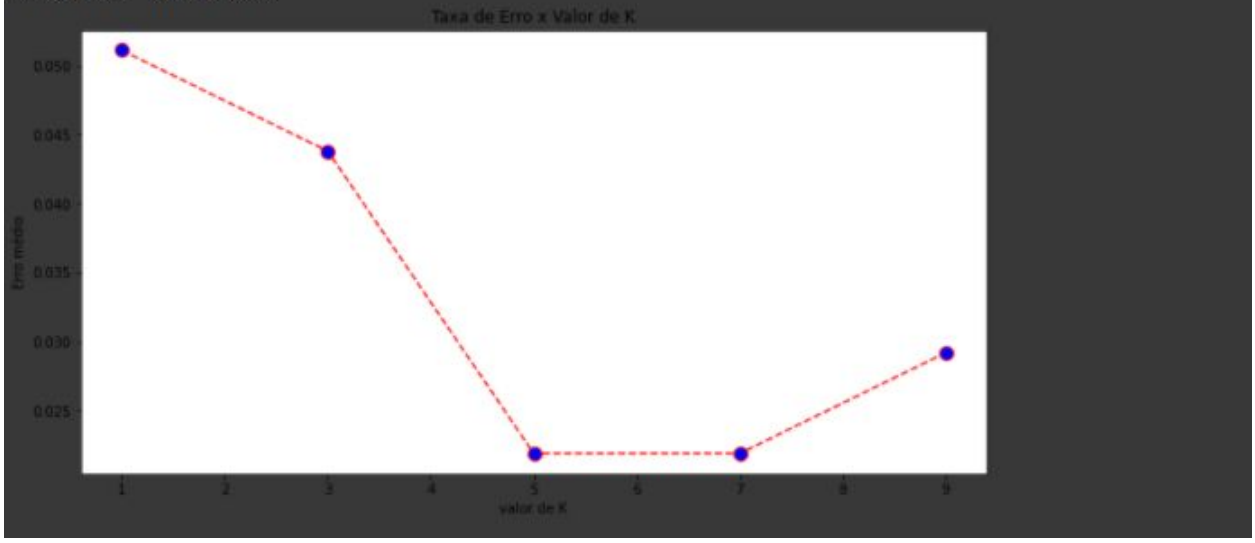
c. Utilizando a distância “euclidean”, podemos analisar, através do gráfico do erro médio abaixo, 5 valores de k e a taxa de erro para cada um deles. O que permite uma melhor escolha dos valores de k.

```
# Calculando erro médio dos valores previstos do conjunto de teste
error = []

for i in range(1, 10, 2):
  knn = KNeighborsClassifier(n_neighbors=i)
  knn.fit(X_train, y_train)
  pred_i = knn.predict(X_test)
  error.append(np.mean(pred_i != y_test))

# Plotando o gráfico dos valores em *error* em relação aos valores de K
plt.figure(figsize=(12,6))
plt.plot(range(1, 10, 2), error, color='red', linestyle='dashed', marker='o', markerfacecolor='blue', markersize=10)
plt.title('Taxa de Erro x Valor de K')
plt.xlabel('valor de K')
plt.ylabel('Erro médio')
```

```
Text(0, 0.5, 'Erro médio')
```



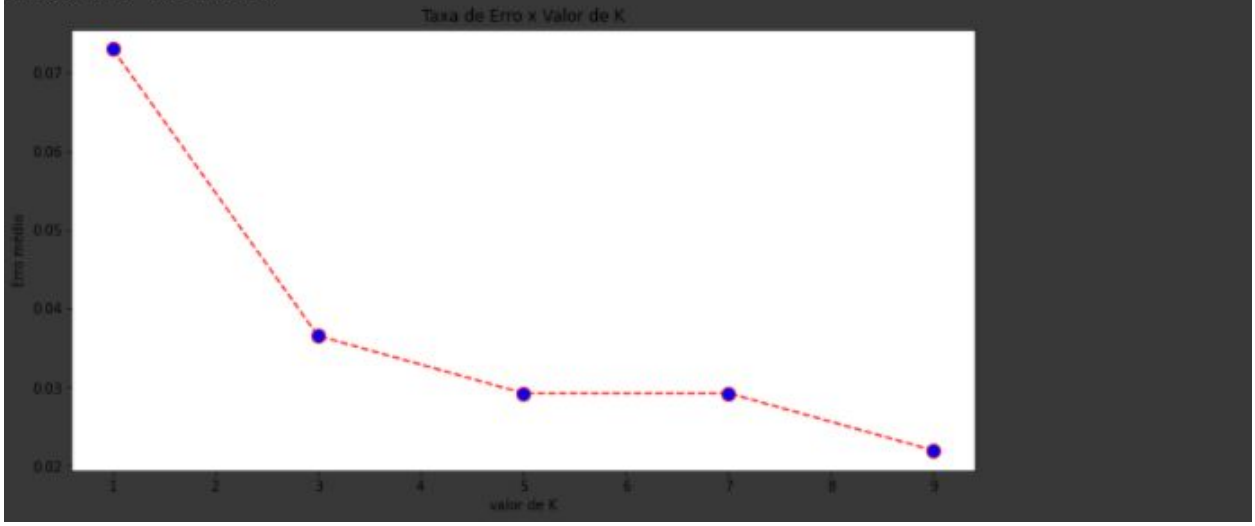
Abaixo temos o gráfico para taxa de erros por valores de k, utilizando a distância *"manhattan"*.

```python
# Calculando erro médio dos valores previstos do conjunto de teste
error = []

for i in range(1, 10, 2):
  knn = KNeighborsClassifier(n_neighbors=i)
  knn.fit(X_train, y_train)
  pred_i = knn.predict(X_test)
  error.append(np.mean(pred_i != y_test))

# Plotando o gráfico dos valores em *error* em relação aos valores de K
plt.figure(figsize=(12,6))
plt.plot(range(1, 10, 2), error, color='red', linestyle='dashed', marker='o', markerfacecolor='blue', markersize=10)
plt.title('Taxa de Erro x Valor de K')
plt.xlabel('valor de K')
plt.ylabel('Erro médio')
```

```
Text(0, 0.5, 'Erro médio')
```



Questão 5

Em quais das bases selecionadas por vocês nas questões 2 e 3 se faz necessário o uso de redução ou agregação de atributos? Mostre nelas onde e qual dos métodos e justifique.

Resposta

Na base de dados ***Credit Card Approval Prediction****,* deve-se fazer um ajuste dos dados. A base possui valores inválidos, tipos de dados diferentes, intervalos entre valores amplos. Também há atributos que podem ser removidos, como a primeira que é referente apenas a identificação do usuário.

Para a base ***Health Insurance Cross Sell Prediction***, poderia ser feitas a normalização dos valores e a exclusão do atributo de identificação, que assim como nas outras bases, não são necessárias para o uso dos métodos de aprendizado.

Já na base ***Cardiac Risk Prediction*** a base está bem estabelecida para que se apliquem os métodos de aprendizagem.

No uso do dataset ***Autism Screening on Adults*** algumas informações estão desfalcadas, fazendo necessário, para uma melhor qualidade do resultado, que sejam removidas juntamente de alguns outros atributos que não são relevantes para o aprendizado.

Consumer Complaints é um dataset com bastante falta de dados em alguns atributos, alguns deles podem ser necessários para uma melhor assertividade do modelo que se implemente. Com isso, podem ser necessárias o preenchimento de algumas informações com base em outras já existentes ou a exclusão destas.

No dataset ***Breast Cancer Prediction****,* foi realizado a remoção da primeira coluna que continha informações sobre identificação do paciente. Uma vez que esta não seria um atributo relevante para o aprendizado, foi utilizado a função *drop()* do Pandas. A questão 4 possui o uso desta função neste dataset.

Questão 6

Descreva 5 problemas comumente encontrados em bases de dados e como resolvê-los, pensando na aplicação de métodos de inteligência artificial.

Resposta

1. **Valores desconhecidos**: este problema pode ser contornado com a substituição destes valores pela média ou moda do atributo.
2. **Valores extremos**: pode-se excluir os outliers, quando este for fruto de um erro de entrada; analisar separadamente os outliers; usar clusterização para encontrar uma aproximação.
3. **Classes desbalanceadas**: para esta categoria de problema, pode-se procurar por uma distribuição da classe que forneça um desempenho aceitável de classificação para a classe minoritária.
4. **Seleção de atributos**: selecionar um subconjunto de atributos e medir a precisão do classificador neste subconjunto.
5. **Construção de atributos**: geração de novos atributos a partir de outros já existentes, porém, pouco relevantes individualmente.

Questão 7

Elabore uma pergunta e responda sobre "Preconceito e Inteligência Artificial", refletindo aspectos importantes do método que um engenheiro de dados/cientista de dados deve conhecer e se preocupar garantindo ética nos modelos criados.

Resposta

Pergunta

Diante do crescente avanço da tecnologia, temos cada vez mais contato com a inteligência artificial no cotidiano. Não é surpresa que estamos delegando diversas funções para sistemas inteligentes que antes era algo impossível de ser feito por estes. Infelizmente, há pontos negativos que precisam de um cuidado constante dos criadores de tais sistemas, isso porque, as máquinas que desempenham tais funções precisam aprender sobre assunto de interesse onde irão atuar e elas não aprendem sozinhas, mas sim com informações providas por seus criadores. Preconceito e inteligência artificial são assuntos que já são debatidos há algum tempo e já temos alguns casos onde se acusa que uma inteligência artificial esta agindo com preconceito.
Uma vez que o aprendizado da IA depende do fornecimento de informações por parte dos cientistas/engenheiros de dados, como poderíamos evitar que a máquina não agisse com preconceito?

Resposta

Casos de preconceito por parte de inteligências artificiais são, basicamente, culpa dos dados que lhes são fornecidos para aprendizagem. Em um primeiro momento, podemos pensar que um problema seja uma quantidade insuficiente de dados para treino, entretanto, pode ser que o problema não seja a quantidade, mas a qualidade. Os dados fornecidos para a máquina podem estar numa quantidade muito grande, mas que não contemple todo o assunto a ser estudado pela IA, ou seja, os dados, ainda que em grandes quantidades, não representam bem o objeto de estudo.
É importante ressaltar, para um maior entendimento, como acontece o preconceito na IA. Este acontece quando estereótipos sociais influenciam fortemente os dados de treinamento da máquina, e isto acontece pelo que foi citado acima.
É importante que os responsáveis pelo sistema, observem os resultados obtidos com os dados que foram fornecidos, pois, mesmo sem a intenção, vieses podem estar sendo passados para a máquina. É preciso entender que um resultado ruim, neste caso, preconceituoso, não é culpa do algoritmo e sim dos dados que foram passados para aprendizado da mesma, e por consequência, das pessoas que forneceram estes dados.
Quando se tem ciência de que preconceito está sendo tratado, a solução deste problema pode estar mais perto de ser encontrada.
O ideal é sempre criar os sistemas sem nenhum viés, entretanto, esta é uma tarefa de extrema complexidade para os cientistas/engenheiro de dados, pois há diversas variáveis e a abertura do preconceito pode não ser óbvia.

O trabalho multidisciplinar é de grande importância nesses casos, pois permite que o treinamento seja visto por diversos pontos de vista, e não só do especialista em tecnologia. Como mencionado, é preciso avaliar os resultados obtidos do treinamento e uma equipe multidisciplinar poderá enriquecer o debate acerca destes resultados.