CLASSIFICAÇÃO BAYESIANA PARA PREDIÇÃO DE CÂNCER DE FÍGADO

Anderson Luis Marchi
Instituto Federal Catarinense

andersonlmarchi@gmail.com

1. Introdução

O presente estudo utiliza o dataset *Predict Liver Cancer from & Clinical Features*, disponibilizado originalmente no Kaggle (2025). Este é um conjunto de dados de classificação de câncer de fígado sintético, porém clinicamente realista, contendo 5.000 registros de pacientes e 14 colunas. Os dados simulam atributos reais de pacientes que influenciam o risco de câncer de fígado, combinando características demográficas, de estilo de vida, clínicas e baseadas em biomarcadores.

Esse dataset pode ser utilizado em pesquisas de aprendizado de máquina, pois combina variáveis demográficas, comportamentais e de percepção subjetiva, o que permite a construção de modelos preditivos voltados à análise do risco de câncer no fígado em contextos clínicos.

2. Desenvolvimento da solução

O dataset contém 5000 registros de indivíduos e um conjunto de variáveis categóricas e numéricas. Entre os principais grupos de atributos, destacam-se:

- **Variáveis demográficas:** idade, gênero.
- **Variáveis de estilo de vida:** índice de massa corporal (BMI), consumo de álcool (nunca, ocasional, regular), status de tabagismo (nunca, ex-fumante, atual) e nível de atividade física (baixo, moderado, alto).
- **Variáveis clínicas:** presença de hepatite B, presença de hepatite C, histórico de cirrose hepática, presença de diabetes e histórico familiar de câncer.
- **Variáveis laboratoriais:** escore de função hepática (0–100) e nível de alfa-fetoproteína (AFP) em ng/mL.
- **Variável alvo:** diagnóstico final de câncer de fígado (0 = não possui câncer, 1 = possui câncer).

De modo geral, o dataset possui um caráter misto, conforme visualizado na Figura 1 a seguir, com dados qualitativos (nominais e ordinais) e quantitativos sendo um modelo perfeito para ser usado em MLP.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	gender	bmi	alcohol_consumption	smoking_status	hepatitis_b	hepatitis_c	liver_function_score	alpha_fetoprotein_level	cirrhosis_history	family_history_cancer	physical_activity_level	diabetes	liver_cancer
2	68	Female	18.1	Regular	Former	0	0	51.9	16.44	0	0	Low	0	0
3	81	Female	19.9	Occasional	Never	0	0	41.6	8.09	0	0	Moderate	1	0
4	58	Female	25.5	Never	Never	0	0		760.64	0	0	Moderate	0	0
5	44	Male	16	Never	Former	0	0	50.3	19.09	0	0	Low	1	0
6	72	Male	21	Occasional	Former	0	0	39.5	4.95	1	0	Low	1	1
7	37	Female	23.1	Regular	Never	0	0	50.8	0.75	0	0	Moderate	0	0
8	50	Male	19.4	Regular	Current	0	0	68.3	0.31	1	0	Moderate	0	0
9	68	Male	15.4	Regular	Former	0	0	70.8	40.18	0	0	High	0	0
10	48	Male	27.4	Occasional	Former	0	0	70.2	4.36	0	0	Low	0	0
11	52	Male	26.2	Occasional	Never	1	0	48.8	8.91	0	1	High	0	0
12	40	Male	29.9	Never	Current	0	0	77.7	45.11	0	0	High	0	0
13	40	Female	20.4	Occasional	Never	0	0	75.3	9.45	0	0	Low	0	0
14	53	Female	29.8	Never	Former	0	0	71.5	1.23	0	0	Moderate	0	0
15	82	Female	33.6	Occasional	Current	0	0	48.4	12.04	1	0	Moderate	0	1
16	65	Female	27.3	Occasional	Never	1	0	55.2	1.38	0	0	Moderate	1	0
17	69	Male	23	Never	Former	1	0	69.2	1.5	0	1	Moderate	0	0
18	53	Male	26.7	Never	Never	0	0	73.2	28.61	0	0	Moderate	0	0
19	32	Female	22.7	Never	Former	0	0	49.3	5.13	1	0	Low	0	0
20	51	Male	21.9	Never	Former	0	0	90.8	17.03	0	0	Low	1	0
21	82	Male	29.4	Never	Current	0	0	62.5	10.67	0	0	Low	1	0
22	31	Female	25.3	Never	Current	1	0	78.8	3.77	1	0	Moderate	0	0
23	53	Male	32	Never	Former	0	1	45.9	10.94	1	1	Low	0	1
24	73	Female	23.8	Never	Former	0	0	67.6	1.07	0	0	High	0	0
25	59	Female	24.3	Regular	Former	0	0	45.6	6.4	0	0	Moderate	0	0
26	67	Male	23.2	Never	Never	0	0	66.8	11.1	0	1	Moderate	0	0
27	31	Male	20.4	Never	Former	1	0	71.5	3.89	1	1	High	0	0
28	50	Female	30.1	Occasional	Current	0	0	72.3	16.19	0	0	Low	0	0
29	62	Female	36.9	Regular	Never	0	0	53.5	0.8	0	0	Low	1	1
30	41	Male	24.1	Occasional	Never	0	0	37.3	9.95	0	0	High	1	0
31	51	Female	22.6	Never	Never	0	0	69.4	18.83	0	1	Moderate	0	0
32	73	Female	22.3	Occasional	Never	1	0	58.6	8.43	0	0	Low	1	0
33	54	Female	22.6	Occasional	Never	1	0	58.6	31.14	0	0	Low	0	0
34	78	Male	34.2	Occasional	Never	1	0	50.7	15	0	0	High	0	0
35	56	Female	22.4	Occasional	Never	0	0	63.7	47.73	1	0	Moderate	1	0
36	71	Male	23.1	Regular	Former	0	0	64.5	5.01	1	0	Low	0	0
37	57	Female	19.8	Regular	Never	0	0	80.4	41.9	1	0	High	1	1
38	45	Male	27.5	Occasional	Former	0	0	34.3	14	1	0	Moderate	0	0

Figura 1: Amostra do dataset aberto em software de planilha

No treinamento de uma rede do tipo Naive-Bayes usamos soluções pré-implementadas em bibliotecas para a linguagem de programação Python e dividimos o processo nos seguintes estágios:

- **Leitura do dataset**: Nele usamos a biblioteca "*pandas"* para leitura e pré-processamento dos dados. Onde removemos a coluna de classificação "liver_cancer" que indica se teve, ou não, o câncer.

```
df = pd.read_csv('liver_cancer_dataset.csv')
X = df.drop('liver_cancer', axis=1)
y = df['liver_cancer']
```

- **Codificação de variáveis categóricas**: Usamos a classe "*LabelEnconde*" da biblioteca "*scikit-learn*" para converter variáveis categóricas (gender, alcohol_consumption, smoking_status, physical_activity_level) em numéricas.

```
label_encoders = {}
for col in X.select_dtypes(include=['object']).columns:
   le = LabelEncoder()
   X[col] = le.fit_transform(X[col])
   label_encoders[col] = le
```

- **Treinamento**: Aqui utilizamos a classe "*GaussianNB*" da biblioteca "*scikit-learn*" para testar diferentes proporções de dados: 0.2, 0.3, 0.4, 0.5.

```
test_sizes = [0.2, 0.3, 0.4, 0.5]
accuracies = []

for test_size in test_sizes:
   X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=test_size, random_state=42)
   model = GaussianNB()
   model.fit(X_train, y_train)
   y_pred = model.predict(X_test)
   acc = accuracy_score(y_test, y_pred)
   accuracies.append(acc)
   print(f'Test size: {test_size}, Accuracy: {acc:.4f}')
```

- **Predição de novos casos**: Incluímos três novos casos com diferentes perfis clínicos e de hábitos para avaliarmos a acurácia do modelo treinado.

```
def encode_case(case_dict):
   case = case_dict.copy()
   for col, le in label_encoders.items():
       case[col] = le.transform([case[col]])[0]
   return case

new_cases = [
   {'age': 60, 'gender': 'Female', 'bmi': 25.0, 'alcohol_consumption':
'Regular', 'smoking_status': 'Never', 'hepatitis_b': 0, 'hepatitis_c':
0, 'liver_function_score': 70.0, 'alpha_fetoprotein_level': 10.0,
'cirrhosis_history': 0, 'family_history_cancer': 0,
'physical_activity_level': 'Moderate', 'diabetes': 0},
   {'age': 45, 'gender': 'Male', 'bmi': 30.0, 'alcohol_consumption':
'Occasional', 'smoking_status': 'Current', 'hepatitis_b': 1,
'hepatitis_c': 0, 'liver_function_score': 55.0,
'alpha_fetoprotein_level': 20.0, 'cirrhosis_history': 1,
'family_history_cancer': 1, 'physical_activity_level': 'Low',
'diabetes': 1},
   {'age': 70, 'gender': 'Female', 'bmi': 22.0, 'alcohol_consumption':
'Never', 'smoking_status': 'Former', 'hepatitis_b': 0, 'hepatitis_c':
1, 'liver_function_score': 80.0, 'alpha_fetoprotein_level': 5.0,
'cirrhosis_history': 0, 'family_history_cancer': 0,
'physical_activity_level': 'High', 'diabetes': 0}
]
new_cases_encoded = pd.DataFrame([encode_case(c) for c in new_cases])
probs = model.predict_proba(new_cases_encoded)
preds = model.predict(new_cases_encoded)
```

4. Resultados

O modelo de classificação foi treinado utilizando quatro proporções de dados para teste. Em cada experimento, avaliamos o desempenho por meio da acurácia mostrada na tabela a seguir.

Proporção	Acurácia
20%	82%
30%	83.4%
40%	84.6%
50%	85.76%

Tabela 1: Acurácia x Proporção de Teste

Através desses valores podemos ver um crescimento de 1.2% na acurácia a cada aumento na proporção de testes tendo um crescimento menor quando alteramos de 40% para 50% do conjunto de dados para teste conforme visto na figura 2 a seguir.

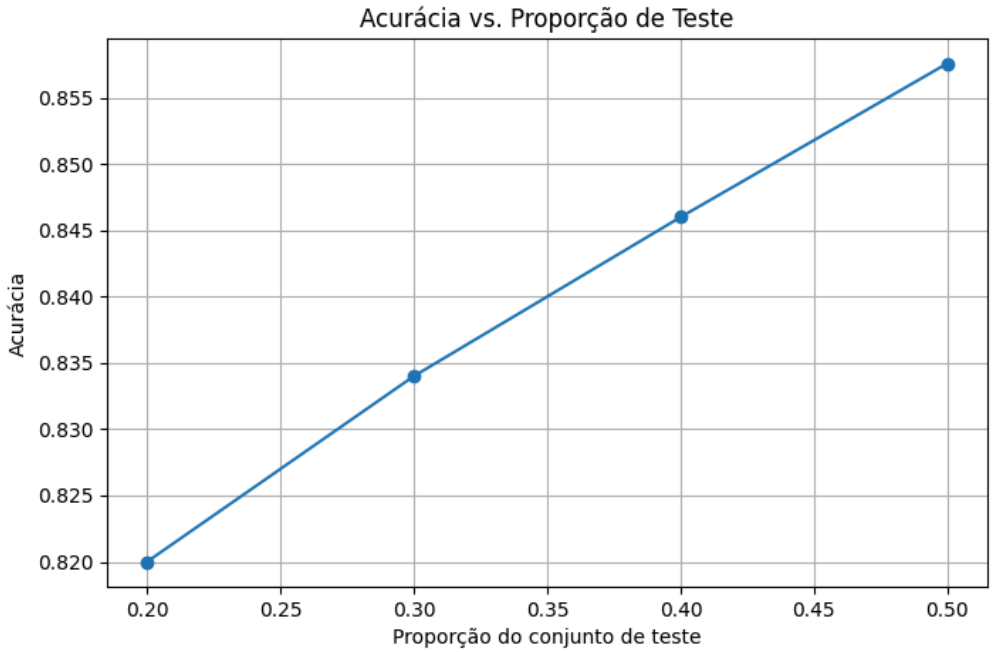

Figura 2: Acurácia X Proporção de Teste

Através desses resultados de acurácia mantemos o treinamento com proporção de 50/50, com maior acurácia verificada, para predição de 3 novos casos conforme tabela a seguir.

Variável	Caso 1	Caso 2	Caso 3
age	60	45	70
gender	Female	Male	Female
bmi	25.0	30.0	22.0
alcohol_consumption	Regular	Occasional	Never
smoking_status	Never	Current	Former
hepatitis_b	0	1	0
hepatitis_c	0	0	1
liver_function_score	70.0	55.0	80.0
alpha_fetoprotein_level	10.0	20.0	5.0
cirrhosis_history	0	1	0
family_history_cancer	0	1	0
physical_activity_level	Moderate	Low	High
diabetes	0	1	0

Tabela 2: Novos casos para testes da rede

E os resultados apontaram que os casos **Caso 1** e **Caso 3** tem Predição de **Não** para câncer de fígado e o **Caso 2** com Predição de **Sim**. No Caso **3** vemos um equilíbrio na predição com 59% para **Não** e 41% para **Sim** como podemos ver no gráfico a seguir.

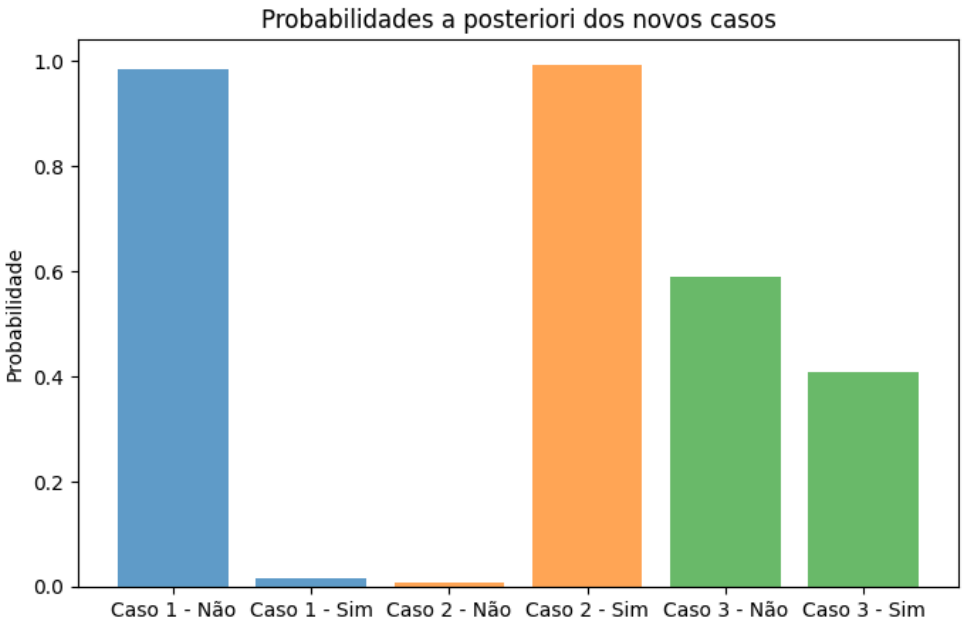

Figura 3: Probabilidades dos novos casos

5. Referências

KAGGLE. **Predict Liver Cancer from & Clinical Features**. Disponível em: https://www.kaggle.com/datasets/miadul/predict-liver-cancer-from-and-clinical-features Acesso em: 03 out. 2025.