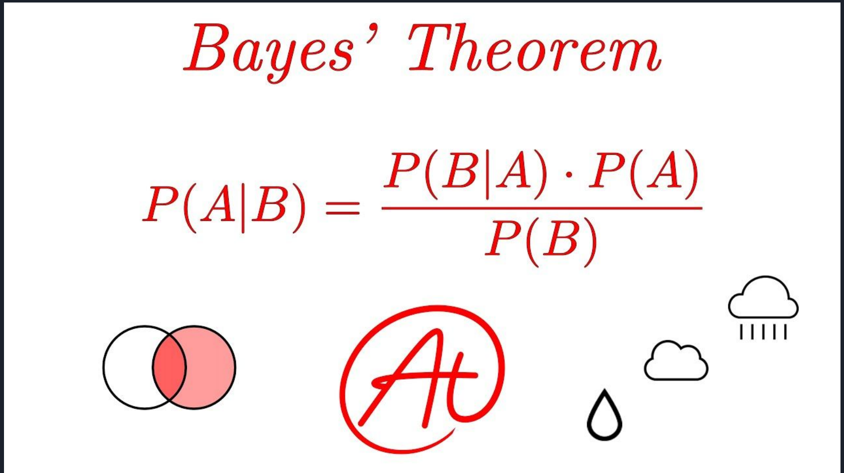
Naive Bayes with Hierarchical MCA

Lakshmi Katrapati, Ipek Sayar

Introduction

- Naive Bayes assumes features are independent of each other
- Want to improve Naive Bayes by making it work for dataset with correlated features
- Goal: Create an algorithm that improves Naive Bayes by accounting for feature dependencies -> allowing it to perform effectively on datasets with correlated features



Related Works

Integrating advanced principal component analysis into naive bayes for enhanced classification performance by Lan Luo, Tianyang Liu

- PCA transforms the data into principal components, reducing the relationship between correlated variables
- This reduction helps meet Naive Bayes' independence assumption, improving classification performance

Dataset

Patient_ID	Age	Sex	Weight_kg	Height_cm	BMI
Patient ID	Age	Sex			
1 – 267k	18 – 60	Female 48% Male 48% Other (10672) 4%	50 – 120	150 – 199	18.5
1	56	Male	118.4	184	35.4
2	46	Male	50.7	164	32.6
3	32	Male	118.0	179	39.8
4	60	Male	96.8	161	23.7
5	25	Male	53.9	169	24.4
6	38	Female	74.3	150	38.3
7	56	Male	63.3	163	33.4

Heart Attack for Individuals in France

- Class Label = Heart Attack
- 26 Features
  - Age
  - Height
  - BMI
  - Blood Pressure Levels
- 266,785 Instances
- Each Instance is a patient

Preprocessing

'Air_Pollution_Level'

0-50	'Good'
51-100	'Moderate'
101-150	'Unhealthy_Sensitive'
151-200	'Unhealthy'

Systolic = Blood_Pressure_Systolic
Diastolic = Blood_Pressure_Diastolic

'BP'

Systolic < 120 & Diastolic < 80	Normal
120 =< Systolic =< 129 & Diastolic < 80	Elevated
130 =< Systolic =< 139 or 80 =< Diastolic =< 89	High_BP_Stage1
Systolic >= 140 or Diastolic >=90	High_BP_Stage2
Systolic > 180 or Diastolic > 120	Hypertensive_Crisis

'Heart_Rate'

50-59	Low Heart Rate (Bradycardia)
60-100	Normal Heart Rate
101-110	Elevated Heart Rate
111-119	High High Rate (Tachycardia)

Preprocessing - Cont.

Patient_ID	Weight_kg
1	118.4
2	50.7
3	118.0
4	96.8
[illegible]	53.9
6	74.3
7	63.3
8	73.5
9	104.3
[illegible]	81.9
11	90.4
12	113.8
13	71.6
14	76.1
15	51.6
16	110.9
17	61.7

‘Physical_Activity_Hours’

0	No Activity
0 - 3.7	Low Activity
3.7 - 7.5	Moderate Activity
7.5 - 11.3	Active
>11.3	Very Active

‘BMI’

0-18.5	Underweight
18.5-25	Healthy_weight
25-30	Overweight
>=30	Obesity

Algorithm

We will improve the Naive Bayes algorithm by addressing its weakness, the assumption of attribute independence, using Multiple Component Analysis.

What is MCA?

MCA is a dimensionality reduction technique specifically designed for categorical variables, similar to how PCA (Principal Component Analysis) works for continuous data.

- Transforms the data into new variables, or "dimensions," that capture the patterns and relationships between the categories of the attributes.
- The output is completely independent.

Chi-Square Test checks if the variables are related, Cramér's V tells you how strong that relationship is.

Chi-Square Test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

- χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.
- O_i = the number of observations of type *i*.
- N = total number of observations
- $E_i = Np_i$ = the expected (theoretical) count of type *i*, asserted by the null hypothesis that the fraction of type *i* in the population is p_i
- n = the number of cells in the table.

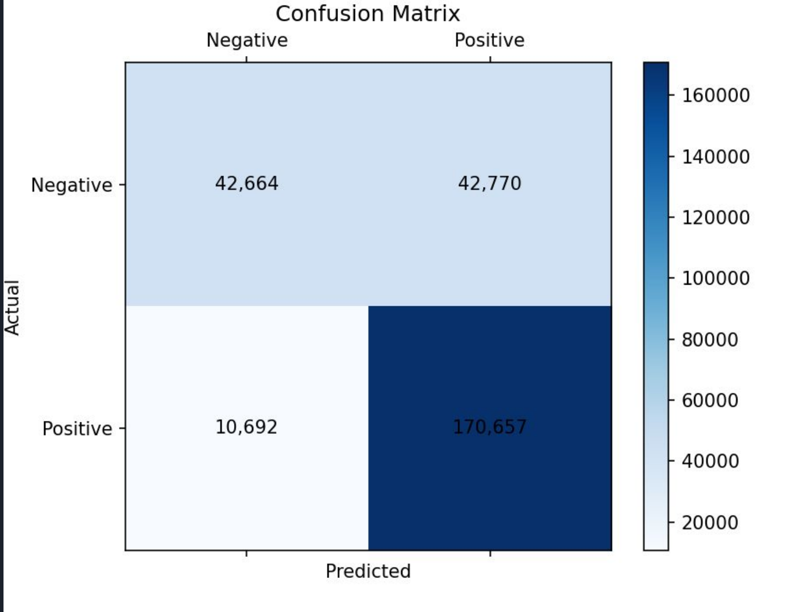
Cramér's V:

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}}$$

- φ is the phi coefficient.
- χ^2 is derived from Pearson's chi-squared test
- n is the grand total of observations and
- k being the number of columns.
- r being the number of rows.

Complement Naive Bayes

Accuracy: 79.96% Recall: 0.9410



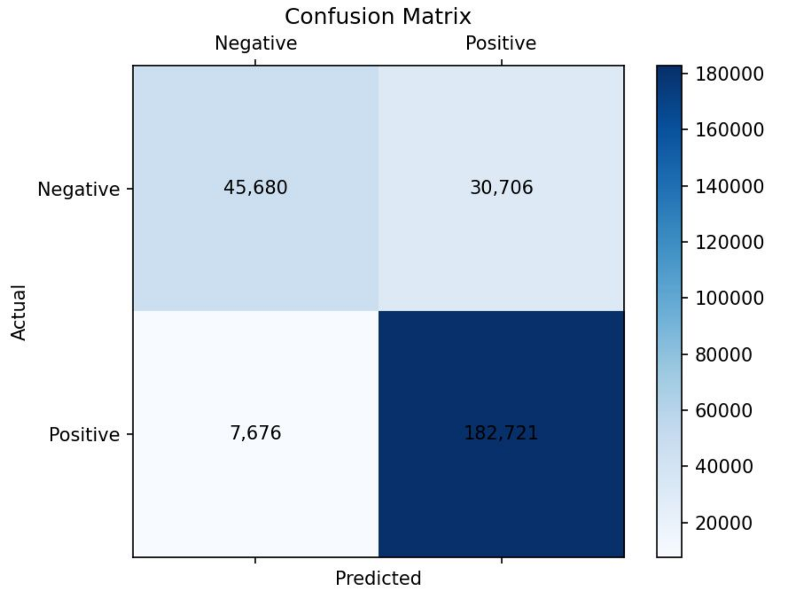
- Designed to do better on imbalance datasets (8-2 split)
  - Focuses on the complement of the minority class
    - focuses on how features <u>contrast</u> with the minority class, rather than just memorizing patterns
  - Strengthens the model's ability to recognize what <u>doesn't belong</u> to the other classes, indirectly improving minority class detection
- Pretty good accuracy on this dataset; attributes are already pretty independent from each other

```
Sex attribute's correlation with the BMI attribute: 0.003318977429497553
Height_cm attribute's correlation with the BMI attribute: 0.0035726780617884365
BP attribute's correlation with the BMI attribute: 0.003308959532608353
Cholesterol_Level attribute's correlation with the BMI attribute: 0.002416541025266376
Smoking_Status attribute's correlation with the BMI attribute: 0.003134053154015418
Alcohol_Consumption attribute's correlation with the BMI attribute: 0.003925400771452204
Physical_Activity_Hours attribute's correlation with the BMI attribute: 0.0035002296561463145
Diabetes attribute's correlation with the BMI attribute: 0.006003349310403892 5
Family_History attribute's correlation with the BMI attribute: 0.002652211677899810 3
Diet_Type attribute's correlation with the BMI attribute: 0.004685597451709236
Stress_Level attribute's correlation with the BMI attribute: 0.002836370011388364
Heart_Rate attribute's correlation with the BMI attribute: 0.002215744126059652
Exercise_Induced_Pain attribute's correlation with the BMI attribute: 0.0009338674596187248
Age_Group attribute's correlation with the BMI attribute: 0.0036602443644796626
Region attribute's correlation with the BMI attribute: 0.0039625937256598095
Air_Pollution_Level attribute's correlation with the BMI attribute: 0.004315541524050 16
Income_Level attribute's correlation with the BMI attribute: 0.0023444983877211785
Education_Level attribute's correlation with the BMI attribute: 0.003373379412577893
Health_Insurance attribute's correlation with the BMI attribute: 0.004294061533107435
Regular_Checkups attribute's correlation with the BMI attribute: 0.003127338268325434 7
Medication_Adherence attribute's correlation with the BMI attribute: 0.006032121814659851
```

Cramer's V values ^^ Normally 0.005 independence threshold

Naive Bayes with Full MCA

Accuracy: 85.61% Recall: 0.9597



- Full MCA is applied to the dataset, which produces a set of components that combine related attributes while maintaining their meaning.
- Ensures the data better aligns with Naive Bayes' assumptions of feature independence.
- Combining all attributes at once may dilute meaningful patterns, making classification less effective

Naive Bayes with Hierarchical MCA

Assume dependencies exist between attributes.

Normalize Cramér’s V values between 0 and 1.

```
BP attribute's correlation with the BMI attribute: 0.5485564838174515
Cholesterol_Level attribute's correlation with the BMI attribute: 0.40061210623987437
Smoking_Status attribute's correlation with the BMI attribute: 0.5195606538314157
Alcohol_Consumption attribute's correlation with the BMI attribute: 0.6507495856453548
Physical_Activity_Hours attribute's correlation with the BMI attribute: 0.5802650814576912
Diabetes attribute's correlation with the BMI attribute: 0.9952301188304863
Family_History attribute's correlation with the BMI attribute: 0.4396813856534108
Diet_Type attribute's correlation with the BMI attribute: 0.776774341712039
Stress_Level attribute's correlation with the BMI attribute: 0.470210996816268
Heart_Rate attribute's correlation with the BMI attribute: 0.36732416786987526
Exercise_Induced_Pain attribute's correlation with the BMI attribute: 0.15481574946798138
Age_Group attribute's correlation with the BMI attribute: 0.6067921830729909
Region attribute's correlation with the BMI attribute: 0.6569154018125972
Air_Pollution_Level attribute's correlation with the BMI attribute: 0.7154267862366621
Income_Level attribute's correlation with the BMI attribute: 0.3886689393479007
Education_Level attribute's correlation with the BMI attribute: 0.5592359564721616
Health_Insurance attribute's correlation with the BMI attribute: 0.7118658516927804
Regular_Checkups attribute's correlation with the BMI attribute: 0.5184474658195847
Medication_Adherence attribute's correlation with the BMI attribute: 1.0
```

Threshold of 0.5:

- Values above 0.5 = moderate dependence
- Values below 0.5 = considered independent

Apply MCA (Local Level)

- Generate combinations of dependent attributes

```
[[0, 1], [2, 0, 1, 3, 5, 6, 7, 8, 10, 14, 15, 16, 18, 19, 20, 21], [3, 0, 1, 2, 4, 5, 6, 7, 10, 11, 12, 14, 15,
16, 17, 18, 21], [4, 0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 18, 19, 20, 21], [5, 1, 2, 4, 6, 8, 9, 11, 1
2, 13, 14, 15, 16, 17, 19, 21], [6, 0, 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20], [7, 0, 1
, 2, 3, 4, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20], [8, 0, 1, 2, 5, 6, 7, 11, 15, 16, 18], [9, 6, 7, 11
, 15, 20, 21], [10, 0, 1, 2, 3, 4, 6, 7, 9, 11, 12, 15, 17, 19, 21], [11, 0, 3, 4, 6, 7, 8, 9, 10, 12, 13, 15, 1
6, 17, 18, 19, 21], [12, 0, 1, 3, 4, 5, 6, 7, 10, 11, 13, 14, 15, 17, 19], [13, 1, 5, 6, 11, 12, 15, 21], [14, 2
, 3, 4, 6, 7, 12, 15, 17, 18], [15, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19], [16, 2, 3
, 5, 6, 7, 8, 11, 15, 17, 18], [17, 0, 1, 3, 5, 6, 7, 10, 11, 12, 14, 15, 16, 18, 19, 21], [18, 1, 2, 6, 7, 8, 1
1, 14, 15, 16, 17], [19, 0, 1, 2, 4, 5, 7, 10, 11, 12, 15], [20, 0, 2, 4, 6, 7, 9, 15, 21], [21, 0, 1, 2, 4, 9,
10, 11, 13]]
```

Apply MCA Again (Global Level)

- Ensures that overlapping attributes don’t reintroduce dependencies
- Create a global combination of the dataset.

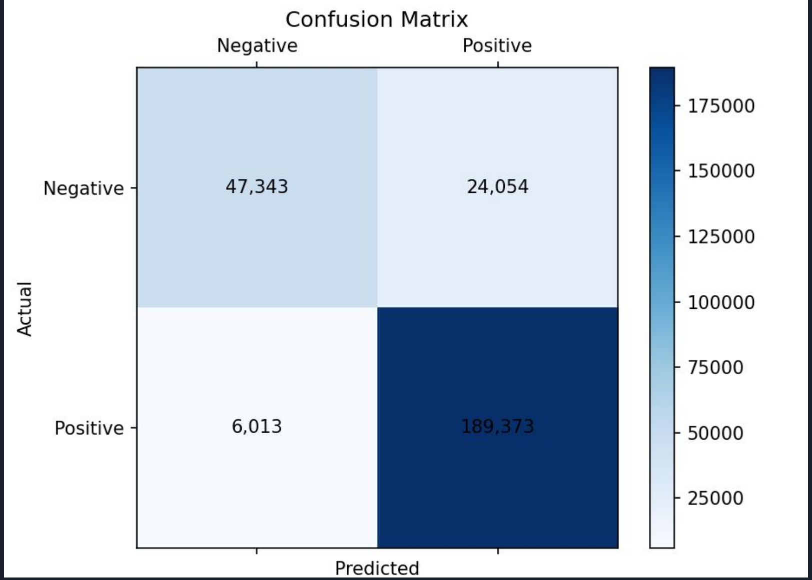
```
[*MCA-output*, *MCA-output*, *MCA-output*, *MCA-output*, *MCA-output*,
 *MCA-output*, *MCA-output*, *MCA-output*, *MCA-output*, *MCA-output*,
 *MCA-output*, *MCA-output*, *MCA-output*, *MCA-output*, *MCA-output*,
```

Breaks down complex dependencies in stages

Ensures complete independence in the final dataset

Naive Bayes with Hierarchical MCA

Accuracy: 88.73% Recall: 0.9692



Concerns

- Second round of MCA might introduce noise.
- Subtle relationships between attributes could be lost during the process.
- Potential for missed patterns leading to reduced performance

Future Improvements

Performance improvements were observed, but they may be limited to the current dataset.

- The dataset used already had many independent attributes.

Should test on datasets with more dependent attributes

- Hierarchical MCA may offer greater improvements in handling complex, dependent relationships

References

https://www.ewadirect.com/journal/aorpm/article/view/18055

Thank You