Naive Bayes with Hierarchical MCA

Lakshmi Katrapati & Ipek Sayar
Dr.Yilmaz
Machine Learning Pd. 6

1. Motivation

In this project, we aim to enhance the performance of the Naive Bayes algorithm by addressing the challenge of feature dependencies within a dataset. Traditional Naive Bayes assumes that attributes are independent, which can lead to inaccuracies when attributes are correlated. To address this, we apply Multiple Correspondence Analysis (MCA) to reduce these correlations and make the features more independent.

Related Works

An article by Luo and Liu discusses how the Naive Bayes algorithm can be improved by combining it with an advanced version of Principal Component Analysis (PCA)—a dimensionality reduction method that reduces complex data into fewer parts while keeping the most important information [2]. The authors introduce an improved PCA method that calculates correlation coefficients for both quantitative and qualitative data. By transforming the data into principal components, the relationship between correlated variables is reduced, helping to meet Naive Bayes' independence assumption and improving classification performance. The result is a more accurate and robust algorithm, particularly when dealing with noisy data.

This idea of improving Naive Bayes with PCA is the basis of our approach as well. Instead of applying traditional PCA, we focus on reducing dependencies between attributes with Multiple Correspondence Analysis (MCA), which works better with categorical data. By handling feature dependence and ensuring the data is more independent, we aim to improve the accuracy and recall of Naive Bayes, similar to how PCA enhances its performance in this study.

2. Intended Experiments

Correlation Analysis: we will calculate correlation coefficients for numerical attributes using Pearson and then use techniques like Chi-square test for categorical features to

quantify relationships. This will help determine the extent of correlation that might violate Naive Bayes' independence assumption.

Classification Model Selection: based on the correlation analysis and the results of MCA, we will choose and run various classification models. We will experiment with multiple models to evaluate how well each performs. These could include:

1. Complement Naive Bayes
2. Naive Bayes with Full MCA
3. Naive Bayes with Hierarchical MCA

Evaluation Metrics: Accuracy and Recall

We will evaluate the models by measuring the overall accuracy to determine how well they predicted the class label.

Since the dataset may be imbalanced, we will also evaluate the recall scores of the class. These metrics will help assess how well the model performs, especially in the context of potential class imbalances. Recall is the best metric to use here because we want to minimize false negatives. Heart attacks are serious, and while misclassifying someone who isn't having one may result in an unnecessary hospital visit, misclassifying someone who is having a heart attack could cost them their life in certain cases.

3. Dataset

Heart Attack in Youth VS Adult in France [1]
Heart attacks are life-threatening emergencies that can cause permanent damage to the heart and even cause death. Analyzing factors that cause heart attacks can lead to better insight on whether or not an individual faces a high risk of getting a heart attack.

This dataset gives different attributes of individuals in France and whether they had a heart attack or not.

The class label of this dataset is "Heart Attack" denoted by the values "Yes" or "No". Out of all the instances, around 20% had a value of "Yes", while the other 80% had a value of "No". if the person had a heart attack, then the value would be "Yes", else "No".

It contains 26 features that describe characteristics of the individual. individuals are detonated as “patients” (one of the features is patient_D) and examples of features include “age”, “weight_kg”, “cholestoral_level”, and “Age_Group”.

Patient_ID Patient ID	# Age Age	Sex Sex	# Weight_kg	# Height_cm	# BMI
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

Figure 1. *A few of the features included in the dataset*

This dataset contains 266,785 instances where each instance represents a patient.

4. Preprocessing

Since our algorithm implements Multiple Correspondence Analysis (MCA), the values in our dataset need to be discretized. Some features in our dataset contain continuous, numerical values, and these values were sorted into different bins based on predefined classifications.

The following attributes were discretized: “Cholestoral_Level”, “Air_Pollution_Level”, “BMI”, “Alcohol_Consumption”, “Physical_Activity_Hours”, “Stress_Level”, and “Heart_Rate”. For each range, the lower number is inclusive while the upper number is exclusive. For example, if the range of values is ‘90-190’, this means that the category contains all numbers from 90 up until but not including 190. The sources where the values of categories are taken from are cited in the bibliography.

0-200	Healthy
200-240	At_Risk
>=240	‘Dangerous’

Table 1: *Cholesterol level discretization values*

0-50	‘Good’
51-100	‘Moderate’
101-150	‘Unhealthy_Sensitive’
151-200	‘Unhealthy’
201-300	‘Very_Unhealthy’
>=300	‘Hazardous’

Table 2: *Air pollution level discretization values*

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

Table 3: *BMI discretization values*

For the ‘Alcohol_Consumption’ attribute we assumed that small numbers meant low levels of consumption, as there are no predefined categories available for this attribute.

0	No Alcohol Consumption
1-4	Low Consumption

5-9	Moderate Consumption
10-15	High Consumption
>15	Very High Consumption

Table 4: *Alcohol consumption discretization values*

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

Table 5: *Physical activity hours discretization values*

1-3	Low Stress
4-5	Moderate Stress
6-8	High Stress
9-10	Very High Stress

Table 6: *Stress level discretization values*

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

Table 7: *Heart rate discretization values*

Several other numerical features also required discretization. However, this process involved combining multiple features to create a new attribute while discarding the original ones.

There is no need for two separate blood pressure attributes, therefore the attributes “Blood_Pressure_Systolic” and “Blood_Pressure_Diastolic” were combined to create a new attribute called “BP” (BP stands for Blood Pressure).

Systolic = Blood_Pressure_Systolic
Diastolic = Blood_Pressure_Diastolic

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

Table 8: *Blood Pressure (BP) discretization values*

After the ‘BP’ attribute was created, the ‘Blood_Pressure_Systolic’ and ‘Blood_Pressure_Diastolic’ attributes were deleted, because they no longer serve a purpose.

In order to discretize the attribute “Height” we considered the gender of the patient, because what is considered “short” for a female is not considered “short” for a male.

Female & height < 160 cm	Short
Female & 160cm <= height <= 170cm	Average
Female & 170cm < height <= 180 cm	Tall
Female & height > 180 cm	Very Tall
Male & height < 165 cm	Short

Male & 165 cm <= height < 176 cm	Average
Male & 176cm <= height < 190 cm	Tall
Male & height >= 190 cm	Very Tall

Table 9: *Height attribute discretization values*

Attributes deleted: “Patient_ID”, “Age”, and “Weight_kg” .

- ‘Patient_ID’ was deleted because it gave no information relevant to the algorithm. It just contains the name-ID of the patient.
- ‘Age’ was deleted because there was another feature in the dataset called “Age_Group” that categories the patient as either a “youth” or “adult”. This feature essentially discretizes the “Age” attribute, so “Age” can be deleted.
- ‘Weight_kg’ was deleted because weight is primarily used to calculate the “BMI” and there is already an attribute for BMI.
    - Height is also used to calculate BMI, but we kept the ‘Height’ attribute because we wanted to see if height affected the chances of an individual receiving a heart attack, whereas weight was already accounted for in the calculation for BMI.

Lastly, the class label ‘Heart_Attack’ was moved to the last column from the 19th column (column numbering starts at 1) to indicate that ‘Heart_Attack’ is a class.

5. Method

We will improve the Naive Bayes algorithm by addressing its weakness, the assumption of attribute independence, using Multiple Component Analysis [9]. MCA is a dimensionality reduction technique specifically designed for categorical variables, similar to how PCA (Principal Component Analysis) works for continuous data. PCA isn't the right choice here to begin with because it combines different values into a single number. if we do PCA before discretizing, it merges different categories into a single, unclear number, leading us to lose the ability to interpret what each value means. Instead, we should discretize first to keep things clear. MCA produces new numerical variables called dimensions that summarize the patterns in the categorical data. These dimensions retain meaning because they are based on how different categories relate to each other. Additionally, Naive Bayes is inherently designed for categorical data since it calculates probabilities using frequency tables. By transforming continuous data into categorical

and applying MCA, we can create components that are not only tailored for this probabilistic approach but also easier to interpret compared to the continuous outputs of PCA. In conclusion, MCA's output is easier to interpret than PCA's continuous values in our context and also helps structure the data in a way that makes sense for Naive Bayes classification.

First, we will discretize our data so that we'll be able to use MCA. Next, we will use the Pearson Chi-Square Test [10] to determine if two categorical attributes are dependent or independent. if the p-value is small, it indicates a significant relationship. Then, Cramér's V [10] will be calculated to measure the strength of this relationship, with values closer to 1 indicating a stronger association. Together, they will help us tell whether two attributes are correlated to each other.

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.

Our Naive Bayes model is Complement Naive Bayes. A model specifically designed to perform better on imbalance datasets, which is a common issue when dealing with real world data. In our case, the dataset has an 8-2 split between the two classes which is a significant imbalance that could greatly affect our classification. Traditional Naive Bayes can struggle in such situations because it tends to be biased towards the majority class. Complement Naive Bayes addresses this issue by modifying the way it calculates probabilities, making it more robust to class imbalance. Instead of assuming equal importance for all classes, it focuses on the complement of the minority class, effectively improving the model's sensitivity to underrepresented classes.

Approach 1: Full MCA Naive Bayes

We assume all the attributes have the utmost dependence with all the other attributes. We will use MCA on the whole entire dataset without doing any kind of processing to see which attributes are more or less correlated with each other than the others.

Approach 2: Hierarchical MCA

We assume that dependencies exist between attributes and then assess the strength of these dependencies. For each attribute, we normalize the Cramér's V values to a scale between 0 and 1. Using a threshold of around 0.5, we classify any value above it as indicating a moderate dependence, while values below 0.5 are considered sufficiently independent. Each attribute will have a list of other attributes it is dependent on based on these correlations.

Next, we apply MCA to generate combinations of these dependent attributes. After creating these individual combinations, we use MCA once again to create a global combination of all the data. This second step is essential because the subsets of data generated from the first MCA may still contain overlapping attributes, which could reintroduce the dependencies we aim to eliminate. By performing MCA a second time on the entire dataset, we can fully remove any remaining dependencies, ensuring that the final dataset is completely independent.

This hierarchical approach to MCA is beneficial because it allows us to break down complex dependencies step by step, ensuring that we don't overlook subtle interactions between attributes. By addressing dependencies at both the local (attribute-specific) and global (dataset-wide) levels, we maximize the clarity and interpretability of the data, ensuring that the relationships between attributes are captured accurately without introducing noise from hidden correlations.

6. Results

Complement Naive Bayes

Accuracy: 79.96% Recall: 0.9410

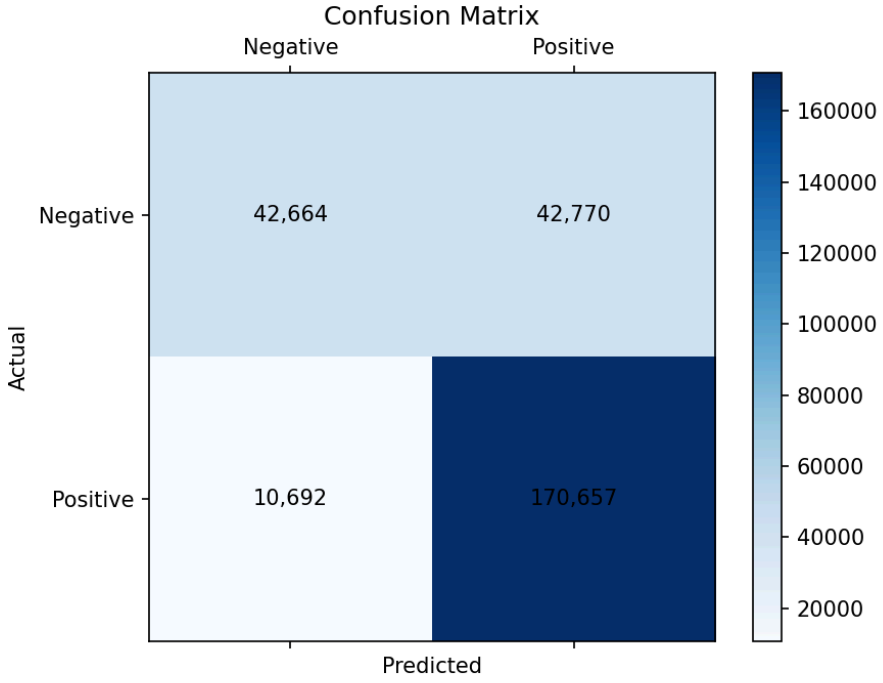

Figure 2. *Confusion matrix for Control: Conditional Naive Bayes. The matrix shows the classification results for the control group.*

	Sex	Height_cm	BMI	BP	Cholesterol_Level	Smoking_Status	Alcohol_Consumption	Physical_Activity_Hours	Diabetes	Family_History	Diet_Type	Stress_Level	Heart_Rate	Exercise_Induced_Pain	Age_Group	Region	Air_Pollution_L	Income_Level	Education_Lev	Health_Insurar	Regular_Checl	Medication_Adherence
Sex		1	0.5502172422	0.9052566168	0.5617960562	0.3946510796	0.684262038	1	0.6922326461	0.3318951659	0.8603642411	0.6998863735	0.6306332097	0.3127507882	0.1997609031	0.5272346397	0.2639549888	0.831131049	0.4738430881	0.5099761247	0.5342843925	0.775433112
Height_cm	1		0.5922755162	0.8200042452	0.6360842512	0.6520589916	0.5696262563	0.7121101003	0.5599151568	0.254845809	0.9063229386	0.4196269189	0.5155911524	0.7650688227	0.4823444493	0.7728769757	0.4496828505	0.7586007618	0.6923698422	0.7262583683	0.4560924323	0.5376426851
BMI	0.02201502987	0.02369784547		0.7289709218	0.5342165033	0.6535296915	0.7038449302	0.6086204612	1	0.4864954095	1	0.4989424792	0.4157040822	0.1731927295	0.612626307	0.6632314454	0.7624567185	0.4910929771	0.593407165	0.7629606674	0.6913500317	1
BP	0.02725636516	0.0246895021	0.5485564838		0.6228165772	0.481566275	0.5115005731	0.5609090014	0.2745487661	0.3195931065	0.8622971039	0.6932582252	0.5454138565	0.312354797	0.738085295	0.5729915136	0.5431501482	0.9508128361	0.4131001305	0.2073528003	0.2250412063	0.4093997104
Cholesterol_Level	0.01685662162	0.01908562978	0.4006121062	0.6206628976		0.6378793148	0.5402547085	0.6565342688	0.4303867224	0.4302850158	0.5833702458	0.6547986085	0.5363167467	0.3014390332	0.7324594094	0.6470499678	0.3636985188	0.2537791499	0.4350484851	0.5187418591	1	0.5819417208
Smoking_Status	0.01255362958	0.02074163094	0.5195606538	0.5087633851	0.6762428422		0.6194370245	0.2887001646	0.5833470539	0.493740365	0.4061957288	0.4840138424	0.899716566	0.6207855267	0.4072317403	0.6134570088	0.554999469	0.7299665366	0.2133992933	0.6567173505	0.4480453627	0.4760536352
Alcohol_Consumption	0.02531304884	0.0210723027	0.6507495856	0.6284516784	0.6660835455	0.7203824806		0.7632994831	0.8380741951	0.6316469143	0.6090235265	0.5867646249	0.7321712893	1	0.8244133786	0.6717339767	0.8123209687	0.997944715	0.9810578393	0.4133266951	0.9388321157	0.3699578237
Physical_Activity_Hours	0.03814740247	0.0271651006	0.5802650815	0.710658819	0.8347002024	0.3462230583	0.787114542		0.5652756398	1	0.9958183822	0.687388888	0.8322456292	0.3256293188	0.6666073446	0.6171748738	0.5568832003	0.997587185	0.5107467072	0.8709033731	0.7750353683	0.4384270786
Diabetes	0.02756516973	0.02229619828	0.9952301188	0.3631046906	0.5711834384	0.73026349	0.9021298763	0.5900704861		0.4713936827	0.1536559921	0.5968444575	0.3440027264	0.3324970251	0.0633729651	0.9219748057	0.5572391726	0.2092699792	0.6493785803	0.41420402	0.2949588436	0.02193018889
Family_History	0.0120017712	0.00921556384	0.4396813857	0.3838358972	0.5185716664	0.561289652	0.6174428962	0.9479369343	0.4280747192		0.5228722527	0.8694119152	0.4863691513	0.3979191993	0.3846877122	0.6113154646	0.1976385654	0.336823849	0.2162092374	0.1949043935	0.9115927122	0.7177158115
Diet_Type	0.02674005564	0.02816844849	0.7767743417	0.8901037922	0.6042718698	0.396880072	0.5116724428	0.7787361973	0.1199280748	0.4493979989		0.5706576332	0.8282857373	0.2097696445	0.2783316581	0.7442500842	0.1172340728	0.8019073365	0.2076588516	0.6483043795	0.3326368608	0.604355408
Stress_Level	0.02639094384	0.0158230691	0.4702109968	0.8682128	0.8228928334	0.5737585261	0.608286893	0.6794616044	0.5651711123	0.9065850559	0.6923457973		0.6109132538	0.5947381575	0.3520371774	0.5379741825	0.8298261126	0.651522189	1	0.6888473292	0.4396505731	0.5793891973
Heart_Rate	0.0222960092	0.01822870236	0.3673241679	0.640442632	0.6319457588	1	0.6997469199	0.7713240077	0.3054243006	0.4755234104	0.9422161103	0.572799276		0.5723307397	0.8921150764	0.7499703005	0.2982757439	0.6699481328	0.4556414553	0.8893111813	0.3705591116	0.3615529801
Exercise_Induced_Pain	0.01118584969	0.02736346374	0.1548157495	0.3710415241	0.3593173617	0.6980011446	0.9668265531	0.3053016108	0.2986412098	0.3935691456	0.2413977698	0.5641167367	0.5789850382		0.3409647194	0.7246210918	0.282892277	0.2815360168	0.0659749991	0.0494852468	0.270619297	0.5361247969
Age_Group	0.00791662204	0.01911554584	0.6067921831	0.9714931449	0.9674335155	0.5073585978	0.8831865625	0.6925235421	0.0630702935	0.4215929273	0.3549049613	0.3699904927	1	0.3778055174		1	0.2766632931	0.7263763682	0.9362471477	0.0717421392	0.2398835561	0.1043784338
Region	0.02089456599	0.03062949161	0.6569154018	0.7541910553	0.8546245936	0.7642888728	0.7196225063	0.6411682477	0.9175714205	0.6699623306	0.5490046845	0.5654099782	0.8406654257	0.8029154659	1		1	1	0.7352997203	1	0.6047485251	0.3605974151
Air_Pollution_Level	0.00990980157	0.0168826808	0.7154267862	0.677265193	0.455076872	0.6550457561	0.8244049688	0.548067727	0.5253734587	0.2051929148	0.1416149105	0.8262182747	0.3167398628	0.2969515587	0.2620940776	0.9473395428		0.7809377886	0.7613051689	0.491378402	0.1509545267	0.1469023393
Income_Level	0.02631907985	0.02402229354	0.3886689393	1	0.267833737	0.7266285092	0.8542513634	0.8281082	0.1664177064	0.294957615	0.8170439656	0.5471457393	0.6000564087	0.2492668873	0.5804081242	0.7990459899	0.658692243		0.7468875195	0.4803920582	0.3658481597	0.427658395
Education_Level	0.01786742362	0.02610751447	0.5582359565	0.5173524287	0.5467303624	0.2529634841	1	0.5048565422	0.6149173541	0.22545362	0.2519404364	1	0.4859597693	0.0695562583	0.8908169529	0.6996202422	0.7646295515	0.8893677803		0.1199116965	0.2673925474	0.3880879056
Health_Insurance	0.01903842154	0.02711266722	0.7118658517	0.2570956733	0.6454172015	0.7707323181	0.417114901	0.8522874536	0.3883173629	0.201214052	0.7787177868	0.6819879301	0.9390409884	0.05165189175	0.06758121536	0.9420016758	0.4886096877	0.5663380201	0.1187176407		0.3204343963	0.2261345775
Regular_Checkups	0.01603113753	0.01368499738	0.5184474658	0.2242630211	1	0.4226275886	0.7614787582	0.6096048345	0.2222514894	0.7563943574	0.3211310288	0.3498421323	0.3144843595	0.2270283263	0.1816196824	0.4578647942	0.120643212	0.3466502615	0.2127716525	0.2575430808		0.4225406363
Medication_Adherence	0.03102625984	0.02151190269	1	0.5440469543	0.7760196619	0.5988546751	0.3936535262	0.4586513668	0.02203529462	0.7941328869	0.778032146	0.614791837	0.4091727768	0.599764024	0.105382001	0.3640644505	0.1565592423	0.5403571346	0.4118013893	0.2423655914	0.5634582123	

Table 10. Normalized Cramer's V outputs

Plain Naive Bayes performed pretty well on this dataset. This tells us that most of the attributes in the dataset are already mostly independent of one another. We can also see this by the calculated Cramer's V outputs, most of them are under 0.015, meaning that they can be considered to be independent. However, in this project, we will try to further improve this accuracy and recall.

Complement Naive Bayes with Full MCA:

Accuracy: 85.61% Recall: 0.9597

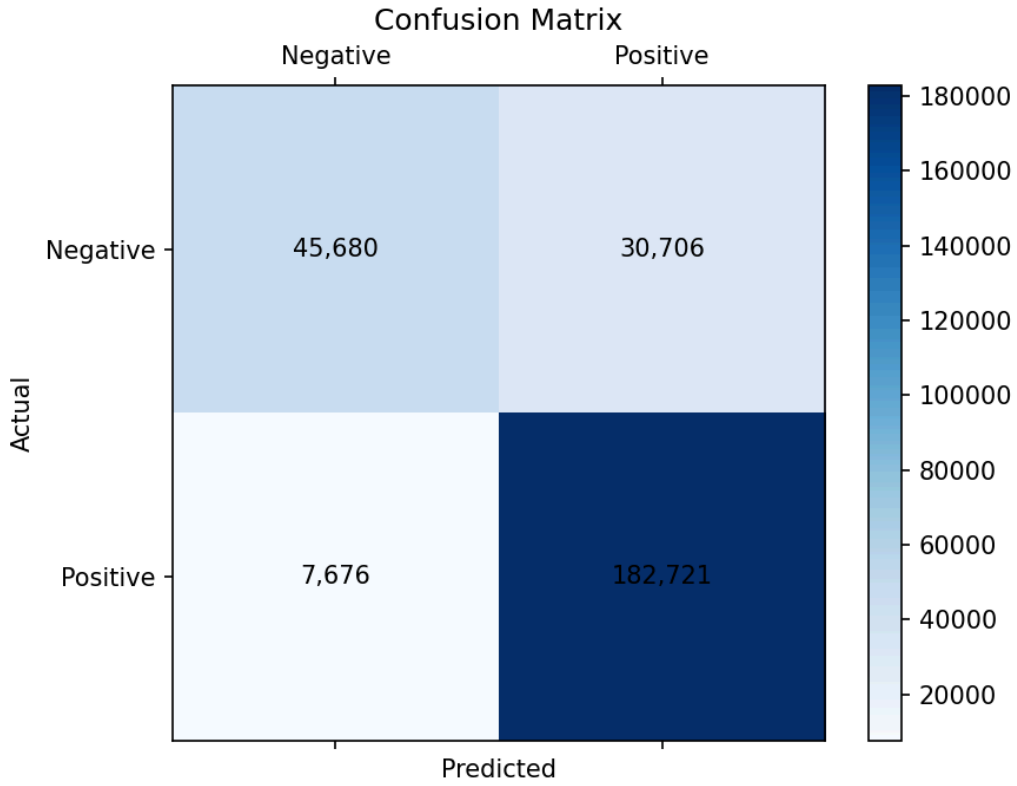

Figure 3: *Confusion Matrix for Approach 2: Full MCA*

Complement Naive Bayes with Cramer's V and Hierarchical MCA:

Accuracy: 88.73% Recall: 0.9692

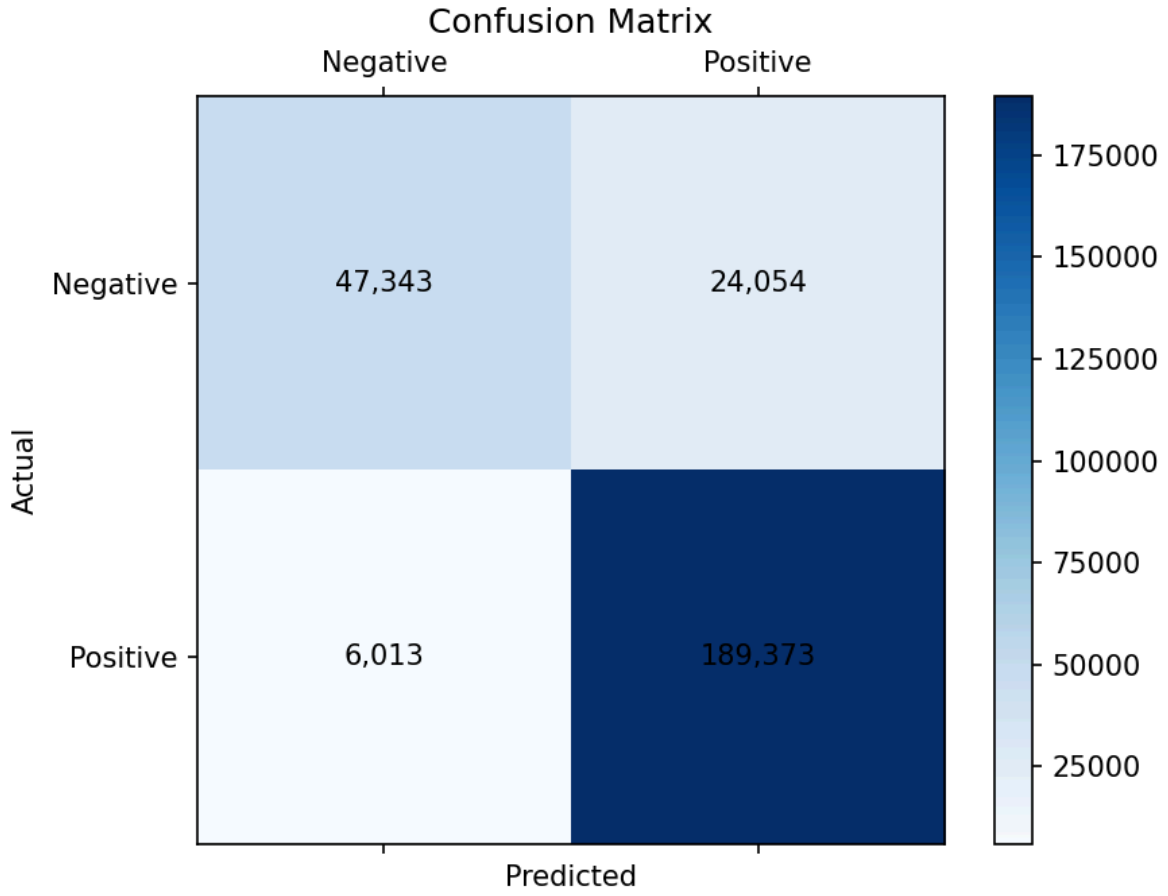

Figure 4: *Confusion Matrix for Approach 3: Hierarchical MCA*

7. Analysis

Overall, each modification of the Naive Bayes model—whether using full MCA or hierarchical MCA—resulted in improvements in performance compared to the original model, especially in terms of recall.

Initially, I had concerns that hierarchical MCA might reduce accuracy and recall, as the second round of MCA could potentially introduce noise by losing subtle relationships during the process of combining dependencies. Specifically, if the global combination fails to fully capture these relationships, the model could miss important patterns, leading to lower performance. However, contrary to this concern, the hierarchical MCA version performed slightly better than the full MCA version, achieving an accuracy of 89 and a recall of 96, compared to the full MCA model's

accuracy of 86 and recall of 95. This suggests that the hierarchical approach helped to further refine the dataset, allowing the model to better handle complex interactions between attributes.

Looking at the numbers, it's clear that each modification didn't dramatically improve the accuracy of the original Naive Bayes model (which had an accuracy of around 80 and recall of 94). However, considering the original model was already performing quite well, the improvements in accuracy and recall with each modification are notable. The full MCA and hierarchical MCA both contributed to slight but meaningful improvements in recall, which is critical when aiming to minimize false negatives, particularly in high-stakes scenarios like detecting heart attacks.

The model's performance suggests that while the dataset we used already contained many independent attributes, further testing on a dataset with more dependent attributes could reveal whether these improvements are consistent across different types of data. it's possible that hierarchical MCA could offer even more substantial gains when dealing with more complex, dependent relationships between attributes.

8. References

[1] Panday, A. (2021). *Heart attack in youth vs adult in France*. Kaggle. https://www.kaggle.com/datasets/ankushpanday1/heart-attack-in-youth-vs-adult-in-france

[2] Luo,L.;Liu,T. (2024). Integrating advanced principal component analysis into naive bayes for enhanced classification performance. Advances in Operation Research and Production Management,3,27-31.

[3] AirNow.gov, U.S. EPA. (n.d.). *Aqi Basics*. AQI Basics | AirNow.gov. https://www.airnow.gov/aqi/aqi-basics/

[4] Centers for Disease Control and Prevention. (n.d.). *Adult BMI categories*. Centers for Disease Control and Prevention. https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

[5] MediLexicon International. (n.d.). *What is the average height for men?*. Medical News Today. https://www.medicalnewstoday.com/articles/318155#what-is-the-link-between-height-and-weight

[6] professional, C. C. medical. (2025, February 6). *What should my cholesterol levels be?*. Cleveland Clinic. https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean

[7] *Target heart rates chart*. www.heart.org. (2024a, August 12). https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates

[8] *Understanding blood pressure readings*. www.heart.org. (2024b, December 18). https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

[9] Abdi, Hervé & Valentin, Dominique. (2007). *Multiple Correspondence Analysis*. Encyclopedia of Measurement and Statistics.

[10] McHugh M. L. (2013). *The chi-square test of independence.* Biochemia medica, 23(2), 143–149. https://doi.org/10.11613/bm.2013.018