Demand for health care in Australia

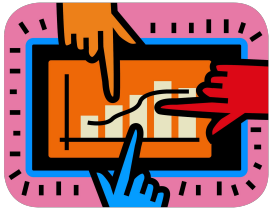
- Final project for Statistical Methods -

group O

Buscema Andrea, Cusma Fait Omar, Derin Tanja, Špringer Christian, Živanović Uroš

UNIVERSITY OF TRIESTE

1st March, 2024



Abstract & Dataset Overview

- Initial data exploration
- Predicting the number of doctor/specialist consultations in the past 2 weeks

Individuals	Variables	Response Variable
5190	19	"doctorco"

Methodologies:

- Negative Binomial Regression
- Zero-Inflated Poisson Regression
- Zero-Inflated Negative Binomial Regression
- Hurdle Poisson Regression
- Hurdle Negative Binomial Regression
- Logistic Regression
- Random Forest
- Naive Bayes
- Neural Networks
- K-Nearest Neighbors

For the project it was used a **dual-language strategy**, combining R and Python to leverage the strengths of both. This approach enabled **efficient handling** of various stages of the project, from data preprocessing to complex modeling and deployment.

The dataset and the models

The dataset contains information divided into these 19 variables:

Categorical (Binary) variables	• '**sex**': gender of the individual (1 if female, 0 if male); • '**levyplus**' • '**freepoor**' ⟩⟩ INSURANCE TYPE • '**freepera**' • '**chcond1**': chronic condition(s) - limited in activity • '**chcond2**': if chronic condition(s) + limited in activity
Categorical (Ordinal) variables	• '**Age**': age divided by 100 (mid point) • '**Agesq**': age squared • '**Income**': Annual income divided by 1000 • '**hscore**': general health score indicates bad health
Discrete variables	• '**illness**': number of illnesses in the past 2 weeks • '**actdays**':number of days in the past 2 weeks with activity limitation due to illness or injury; • '**doctorco**': (Response variable) Number of consultations with a doctor or specialist in the past 2 weeks; • '**nondocco**': consultations with non-doctor in the past 2 weeks • '**hospadmi**': admissions to a hospital past 12 months • '**hospdays**': nights in a hospital past 12 months • '**medicine**': prescribed and non prescribed medications past 2 weeks • '**prescrib**': prescribed medications used in past 2 weeks; • '**nonpresc**': nonprescribed medications used in past 2 weeks

Considerations and insights

The Australian healthcare system in 1977-1978 was undergoing significant changes, primarily due to the implementation of Medibank, a universal health insurance system.

Medibank Introduction:

Introduced in 1975 by the Whitlam Government through the Health Insurance Act 1973.

Aimed to provide universal health coverage to all Australians and free treatment in public hospitals.

Funded by a 2.5% levy on taxable incomes, an additional levy for high-income earners, and government funds.

Transition in Healthcare System:

In 1981, the Fraser Government abolished Medibank, replacing it with Medibank Private, a government-subsidized private insurance scheme.

The dataset analyzed in this project corresponds to the period when the original Medibank scheme was operational.

Challenges and Debates:

The cost of running Medibank was a significant concern for the government.

There were debates on whether Medibank achieved equitable access for all Australians.

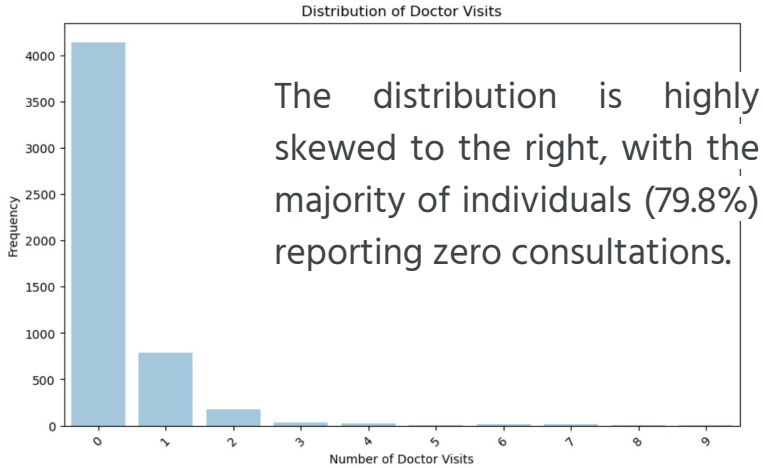
Concerns were raised about longer waiting times in the public healthcare system.

This period represents a pivotal moment in Australian healthcare history, with Medibank being a crucial element in shaping the future direction of the nation's health policy and insurance systems.

Exploratory Data Analysis
response variable

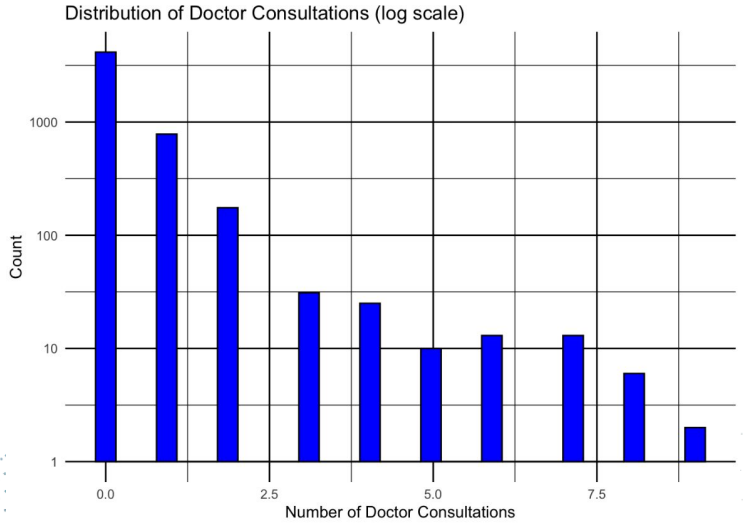
visits	0	1	2	3	4	5	6	7	8	9
freq.	4141	782	174	30	24	9	12	12	5	1

The variable 'doctorco' (doctor consultations) is the response variable in the dataset, reflecting the number of consultations with a doctor or specialist in the past 2 weeks.

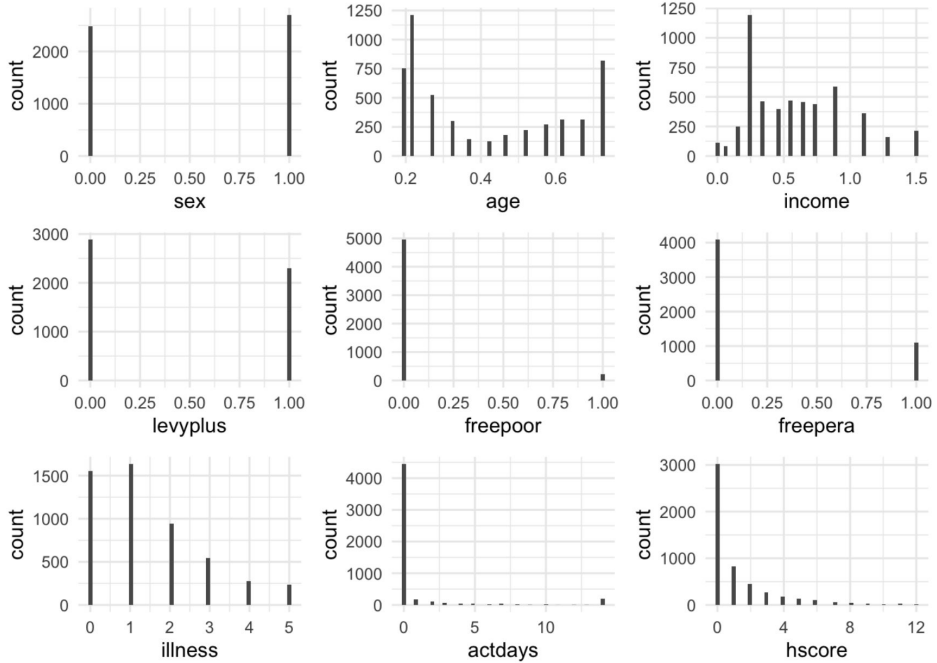


The distribution is highly skewed to the right, with the majority of individuals (79.8%) reporting zero consultations.

FOR A BETTER VISUALISATION



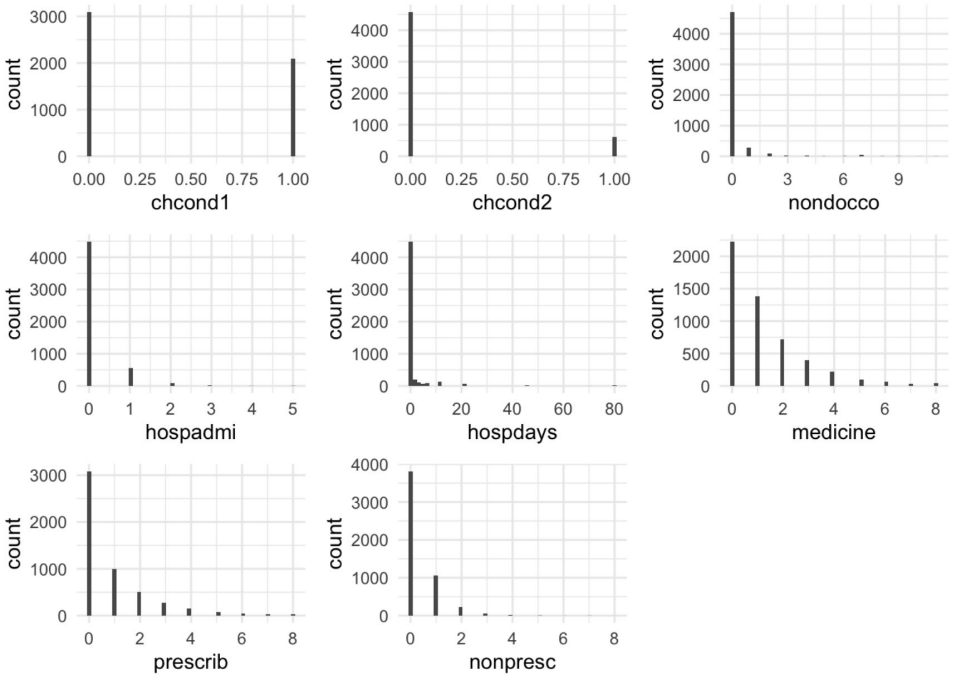
Exploratory Data Analysis
variables



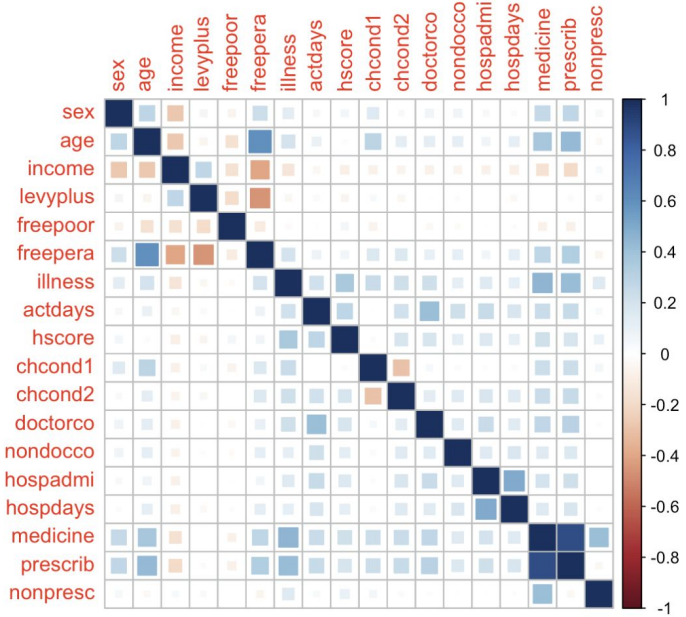
- **Gender (‘sex’)**: Slight female majority with approximately 52% of the sample.
- **Age (‘age’)**: Range from 19 to 72 years old; median age of 32 suggesting a distribution slightly skewed towards older individuals.
- **Income (‘income’)**: Data divided by 1000 and coded into ranges, with the most common income level being 0.25.
- **Private Health Insurance (‘levyplus’)**: 2,298 individuals (about 44.2% of the sample) have private health insurance coverage for a private patient in a public hospital.
- **Government Health Care (‘freepoor’)**: 222 individuals (approximately 4.2%) covered by government healthcare due to low income, recent immigration, or unemployment.
- **Old-age/Disability Pension (‘freepera’)**: 1,091 individuals (approximately 21.02%) covered for health care due to old-age or disability pension.
- **Illnesses (‘illness’)**: On average, individuals reported about one to two illnesses in the past two weeks.
- **Reduced Activity Days (‘actdays’)**: Average of 0.86 days of reduced activity in the past two weeks, with 85.8% reporting no days of reduced activity.
- **Health Score (‘hscore’)**: Mean score is 1.218, suggesting generally good health with 58.3% reporting no health issues.

Exploratory Data Analysis
variables

- **Chronic Conditions without Limitation (‘chcond1’)**: About 40.31% of individuals have chronic conditions without limiting their activities.
- **Chronic Conditions with Limitation (‘chcond2’)**: 605 individuals (about 11.66%) have chronic conditions that limit their activities.
- **Non-doctor Consultations (‘nondocco’)**: Very few consultations with non-doctor health professionals; heavily skewed towards zero consultations.
- **Hospital Admissions (‘hospadmi’)**: 86.5% reported no hospital admissions in the past year, with 10.8% having one admission.
- **Hospital Days (‘hospdays’)**: 86.5% had no nights in a hospital; smaller proportions reported 1 to 80 nights with higher frequencies for shorter stays.
- **Medication Use (‘medicine’)**: 42.9% did not use any medication in the past two days; on average, individuals used just over one medication.
- **Prescribed Medications (‘prescrib’)**: 59.4% did not use any prescribed medications in the past two days; average use is less than one prescribed medication.
- **Non-prescribed Medications (‘nonpresc’)**: 73.5% did not use any non-prescribed medications; on average, one-third of a medication was used in the past two days.



Exploratory Data Analysis
Correlation Matrix

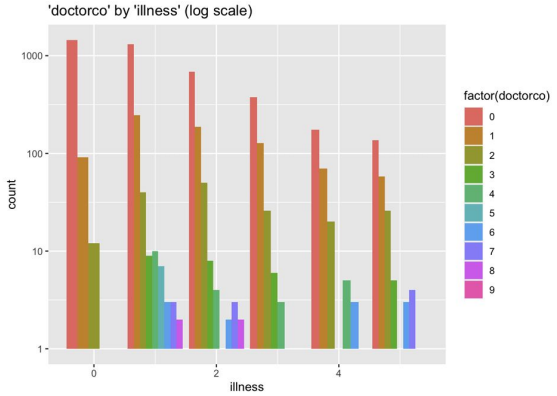


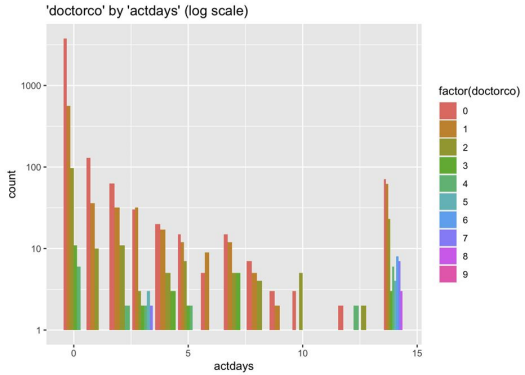
Considerations:

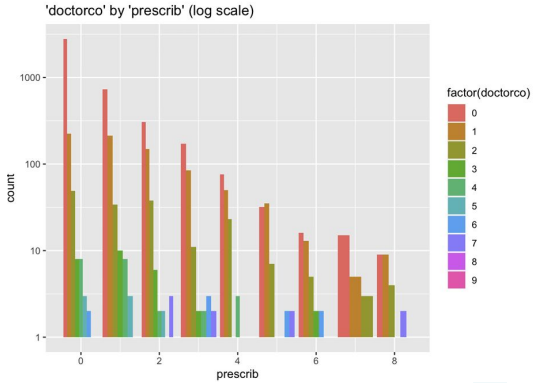
- **Correlation and Causation:** The observed correlations do not imply causation, and there may be other underlying factors affecting these patterns.

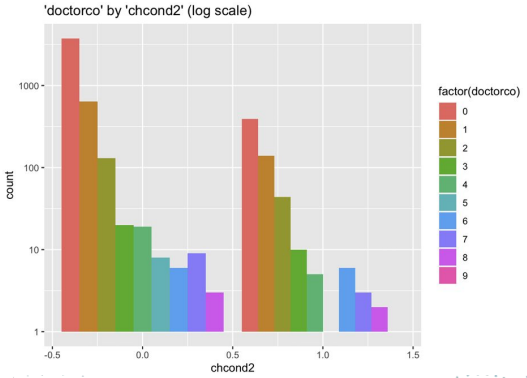
- **Healthcare Disparities**: The disparities related to income and healthcare access highlight a need for further research into the socioeconomic factors affecting healthcare utilization.

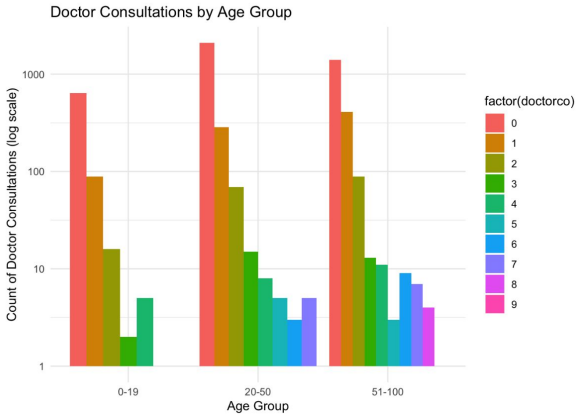
Bivariate analysis

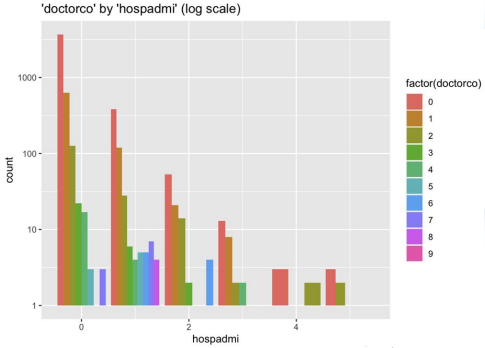












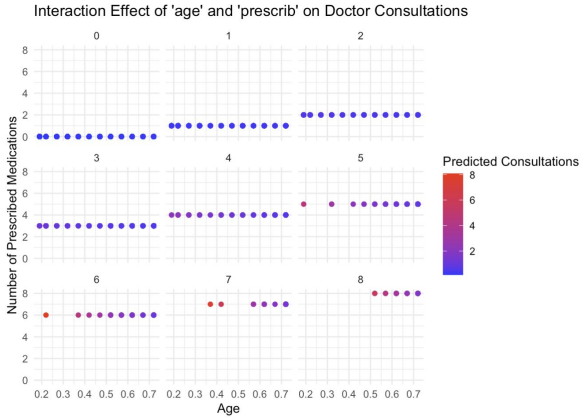
Bivariate analysis

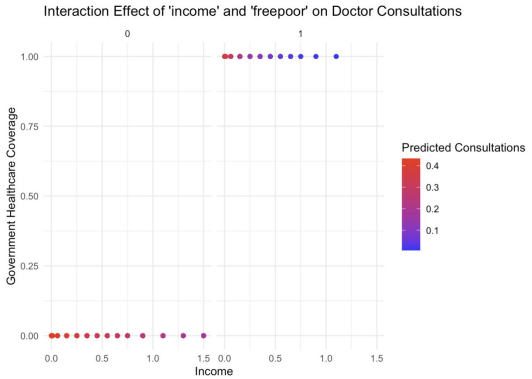
Age: Older individuals have more doctor consultations, reflecting greater healthcare needs as age increases.
Sex and Income: Certain demographic groups, differentiated by sex and income levels, show varying frequencies of doctor visits, hinting at disparities in health conditions, access to care, or health-seeking behaviors.
Healthcare Coverage:

- **Private Insurance ('levyplus')**: No significant correlation with doctor consultations, indicating that private insurance status does not strongly influence medical visit frequency.
- **Government Coverage ('freepoor', 'freepera')**: Significant associations, particularly with 'freepera', suggest that government-supported individuals, such as older adults, those with disabilities, or veterans, are more likely to have more doctor consultations.

Health Episodes ('illness', 'actdays'): A clear link exists between the number of illnesses and activity-limiting days with the frequency of medical visits, underlining that health problems significantly drive healthcare seeking behavior.
Perceived Health ('hscore'): Poorer health scores correlate with more frequent doctor consultations, aligning with the likelihood of seeking medical attention when feeling unwell.
Non-doctor Professional Interactions ('nondocco'): Increased interactions with non-doctor health professionals are related to more doctor consultations, possibly reflecting comprehensive care needs or more complex health conditions.
Chronic Conditions ('chcond1', 'chcond2'): A strong association, especially with 'chcond2', suggests that chronic conditions, particularly those that limit activity, lead to more doctor visits.
Hospital Utilization ('hospadmi', 'hospdays'): Both the number and length of hospital stays are associated with a higher frequency of doctor consultations, indicating that hospital events are typically followed by increased outpatient care.
Medication Use ('medicine', 'prescrib', 'nonpresc'): Prescribed medication use correlates strongly with more doctor consultations, implying the necessity of ongoing medical supervision. However, the use of non-prescribed medication is not significantly associated with doctor visit frequency, suggesting that self-medication practices might not result in higher healthcare utilization.

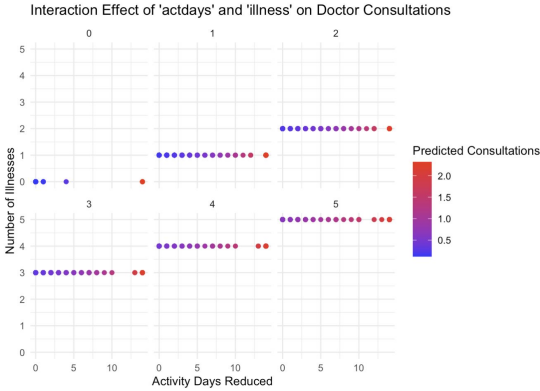
Combined variables





Quick Summary of Variable Effects on Doctor Consultations

- **Activity Days and Illness**: More 'actdays' and illnesses lead to more doctor visits. Interaction is highly significant ($p < 0.001$).
- **Age and Medications**: Older age and more prescriptions increase consultation frequency. Interaction greatly improves model fit.
- **Income and Government Coverage**: Low income with government healthcare correlates with more doctor visits. Interaction is significant ($p = 0.0080$).
- **Income and Private Insurance**: Higher income and private insurance impact consultation frequency. Significant model improvement ($p = 0.0050$).
- **Sex and Health Score**: The effect of health score on doctor visits differs by sex. Interaction effect is significant ($p = 0.0149$), requiring further analysis for detailed interpretation.



These results underscore the importance of considering how different variable combinations influence the frequency of doctor consultations.

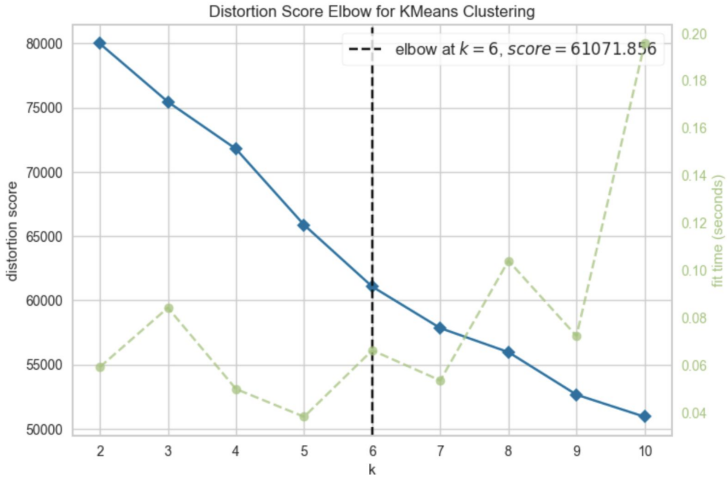
Cluster analysis

Cluster Analysis in Healthcare:

- **Patient Grouping**: Identifies similar patient groups for better disease and treatment insights.
- **Disease Subtyping**: Aids in tailoring treatments, especially in complex diseases like cancer.
- **Resource Management**: Helps allocate resources effectively, prioritizing high-risk patients.
- **Predictive Analytics**: Uncovers patterns for forecasting health trends.
- **Healthcare Delivery**: Informs the design of targeted healthcare programs.
- **Risk Stratification**: Segregates patients into risk categories for proactive healthcare.
- **Medical Insights**: Discovers new data correlations, useful in genomics and epidemiology.
- **Cost Efficiency**: Reduces healthcare costs, especially in chronic disease management.
- **Performance Benchmarking**: Allows comparison and improvement of healthcare outcomes.

Elbow Method in Cluster Analysis:

- **Optimal Clusters**: Determines the best number of clusters for data analysis.
- **K-Means Clustering**: Uses K-Means with a range of cluster values and identifies the 'elbow' point as the optimal k.
- **Python Implementation**: Employs feature selection, standardization, and K-Elbow Visualizer for finding optimal k.
- **Elbow Point**: At k=6, the rate of decrease in distortion score changes, indicating the optimal cluster count.
- **Centroids and Clustering**: Re-runs K-Means with optimal k and assigns labels to data, using centroids for cluster profiling.



Cluster analysis

From centroids to Cluster Profiling

"The Healthy Young Adults"

- Age: Younger age group.
- Health: Relatively healthy with low actdays, chcond1, chcond2, and hscore.
- Healthcare Utilization: Lower hospadmi and hospdays, indicating fewer hospital admissions and shorter stays.
- Income: Higher than average income, possibly indicating better access to health resources.
- Insurance: Almost all have private health insurance (levyplus).
- Gender: Slightly more females than males (sex).

"The Stable Middle-Aged"

- Age: Middle-aged group.
- Health: High chcond1, but low chcond2, suggesting chronic conditions without severe limitations.
- Healthcare Utilization: Low hospadmi and hospdays.
- Income: Higher income, possibly at peak career stage.
- Insurance: Mostly covered by private insurance.

"The Aging with Care Needs"

- Age: Older individuals, but not as old as Cluster 1.
- Health: Many chronic conditions (chcond1 and chcond2).
- Healthcare Utilization: Moderate hospadmi and higher hospdays.
- Income: Lower income, which may impact their healthcare options.
- Insurance: Some with private health insurance.
- Gender: A higher proportion of females.

"The High-Needs Elderly"

- Age: Older age group.
- Health: Higher actdays, chcond1, chcond2, indicating more chronic conditions and health issues.
- Healthcare Utilization: Highest hospadmi and hospdays, suggesting frequent and longer hospital stays.
- Income: Lower income, which could be related to retirement.
- Insurance: Mixed insurance coverage.
- Gender: More females than males.

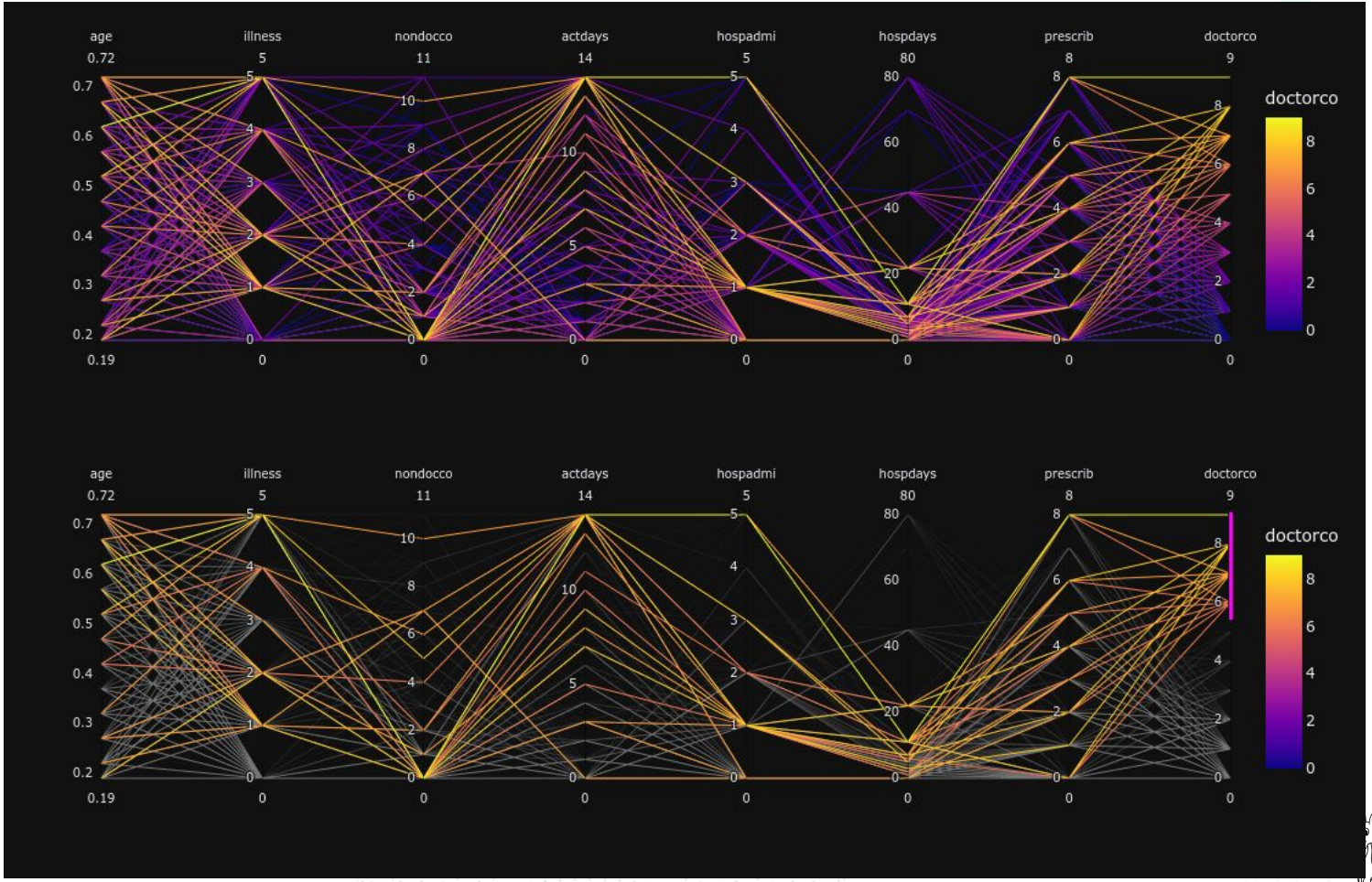
"The Young and Occasionally Unwell"

- Age: Young, similar to Cluster 0.
- Health: Moderate health issues, higher than Cluster 0 but less severe than other clusters.
- Healthcare Utilization: Moderate hospadmi and hospdays.
- Income: Similar to Cluster 0, relatively higher income.
- Insurance: Lacks private health insurance.
- Gender: More females than males.

"The Economically Disadvantaged"

- Age: Younger, but with health issues.
- Health: Moderate actdays, some chronic conditions.
- Healthcare Utilization: Low to moderate hospadmi and hospdays.
- Income: Low income, suggesting economic challenges.
- Insurance: Lacks private health insurance, possibly relying on public assistance (freepoor).
- Gender: Balanced gender distribution.

Visualization, Dim. Reduction, and Clustering


age
illness
nondocco
actdays
hospadmi
hospdays
prescrib
doctorco
doctorco

Dim. Reduction

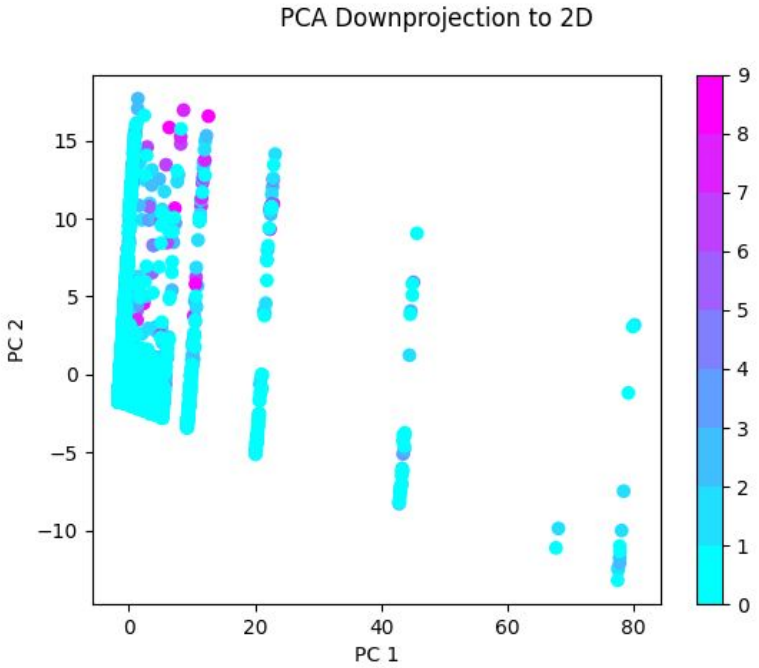
2 Methods were tried:

1. **PCA**

   PCA works very well for continuous data, but our dataset is all categorical variables.

2. **MCA**

   Multiple correspondence analysis. Like PCA but for categorical data.



PCA has issues with categorical data

- We can still clearly see that some points with high amounts of doctors visits do stand out.
- Doctors visits are not included in the down-projection.

MCA

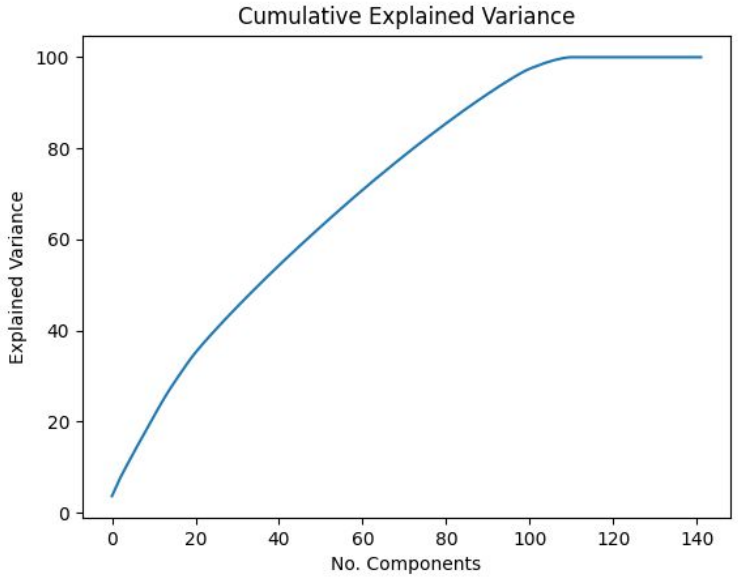
How does MCA work?

1. Let y_{ik} be a value in the indicator matrix and let p_k be the sum of row k in the indicator matrix.
2. We normalize the indicator matrix: $x_{ik} = y_{ik}/p_k - 1$
3. Apply un-standardized PCA to this matrix.

The *indicator matrix* is the one-hot encoding of the dataset.

The method presented here is based on PCA, but there is much more theory around MCA than is presented here.

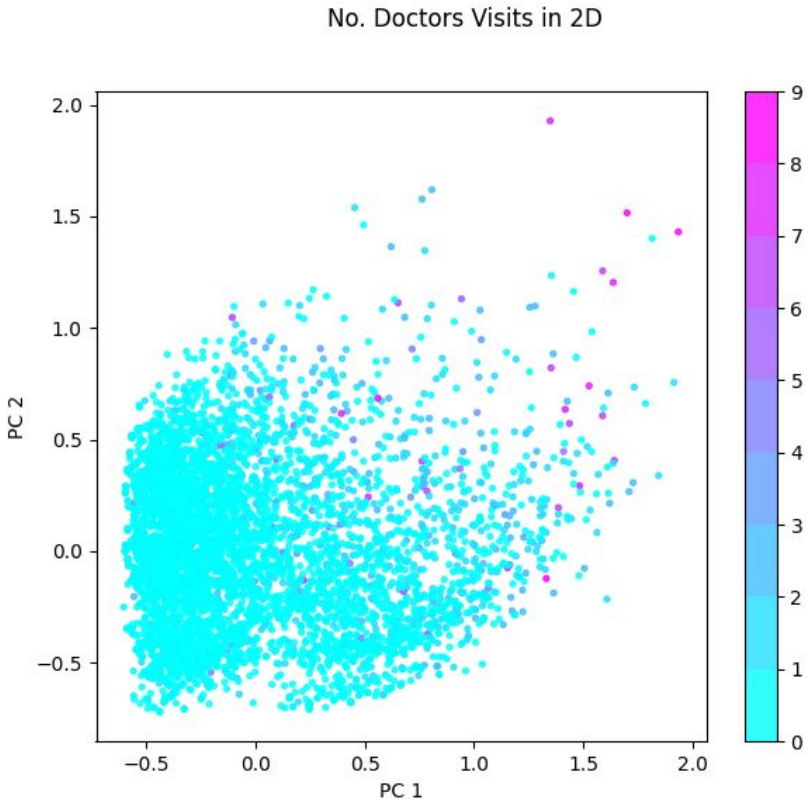
Applying MCA



component	eigenvalue	% of variance	% of variance cumulative
0	0.236	3.62%	3.62%
1	0.134	2.05%	5.67%
2	0.129	1.98%	7.64%
3	0.115	1.76%	9.41%
4	0.111	1.71%	11.11%
5	0.111	1.70%	12.82%

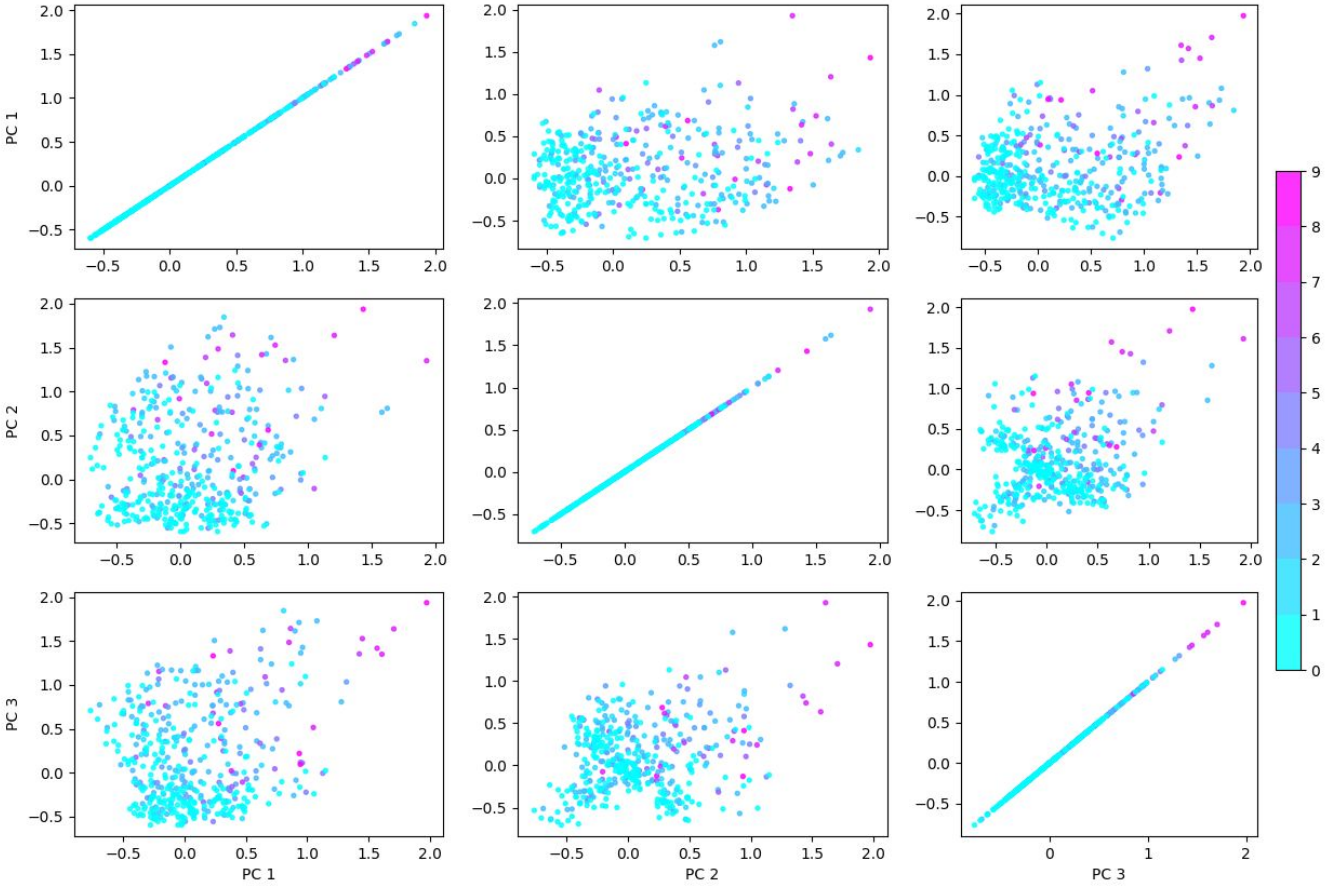
- A little disappointing, as there's no clear elbow in the graph.
- Some variables get completely removed.
- There are many components because the one-hot encoding introduces extra dimensions.

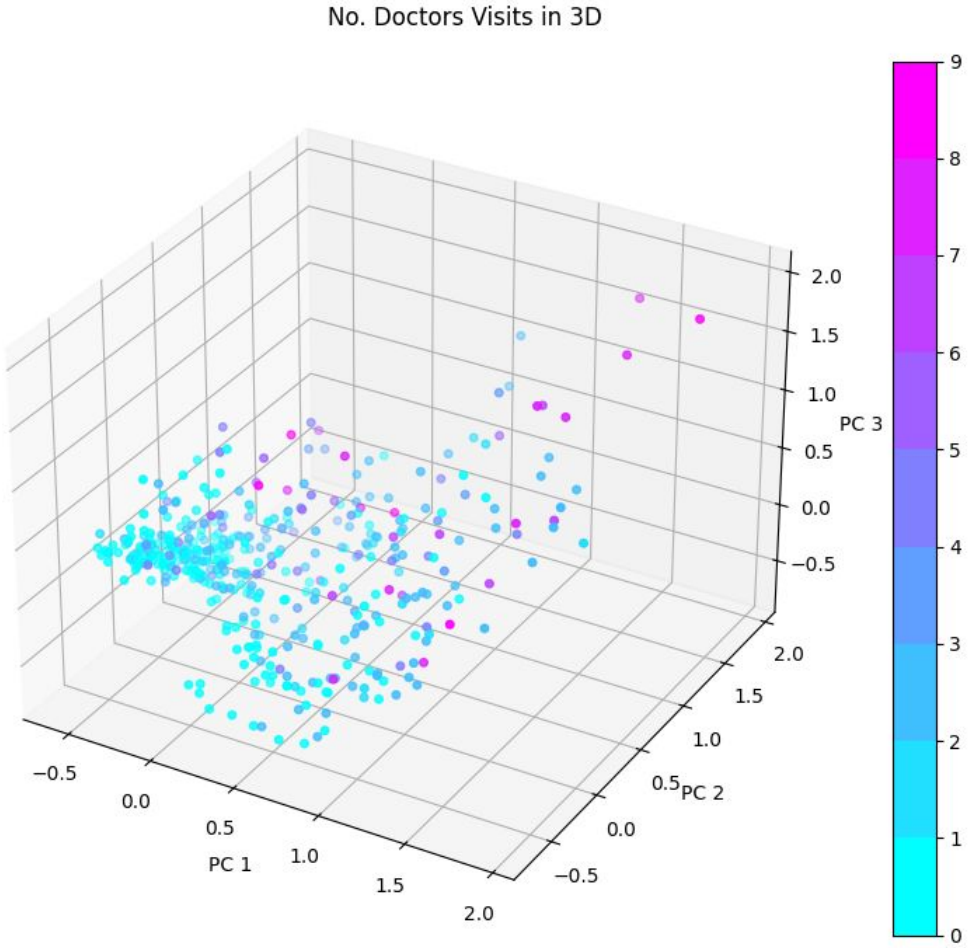
Plotting MCA



- Doctors visits has not been included in the downprojection.
- We can see that values with a high number of doctors visits are slightly separated.

Principal Component Combinations and No. Doctors Visits

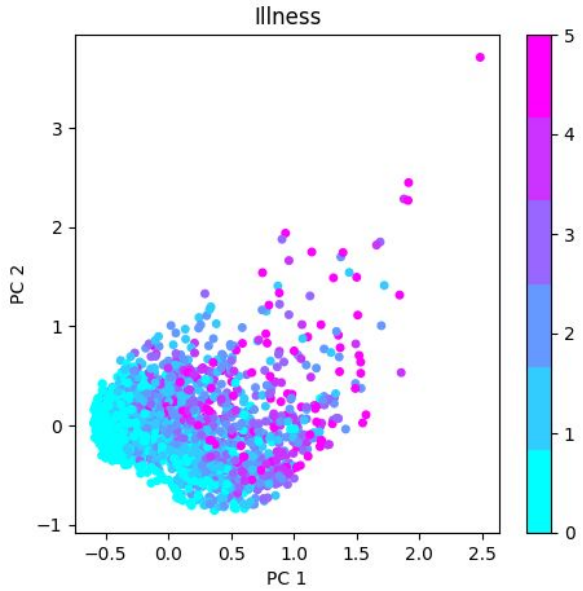


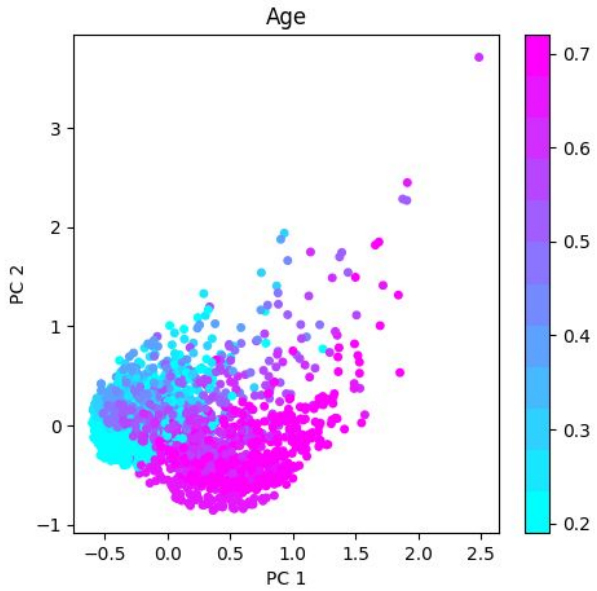


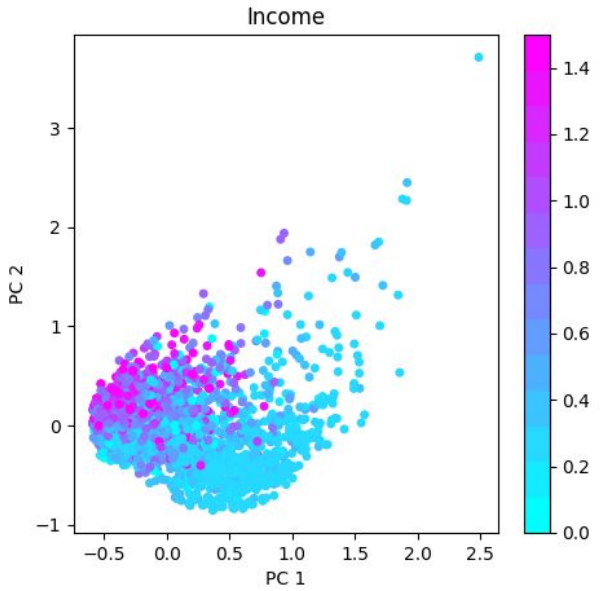
In 3D we can see that there do seem to be some clusters in the data. However, they don't seem to be strictly related to doctors visits.

Exploring Other Variables

Doctors visits is included in this downprojection.







Clustering Continued

Finding Meaning

We have some clusters, but now what?

- Analyze the cluster centroids to see what they may represent.
- Use clusters to help us identify groups in our data.

```
Cluster 0 centroid values:
actdays     3.000838e-01
age         3.372590e-01
chcond1     9.723386e-02
chcond2     1.274099e-01
cluster     6.694049e+00
freepera    2.498002e-16
freepoor    0.000000e+00
hospadmi    1.089690e-01
hospdays    3.914501e-01
hscore      1.010897e+00
illness     1.062867e+00
income      6.539648e-01
levyplus    9.983236e-01
medicine    8.365465e-01
nondocco    1.047779e-01
nonpresc    3.989941e-01
prescrib    4.375524e-01
sex         5.389774e-01
Name: 0, dtype: float64
```

Groups We Found

Cluster 0: "Healthy Young Adults"

- Strongest features: actdays, income, and hscore.

Cluster 1: "High-Needs Elderly"

- Strongest features: actdays, hospdays, and hscore.

Cluster 2: "Stable Middle-Aged"

- Strongest features: actdays, income, and age.

Cluster 3: "Young and Occasionally Unwell"

- Strongest features: actdays, income, and hospdays.

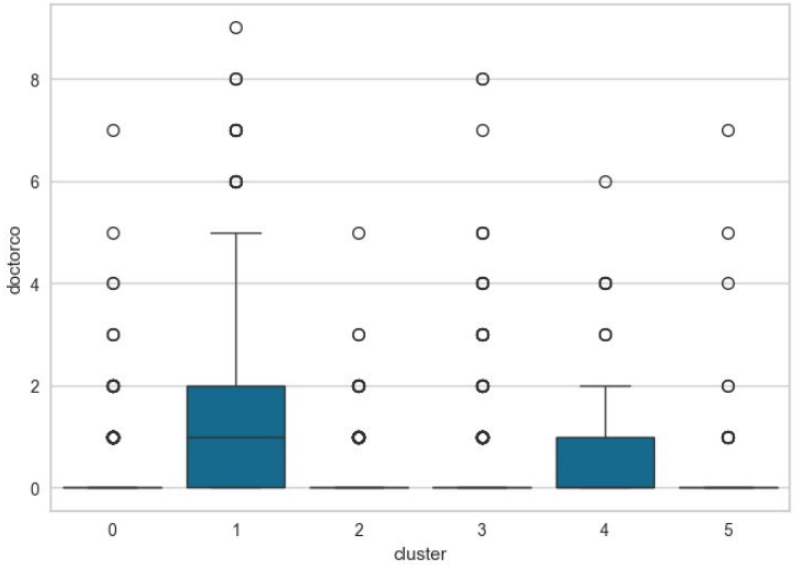
Cluster 4: "The Aging with Care Needs"

- Strongest features: illness, income, and hscore.

Cluster 5: "Economically Disadvantaged"

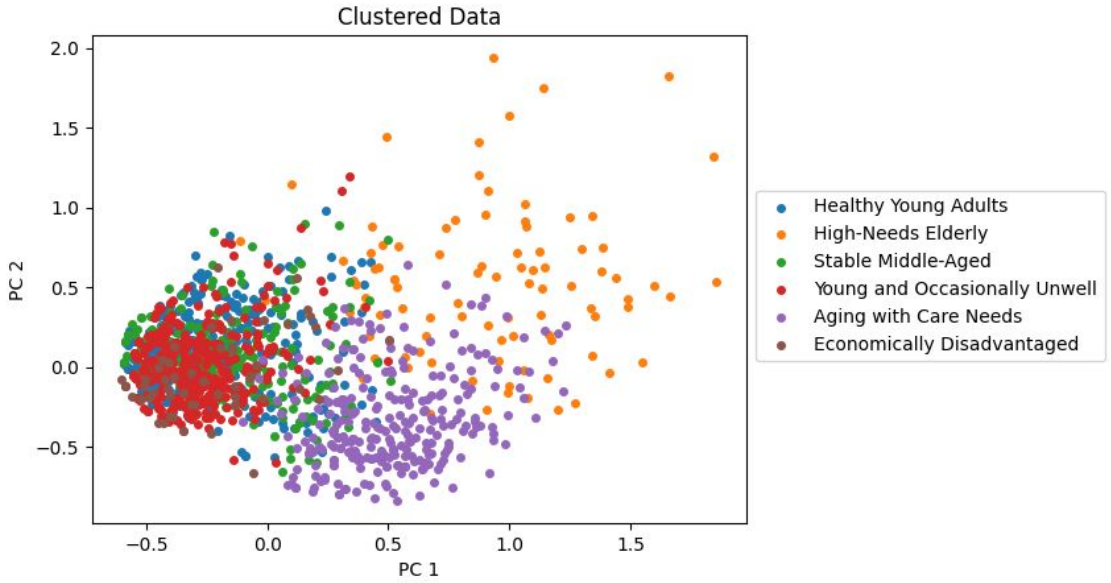
- Strongest features: prescrib, hscore, and actdays.

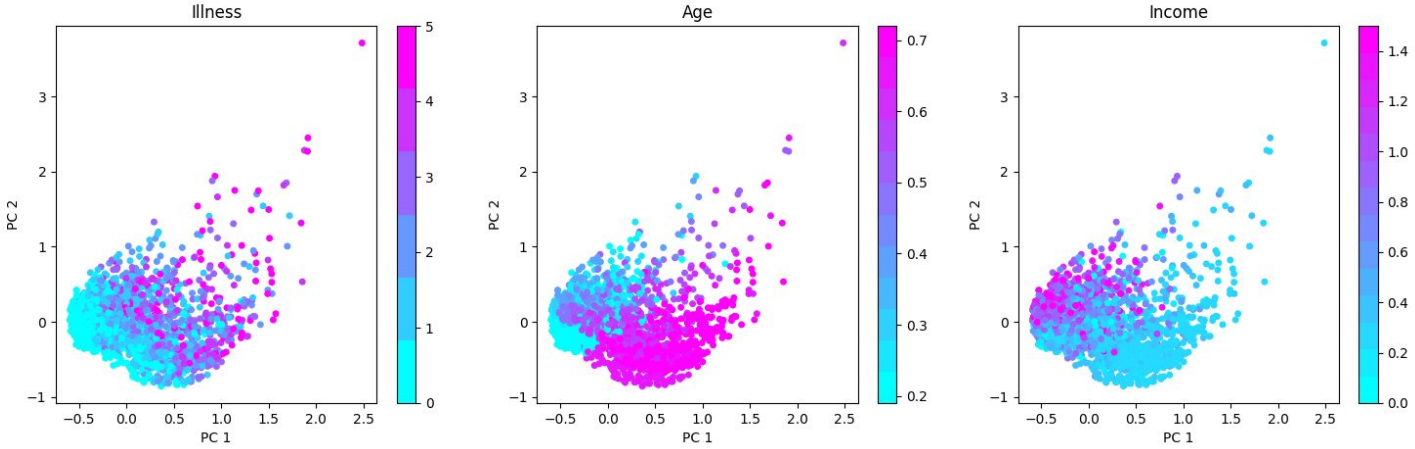
Looking at doctors visits



- **Cluster 0**: "Healthy Young Adults"
- **Cluster 1**: "High-Needs Elderly"
- **Cluster 2**: "Stable Middle-Aged"
- **Cluster 3**: "Young and Occasionally Unwell"
- **Cluster 4**: "The Aging with Care Needs"
- **Cluster 5:** "Economically Disadvantaged"

A more visual approach

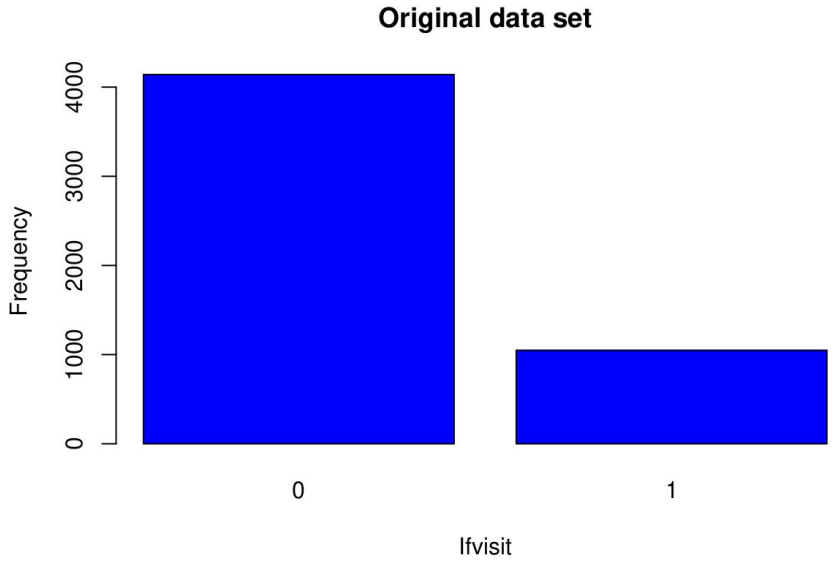




Binary classification problem

We can transform the multi-class classification problem of the variable “doctorco” into a binary classification problem of the “ifvisit” variable.

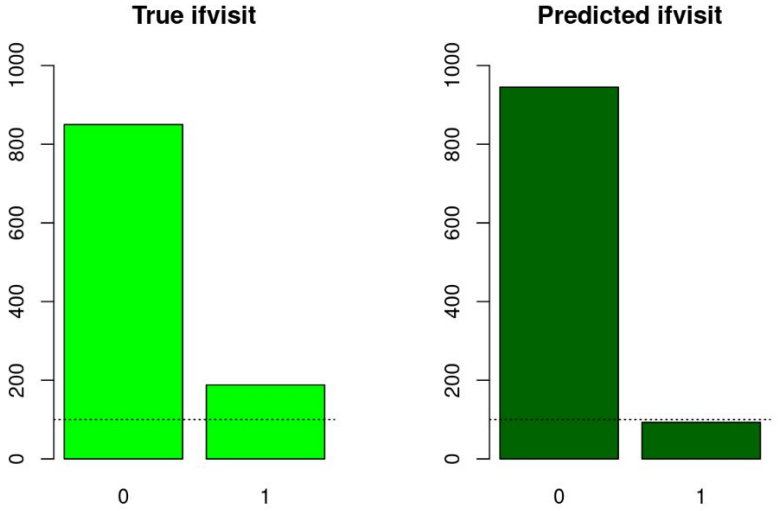
➔ data$ifvisit = ifelse(data$doctorco == 0, 0, 1)



GAM - first attempt

```
model_gam <- gam(ifvisit ~ s(hospdays) + s(actdays) + age*prescrib + freepoor + hscore + nonpresc + illness, d
ata = train_data, family = binomial(link = "logit"))
summary(model_gam)
```

```
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.72414    0.13763 -19.794  < 2e-16 ***
age           1.52680    0.27834   5.485 4.13e-08 ***
prescrib      0.91275    0.10486   8.704  < 2e-16 ***
freepoor     -0.91922    0.30227  -3.041  0.00236 **
hscore        0.06140    0.02004   3.064  0.00219 **
nonpresc     -0.18891    0.06426  -2.940  0.00328 **
illness       0.16493    0.03493   4.722 2.34e-06 ***
age:prescrib -1.01709    0.17178  -5.921 3.20e-09 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(hospdays) 4.286  4.955  17.17 0.00514 **
s(actdays)  3.259  3.931 172.31 < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

R-sq.(adj) =  0.205   Deviance explained = 18.8%
UBRE = -0.16331  Scale est. = 1         n = 4152
```

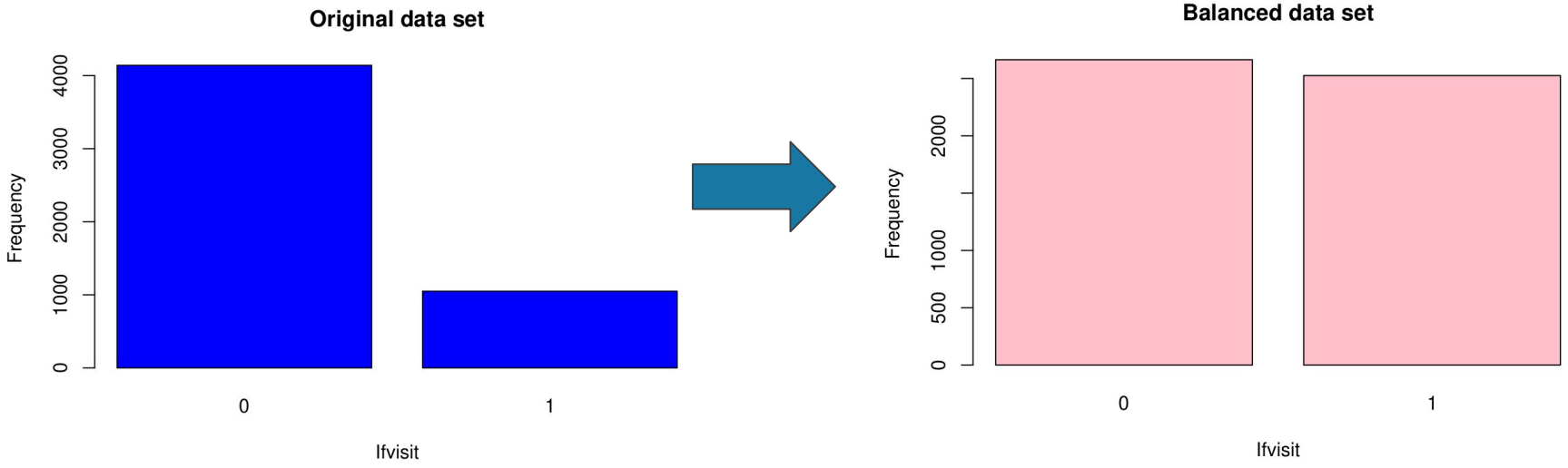



```
"AIC (GAM): 3473.93"
"MAE (GAM): 0.1686"
"Number of true ifvisit: 188"
"Number of ifvisit predicted: 93"
```

Balancing with ROSE

Since we transformed the problem into a binary classification problem we can use techniques to balance the data set like ROSE (Random Over-Sampling Examples)

```
data.rose <- ROSE(ifvisit ~ ., data = data, seed = 1, hmult.majo = 0)$data
```

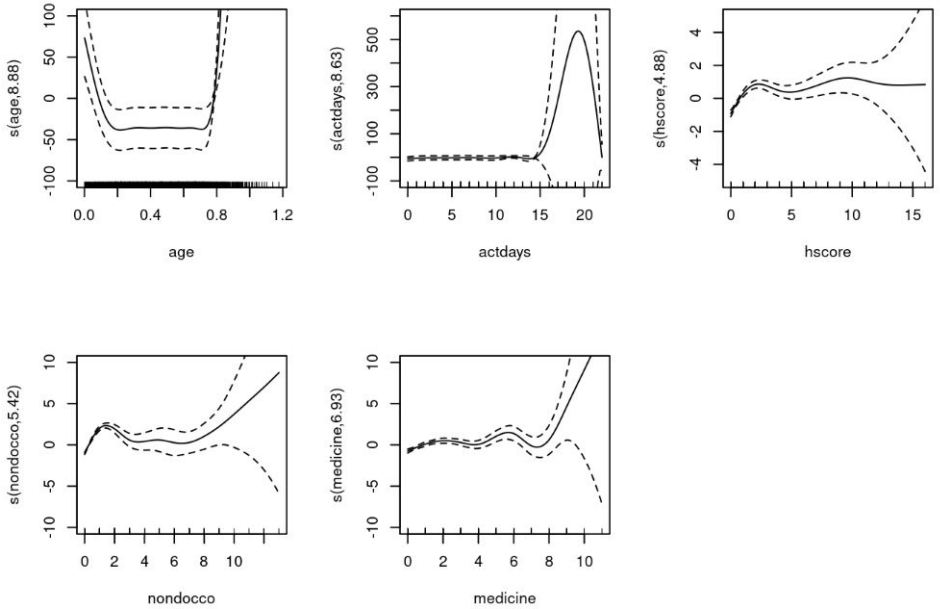


GAM after ROSE

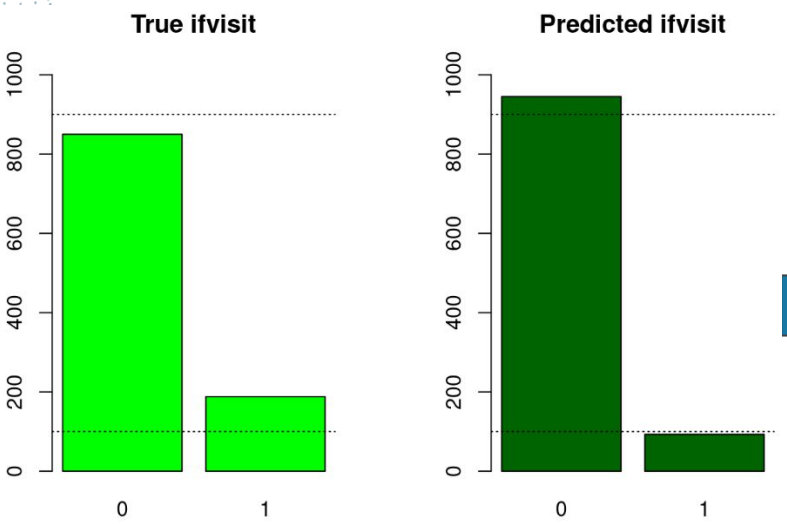
```
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    40.91      13.05   3.135  0.00172 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(age)      8.882  8.986 140.36  <2e-16 ***
s(actdays)  8.626  8.860 596.54  <2e-16 ***
s(hscore)   4.882  5.811  85.30  <2e-16 ***
s(nondocco) 5.421  6.249 287.02  <2e-16 ***
s(medicine) 6.930  7.430  49.68  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.843   Deviance explained = 79.4%
UBRE = -0.69709  Scale est. = 1         n = 4152
```

"AIC (GAM): 1257.7" "Number of true ifvisit: 529"

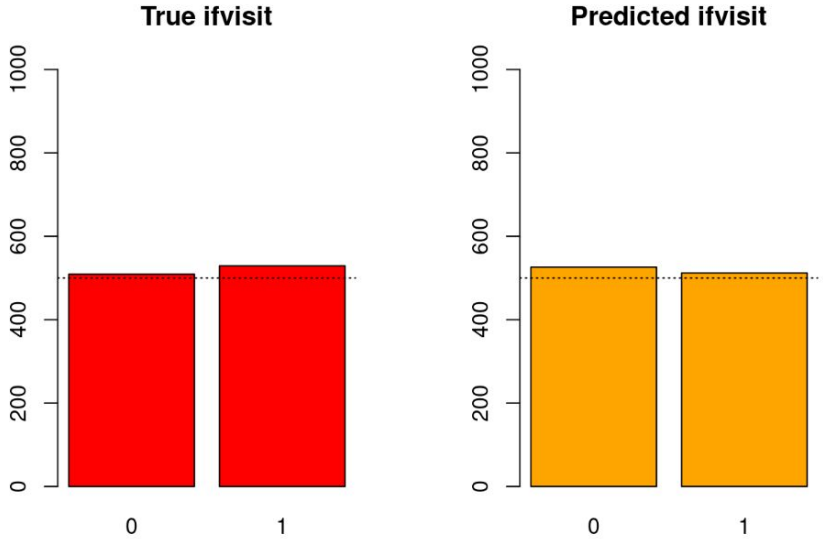
"MAE (GAM): 0.05491" "Number of ifvisit predicted: 512"



GAM after ROSE



GAM before ROSE



GAM after ROSE

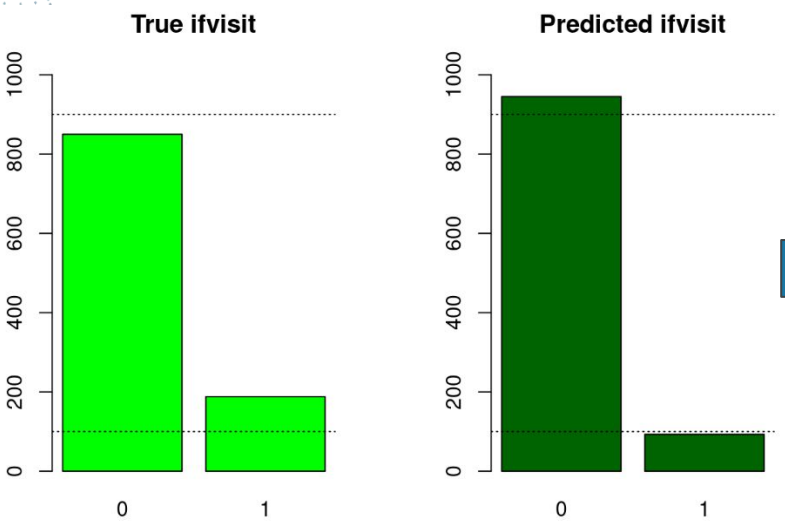
GLM after ROSE

```
model_glm <- glm(ifvisit ~  hospadmi + nondocco + illness + actdays + prescrib + nonpresc, d
ata = train_data, family = binomial)
summary(model_glm)
```

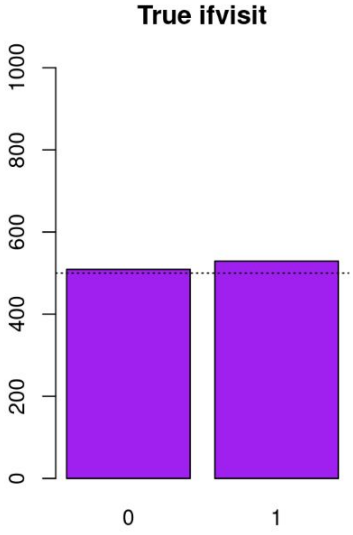
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.58171    0.09089 -28.404  < 2e-16 ***
hospadmi     0.95475    0.09502  10.048  < 2e-16 ***
nondocco     1.22241    0.08169  14.965  < 2e-16 ***
illness      0.11996    0.03324   3.609 0.000307 ***
actdays      0.53808    0.02988  18.005  < 2e-16 ***
prescrib     0.47077    0.03747  12.563  < 2e-16 ***
nonpresc     0.25345    0.05829   4.348 1.37e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5749.9  on 4151  degrees of freedom
Residual deviance: 2947.5  on 4145  degrees of freedom
AIC: 2961.5

Number of Fisher Scoring iterations: 6
```

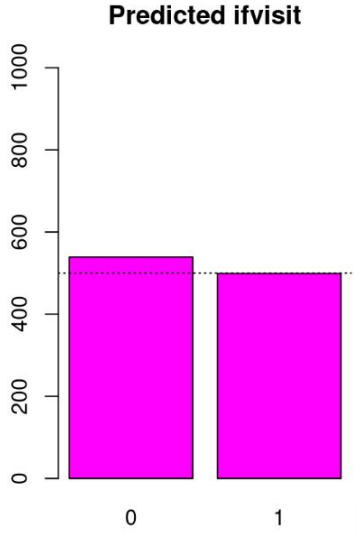
	GAM	GAM R	GLM R
AIC	3473	1257	2961
MAE	0.1686	0.0549	0.1021
Predicted “ifvisit”	93	512	499
True “ifvisit”	188	529	529

GLM after ROSE



GAM before ROSE





GAM after ROSE

Zero Inflated Negative Binomial model

The ZINB model is combining the principles of NB regression with a mechanism to account for excess zeros.

Differentiates between two sources of zeros:

- those arising from the data's natural variability "sampling zeros" and
- those that are structurally inherent or "excess zeros"

Another key feature is the dispersion parameter r of the Negative Binomial distribution, which is used to model overdispersion in the count data part of the model

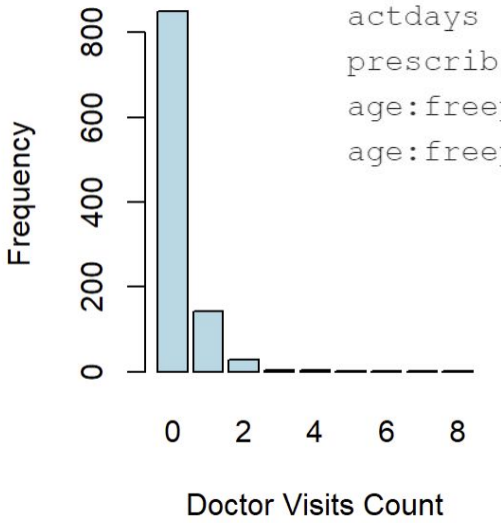
```
ZINB_model <- zeroinfl(doctorco ~income_factor*levyplus + illness*actdays
+ age:chcond2 +  hospadmi + prescrib + nonpresc|age:freepoor + freepera +
illness + actdays + prescrib, data = train_data, dist = "negbin")
```

```
Count model coefficients (negbin with log link):
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.533666   0.187994  -2.839 0.004529 **
income_factorMiddle          -0.425590   0.153377  -2.775 0.005523 **
income_factorHigh            -0.491655   0.162270  -3.030 0.002447 **
levyplus1                    -0.588873   0.248310  -2.372 0.017715 *
illness                       0.107232   0.035167   3.049 0.002294 **
actdays                       0.106329   0.013240   8.031 9.65e-16 ***
hospadmi                      0.195201   0.043981   4.438 9.07e-06 ***
prescrib                      0.086300   0.024309   3.550 0.000385 ***
nonpresc                     -0.157661   0.048440  -3.255 0.001135 **
income_factorMiddle:levyplus1  0.756097   0.266787   2.834 0.004596 **
income_factorHigh:levyplus1    0.692485   0.269249   2.572 0.010114 *
illness:actdays              -0.007566   0.004499  -1.682 0.092621 .
age:chcond20                 -0.512984   0.228037  -2.250 0.024477 *
age:chcond21                 -0.664422   0.263900  -2.518 0.011812 *
Log(theta)                    0.903989   0.167827   5.386 7.19e-08 ***
```

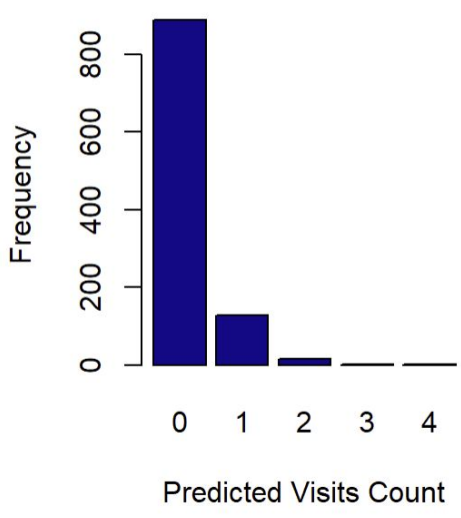
```
ZINB_model <- zeroinfl(doctorco ~income_factor*levyplus + illness*actdays
+ age:chcond2 +  hospadmi + prescrib + nonpresc|age:freepoor + freepera +
illness + actdays + prescrib, data = train_data, dist = "negbin")
```

```
Zero-inflation model coefficients (binomial with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.9429     0.3034   6.404 1.51e-10 ***
freepera1       -1.0961     0.4214  -2.601 0.009290 **
illness         -0.3753     0.1119  -3.355 0.000795 ***
actdays         -1.4066     0.5394  -2.608 0.009119 **
prescrib        -1.6273     0.2566  -6.341 2.28e-10 ***
age:freepoor0   -1.1080     0.7340  -1.510 0.131133
age:freepoor1    4.1997     2.2725   1.848 0.064589 .
```

Actual Visits



Predicted Visits



Hurdle Negative Binomial model

hurdle models decompose the prediction process into two components:

- a binary process for distinguishing between zero and non-zero counts
- a truncated count distribution model exclusively for the positive counts

It first decides if a visit happens at all and then predicts how many visits will happen if it does.

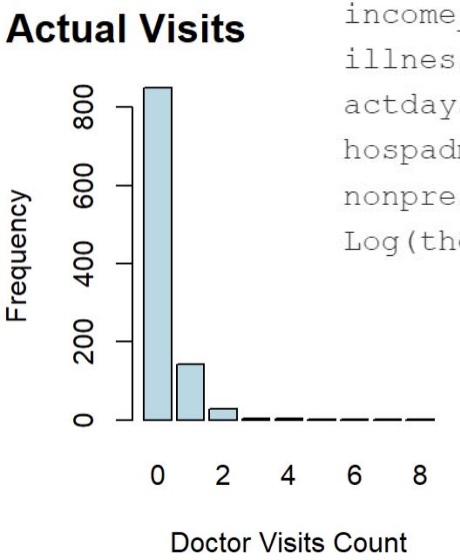
The model uses different factors to predict both the chance of no visits and the expected number of visits, helping us understand what influences these outcomes

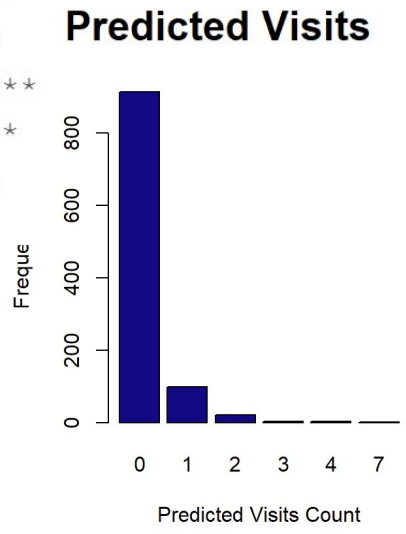
```
hurdle_model <- hurdle(doctorco ~ illness + actdays + hospadmi|income:freepoor + actday
s *illness + sex*hscore + hospadmi + age*prescrib + nonpresc, data = train_data, dist
="negbin")
```

```
Zero hurdle model coefficients (binomial with logit lin
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.85697    0.18019 -15.855  < 2e-16 *
actdays           0.22737    0.02465   9.225  < 2e-16 *
illness           0.24044    0.03548   6.778 1.22e-11 *
sex1              0.20723    0.11168   1.855 0.063525 .
hscore            0.09858    0.03060   3.222 0.001273 *
hospadmi          0.17192    0.07998   2.150 0.031594 *
age               1.14317    0.28541   4.005 6.19e-05 *
prescrib          0.95751    0.10450   9.163  < 2e-16 *
nonpresc         -0.18286    0.06326  -2.891 0.003843 *
income:freepoor0 -0.12541    0.13215  -0.949 0.342616
income:freepoor1 -4.14804    1.18397  -3.504 0.000459 *
actdays:illness  -0.03808    0.00884  -4.307 1.65e-05 *
sex1:hscore      -0.07548    0.03769  -2.003 0.045178 *
age:prescrib     -1.08499    0.17035  -6.369 1.90e-10 *
```

```
hurdle_model <- hurdle(doctorco ~ illness + actdays + hospadmi|income:freepoor + actday
s *illness + sex*hscore + hospadmi + age*prescrib + nonpresc, data = train_data, dist
="negbin")
```

```
Count model coefficients (truncated negbin with log link):
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.04733    0.87710  -2.334  0.01959 *
income_factorMiddle -0.86573    0.27713  -3.124  0.00178 **
income_factorHigh   -0.79819    0.27834  -2.868  0.00413 **
illness              0.13945    0.05537   2.519  0.01178 *
actdays              0.14213    0.01708   8.322  < 2e-16 ***
hospadmi             0.30483    0.10264   2.970  0.00298 **
nonpresc            -0.21266    0.11668  -1.823  0.06837 .
Log(theta)          -1.64343    1.02746  -1.600  0.10971
```



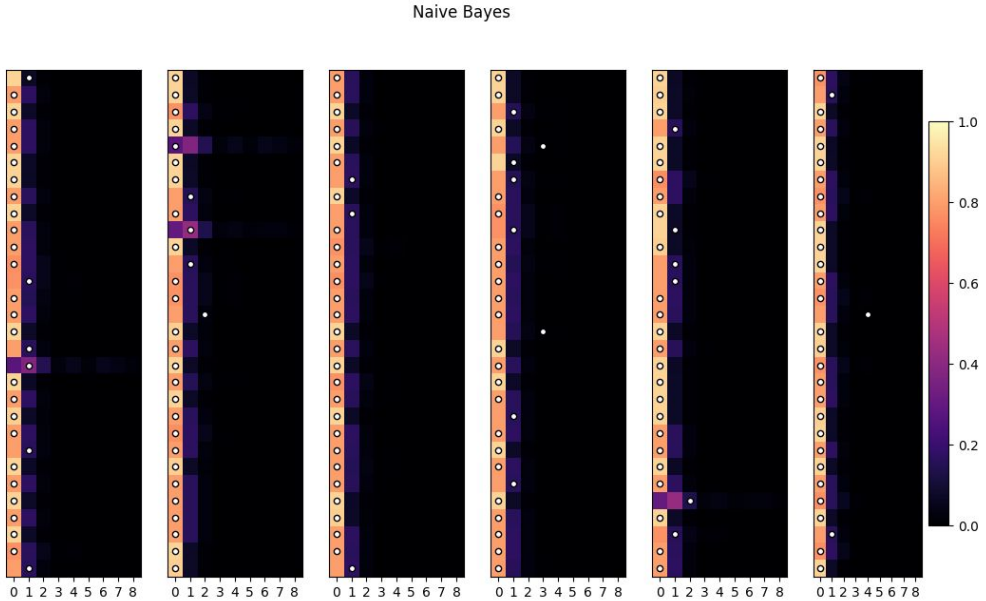


ZINB vs HNB

	ZINB	HNB
AIC	4908.415	4980.611
MAE	0.2649326	0.2736031
RMSE	0.6988843	0.731878
ACCURACY	0.830	0.8237
BALANCED ACCURACY	0.538	0.520
AUC	0.681	0.6355

n.visits	TRUE	ZINB	HNB
0	850	882	915
1	142	134	96
2	29	16	19
3	5	4	4
4	4	1	1
5	2	1	2
6	2	0	1
7	3	0	0
8	1	0	0
9	0	0	0

Naive Bayes



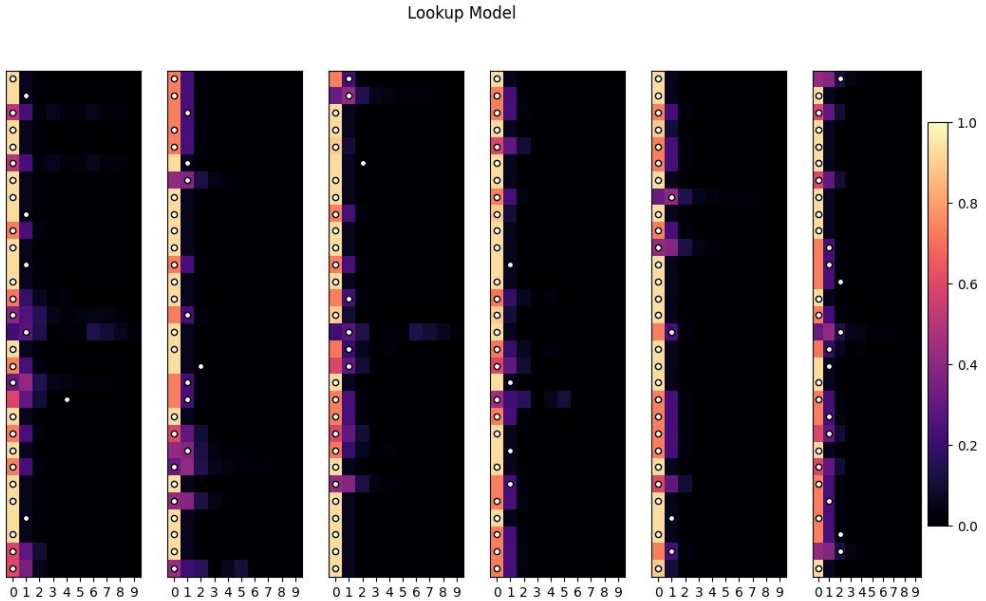
"Gaussian" because this is a normal distribution

This is our prior belief

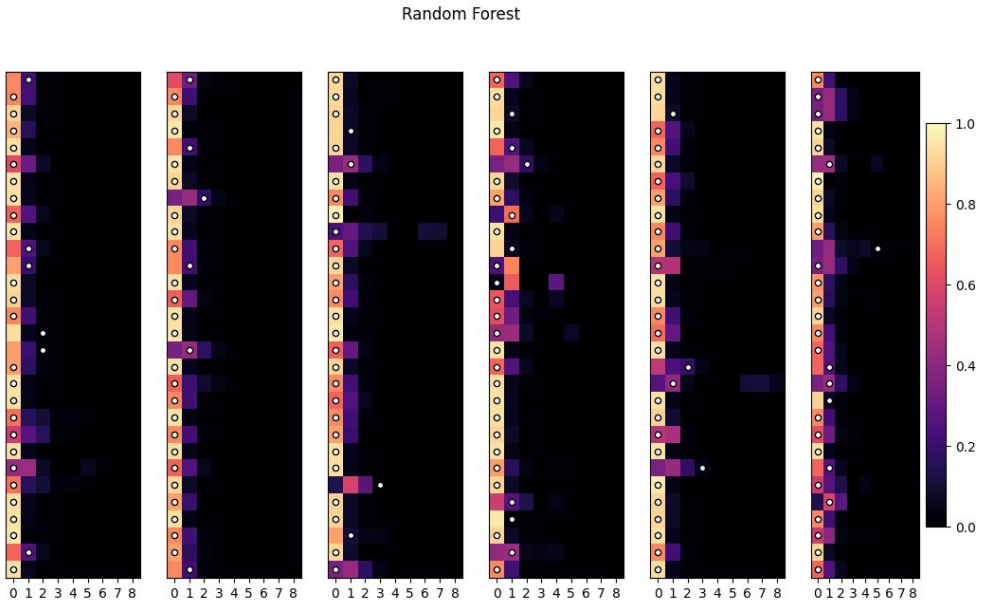
$$P(\text{class} \mid \text{data}) = \frac{P(\text{data} \mid \text{class}) \times P(\text{class})}{P(\text{data})}$$

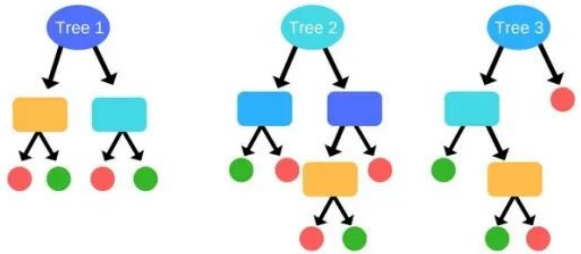
We don't calculate this in naive bayes classifiers

Lookup Model

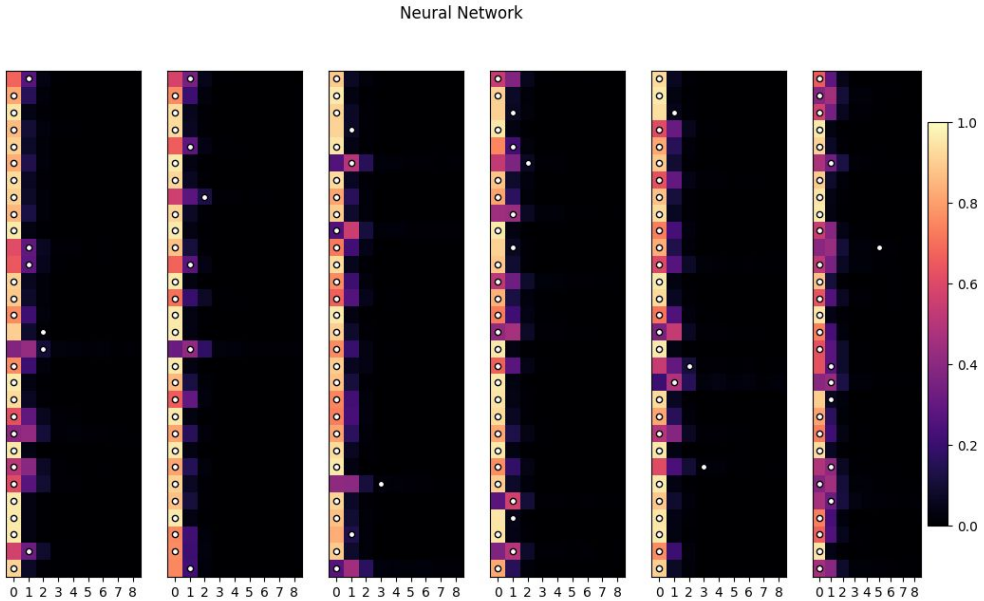


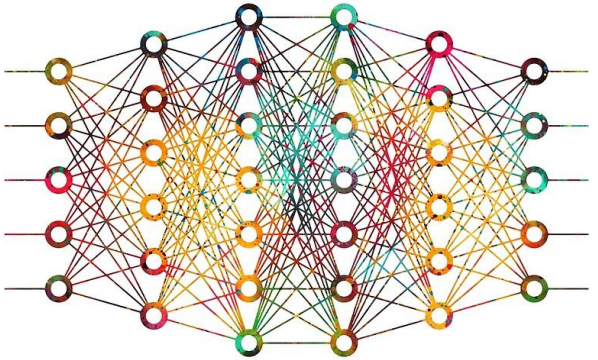
Random Forest



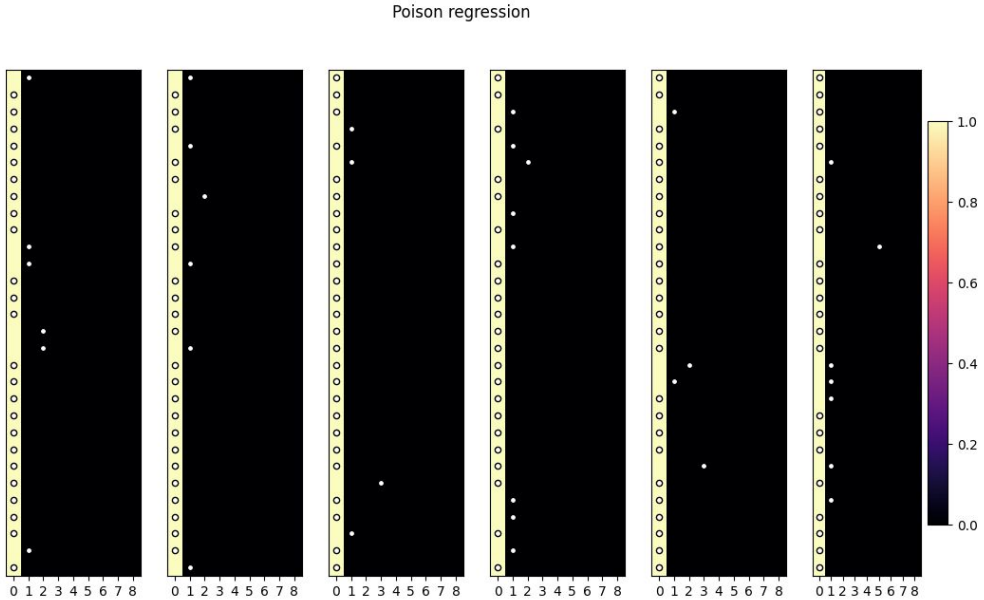


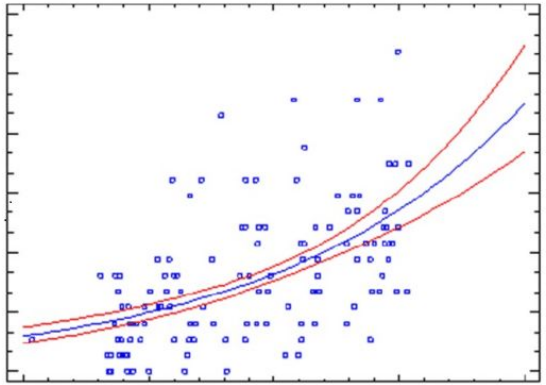
Neural Networks



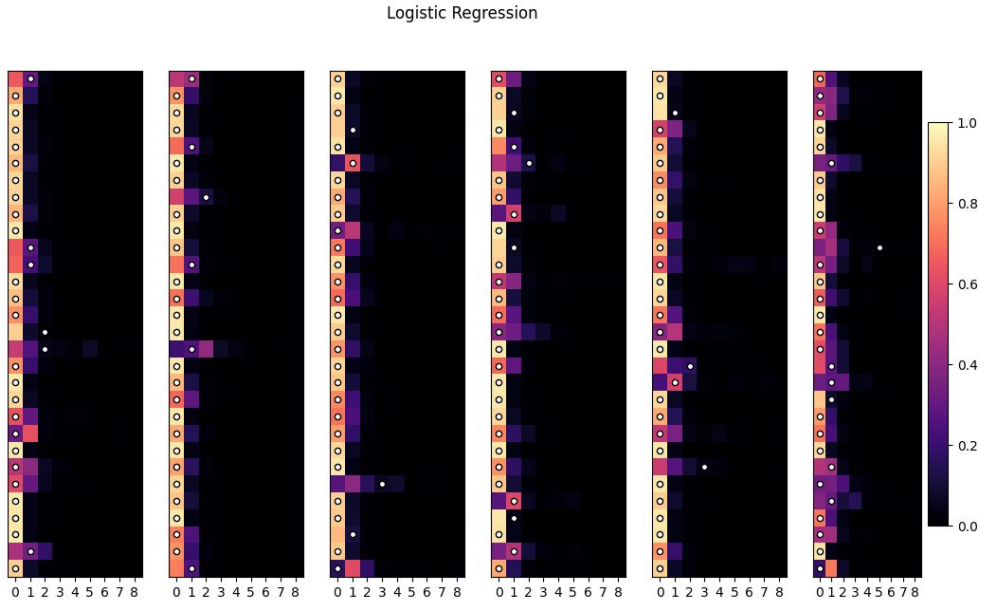


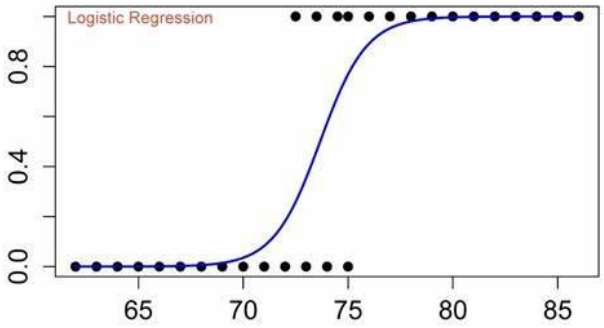
Poisson Regression



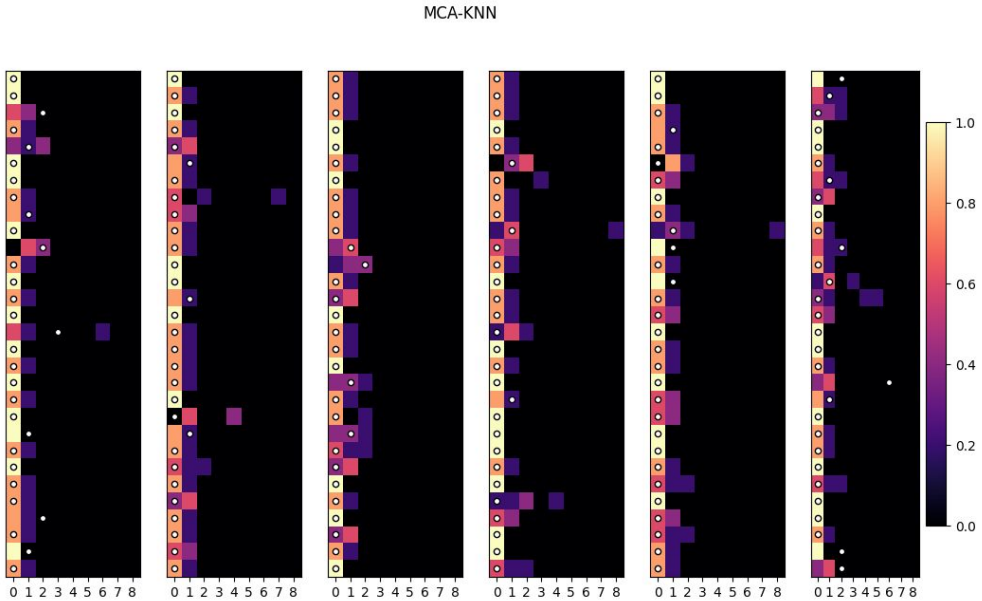


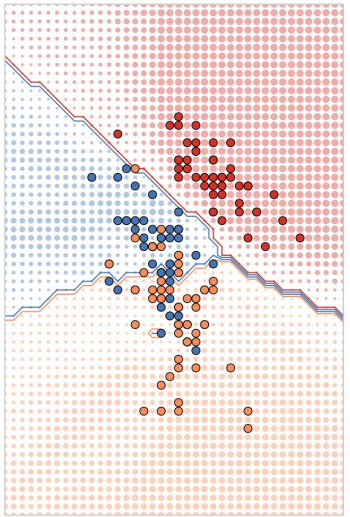
Logistic Regression



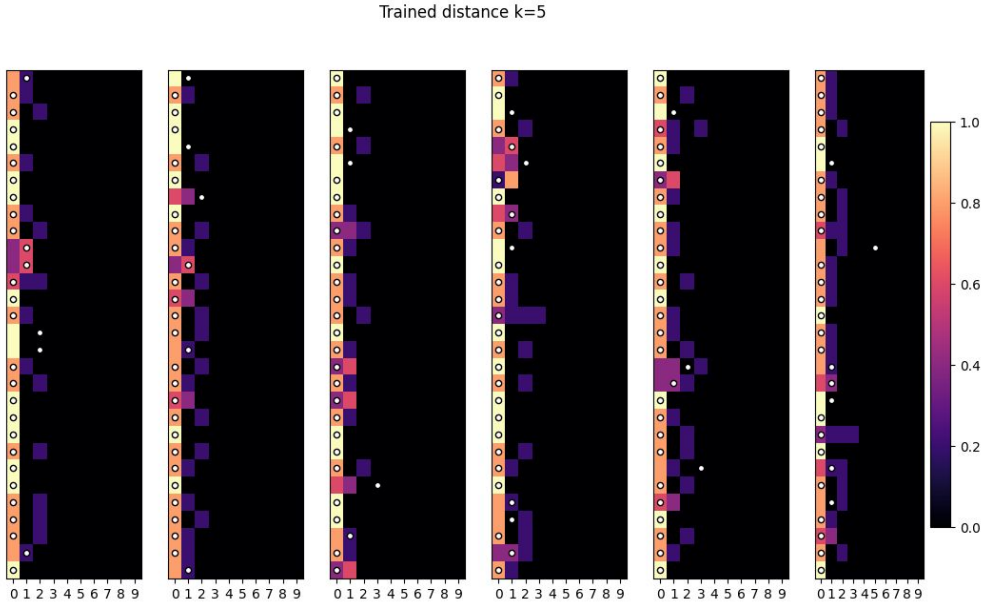


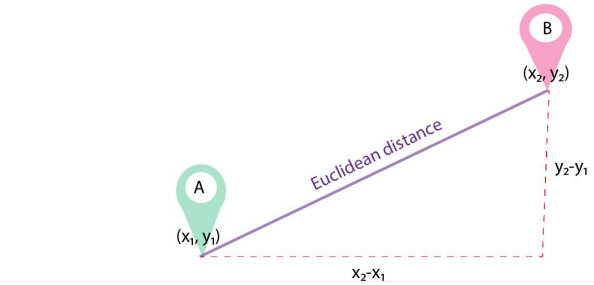
MCA-KNN





Trained distance





DATA SCIENCE & ARTIFICIAL INTELLIGENCE

SCIENTIFIC & DATA-INTENSIVE COMPUTING

Thanks!

Buscema Andrea, Cusma Fait Omar, Derin Tanja,
Špringer Christian, Živanović Uroš

Do you have any questions?

... AND, HOW MANY KANGOROOOSS WERE IN THE PRESENTATION??

Conclusion