

Dementia Disease: Nonparametric Models for prediction and Survival Analysis

Francesca Di Filippo, Erica Manfrin,
Elena Musiari, Edoardo Palli

Nonparametric Statistics
2021-2022

PROJECT DEVELOPMENT

Introduction to
the dataset

Questions &
first analysis

Outlier
detection

Survival
Analysis

Conclusion

Logistic
regression

Robust
regression

Conformal
classification

DATASET



- ID: the identification number for each patient
- M.F: the gender
- Hand: the dominant hand
- Age
- EDUC: the educational level, with values ranging from 1 to 5
- SES: the socio-economic status, with values ranging from 1 to 5
- MMSE: the Mini Mental State Examination, with values ranging from 1 to 30
- CDR: the Clinical Dementia Rating, ranging between 0 and 3
- eTIV: the Estimated Total Intracranial Volume
- nWBV: the Normalize Whole Brain Volume
- ASF: the Atlas Scaling Factor
- Delay: the days passed from the first visit
- Group: the label pointing to the type of patient

DATASET



- ID: the identification number for each patient
- M.F: the gender
- Hand: the dominant hand
- Age
- EDUC: the educational level, with values ranging from 1 to 5
- SES: the socio-economic status, with values ranging from 1 to 5
- MMSE: the Mini Mental State Examination, with values ranging from 1 to 30
- CDR: the Clinical Dementia Rating, ranging between 0 and 3
- eTIV: the Estimated Total Intracranial Volume
- nWBV: the Normalize Whole Brain Volume
- ASF: the Atlas Scaling Factor
- Delay: the days passed from the first visit
- Group: the label pointing to the type of patient

DATASET



- ID: the identification number for each patient
- M.F: the gender
- Hand: the dominant hand
- Age
- EDUC: the educational level, with values ranging from 1 to 5
- SES: the socio-economic status, with values ranging from 1 to 5
- MMSE: the Mini Mental State Examination, with values ranging from 1 to 30
- CDR: the Clinical Dementia Rating, ranging between 0 and 3
- eTIV: the Estimated Total Intracranial Volume
- nWBV: the Normalize Whole Brain Volume
- ASF: the Atlas Scaling Factor
- Delay: the days passed from the first visit
- Group: the label pointing to the type of patient

RESEARCH QUESTIONS

Is there statistical difference between demented and non demented patients?

The Dementia diagnosis can be predicted using non-specific medical tests?

Which type of patients are more likeable to become demented?

EXPLORATORY ANALYSIS

USING ANOVA AND PERMUTATION TESTS

PERMUTATION TEST

Study the statistical difference between the two groups using
Tukey median

$$H_0 : X_{Demented} \stackrel{d}{=} X_{Nondemented} \quad vs \quad H_1 : X_{Demented} \stackrel{d}{\neq} X_{Nondemented}$$

p-value = 0.002



SIGNIFICANCE TESTS



- ANOVA TEST:

Study the importance of M.F variable in the grouping, using
permutational ANOVA

p-value = 0.011

- PERMUTATION TEST:

Study the importance of Age variable in the grouping

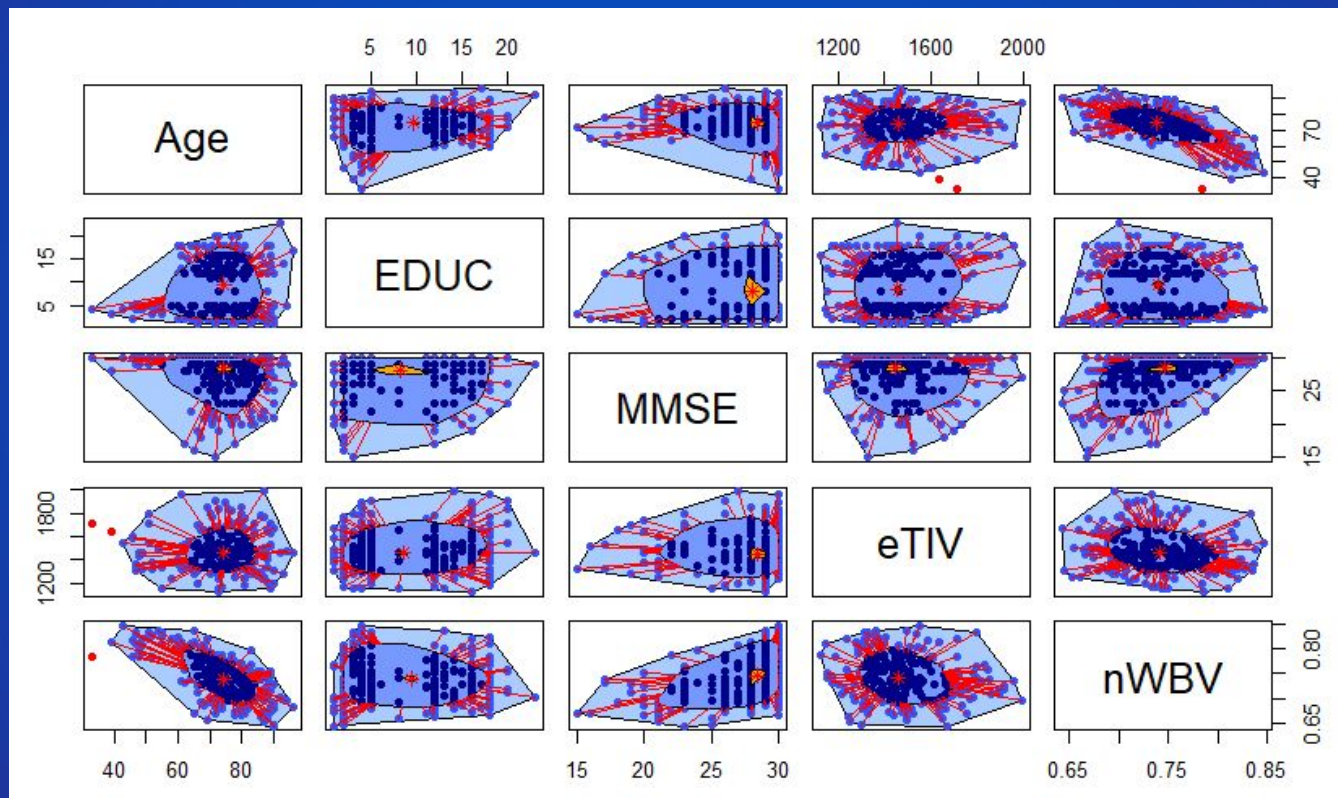
p-value = 0

The background is a solid dark blue color. It is decorated with numerous abstract shapes, including spheres and organic, bubble-like forms. These shapes are rendered in various shades of blue and purple, with some having a gradient or a bright highlight, giving them a three-dimensional appearance. They are scattered across the frame, with a higher concentration on the right side.

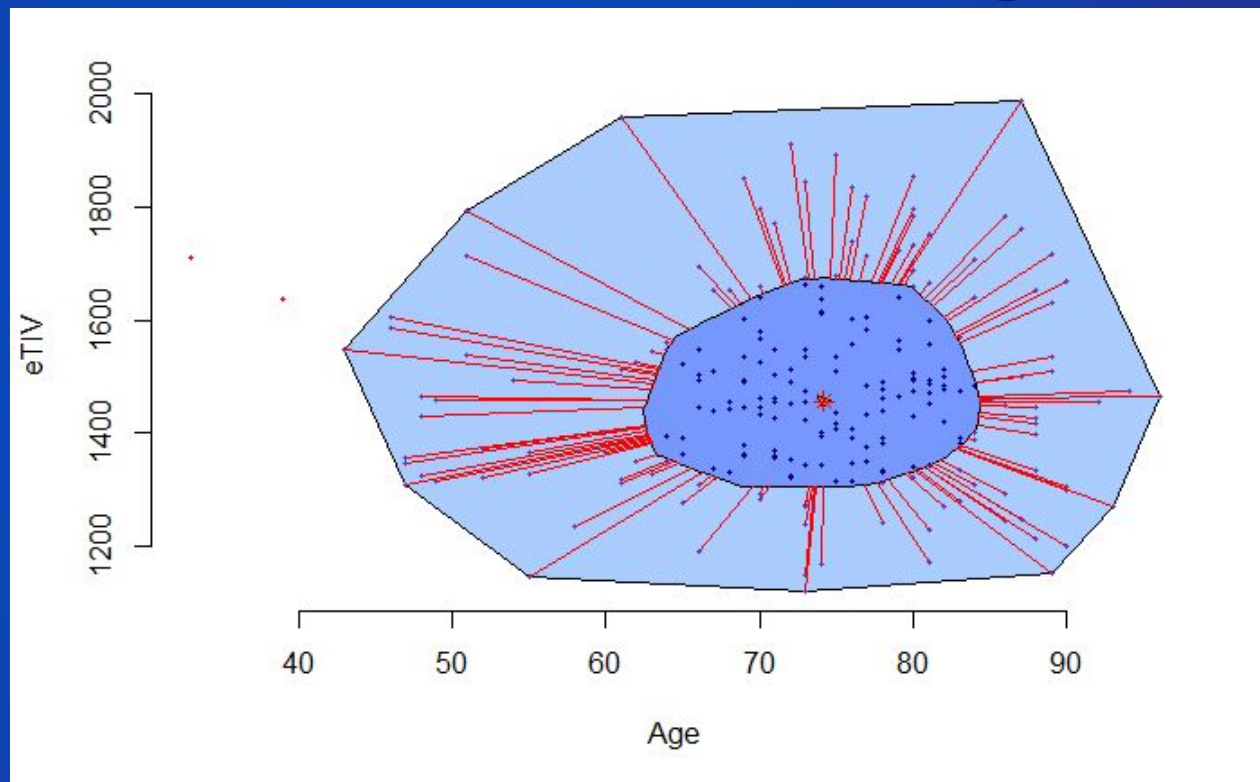
OUTLIER DETECTION

USING DEPTH MEASURES

BAGPLOT MATRIX TRAINING DATASET

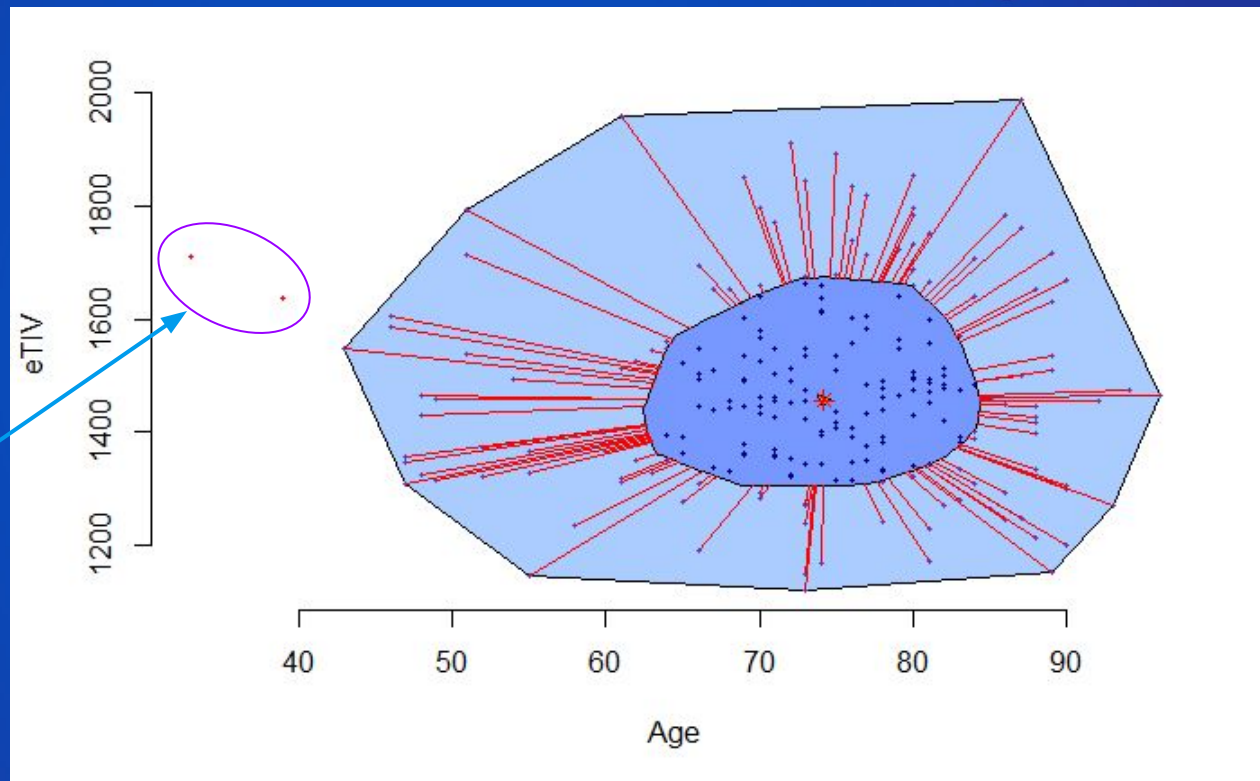


BAGPLOT MATRIX TRAINING DATASET

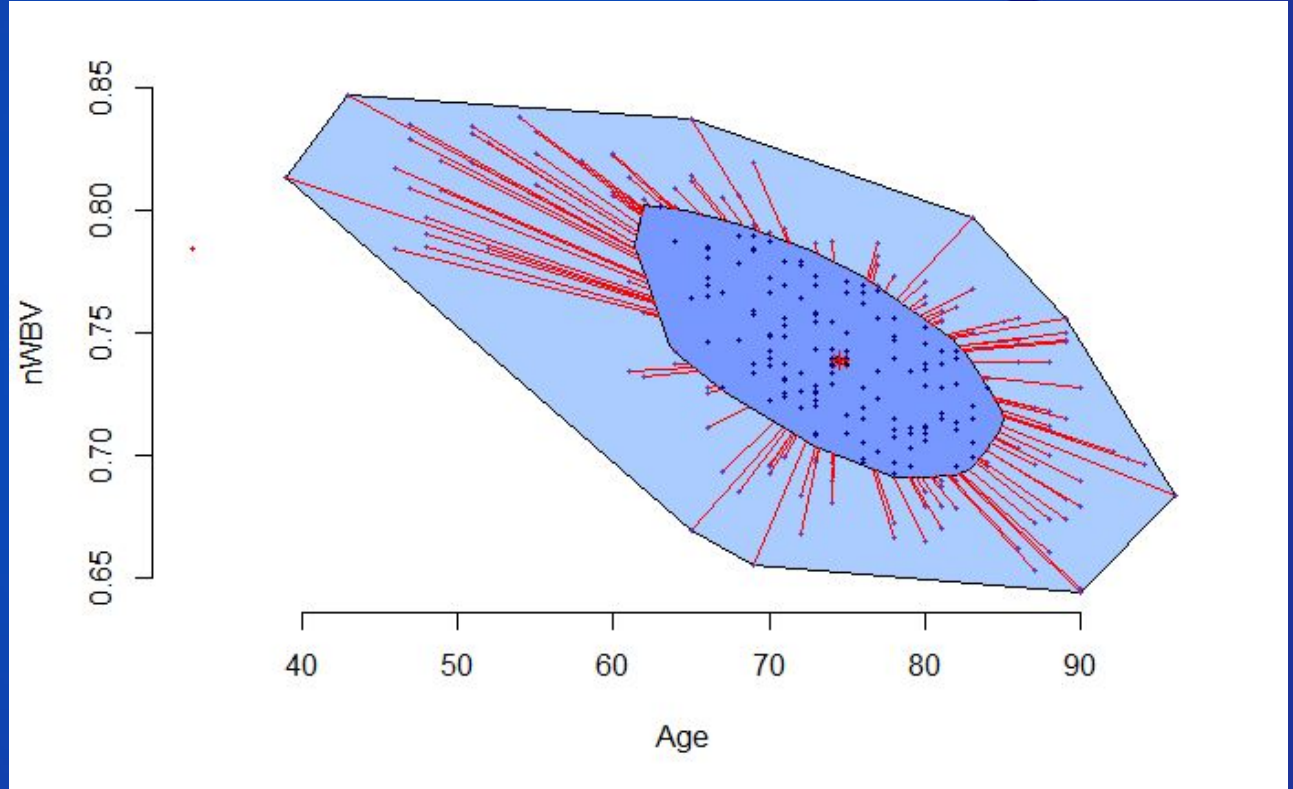


BAGPLOT MATRIX TRAINING DATASET

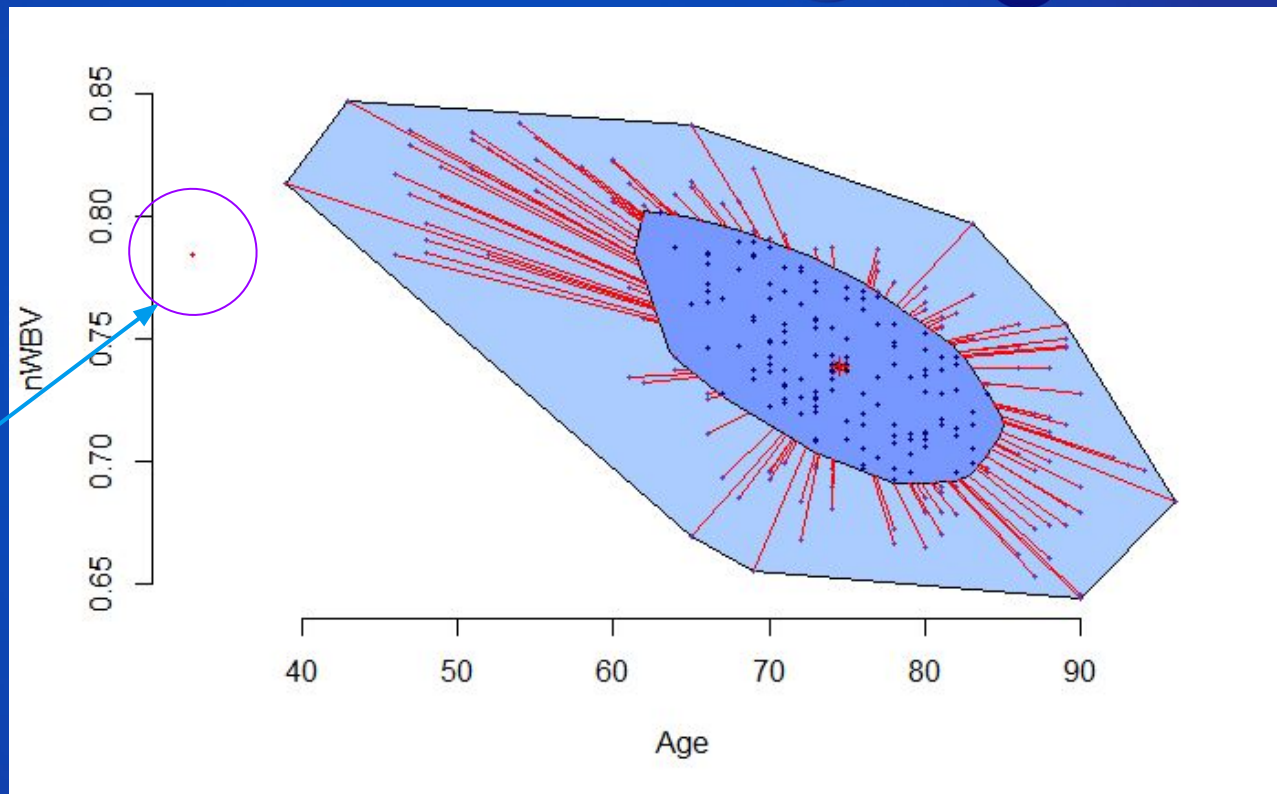
patients at lines
145 and 175



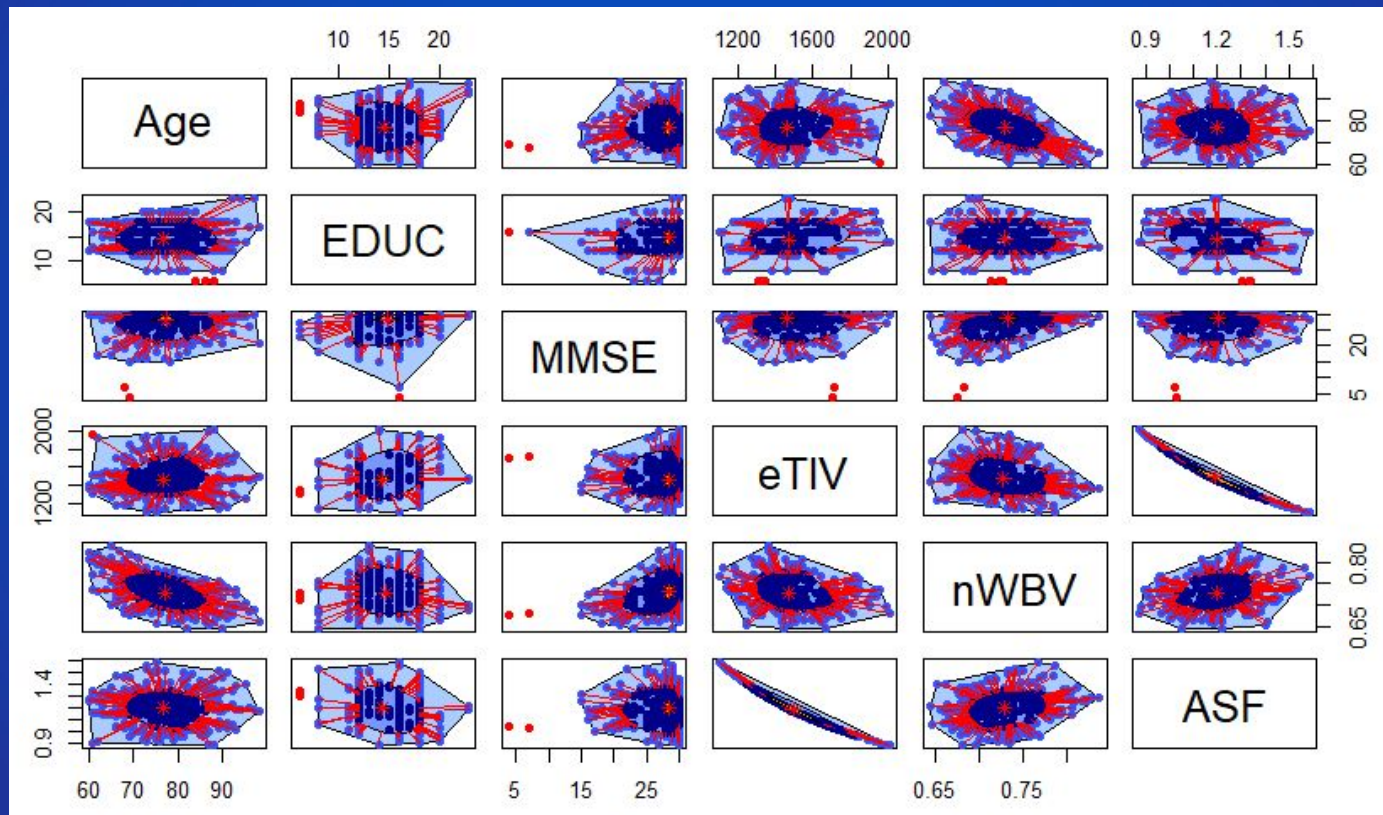
BAGPLOT MATRIX TRAINING DATASET



BAGPLOT MATRIX TRAINING DATASET



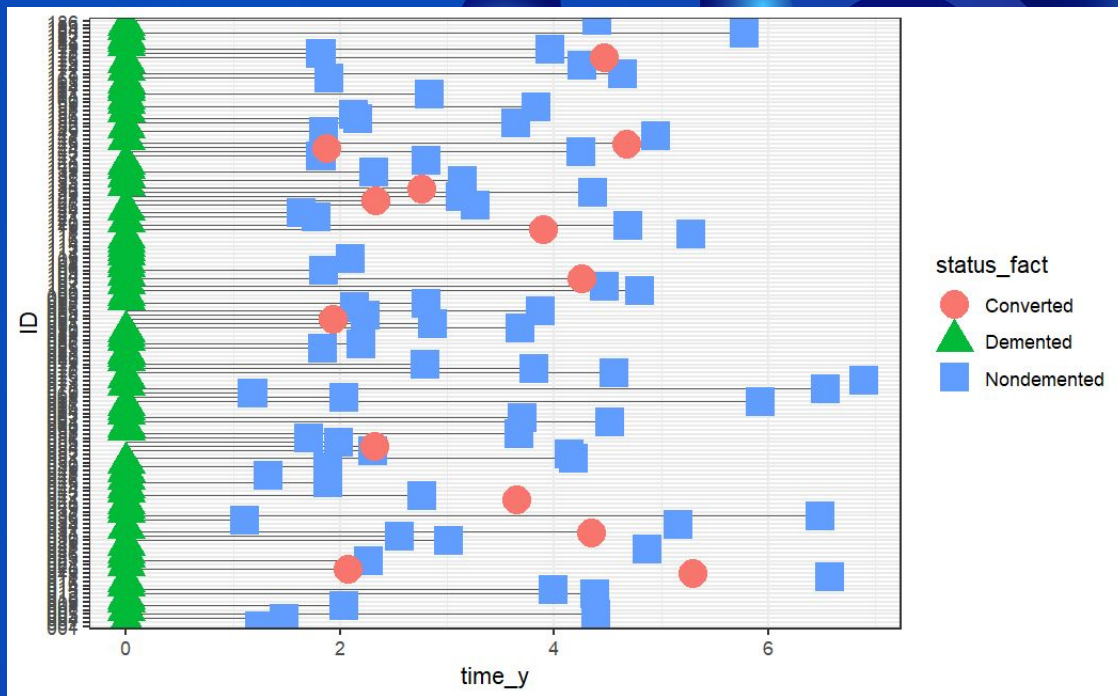
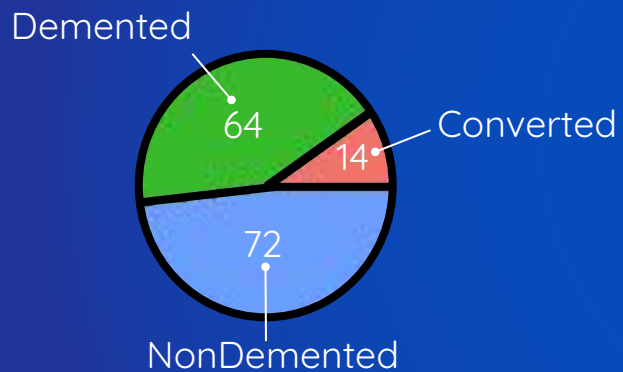
BAGPLOT MATRIX LONGITUDINAL DATASET



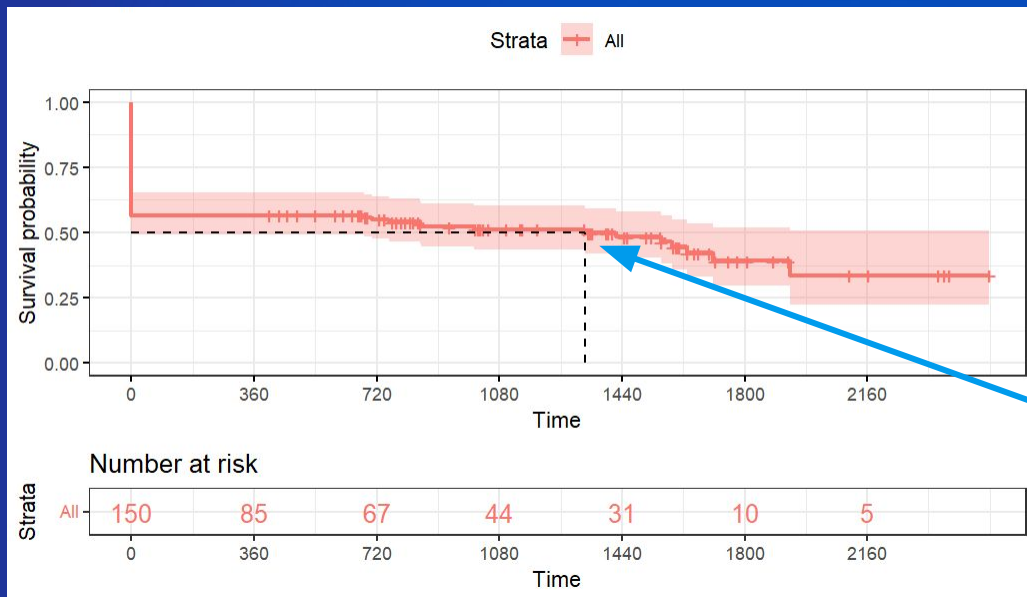
SURVIVAL ANALYSIS

The background is a solid dark blue. It is decorated with numerous abstract elements: spheres of varying sizes in shades of blue and purple, and larger, more complex organic shapes that resemble molecular structures or fluid droplets. These elements are scattered across the frame, with a higher concentration on the right side.

TIME TO EVENT DATASET



KAPLAN MEIER ESTIMATOR OF SURVIVAL PROBABILITY



$$S(t) = P(T < t)$$

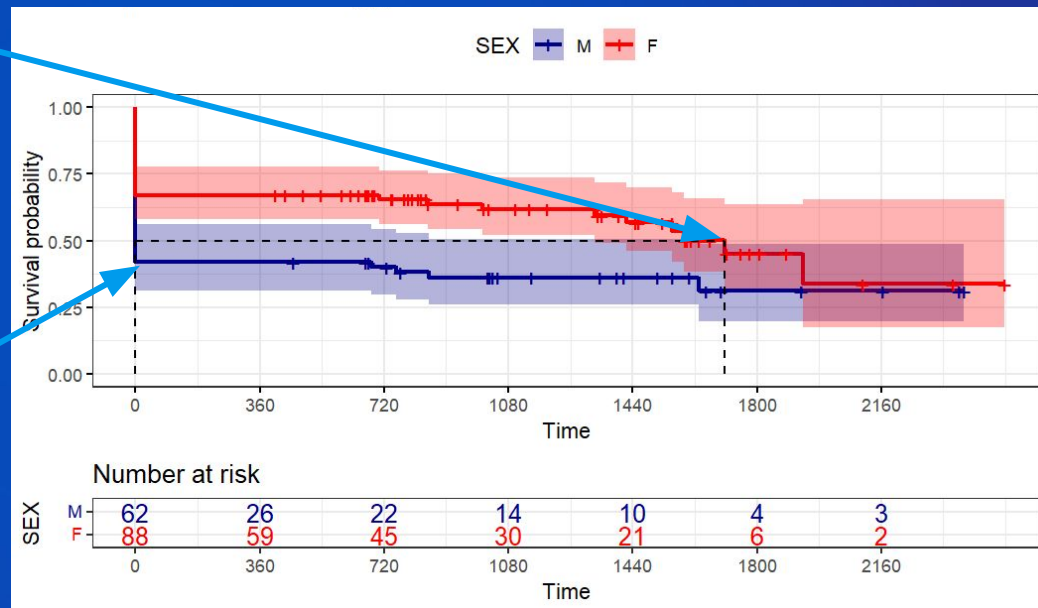
median survival time:
3 and a half years

KAPLAN MEIER ESTIMATOR WITH GENDER CLASSIFICATION

$$S(t) \sim \text{Gender}$$

Female median survival time:
almost 5 years

Male median survival time:
at the first visit the Survival probability
is already below 50%



LONG RANK TEST

$$H_0 : S_{male}(\cdot) = S_{female}(\cdot) \text{ vs } H_1 : S_{male}(\cdot) \neq S_{female}(\cdot)$$

p-value=0.01

$$HR_{MF} = \frac{O_{Male}/E_{Male}}{O_{Female}/E_{Female}} = 1,57$$

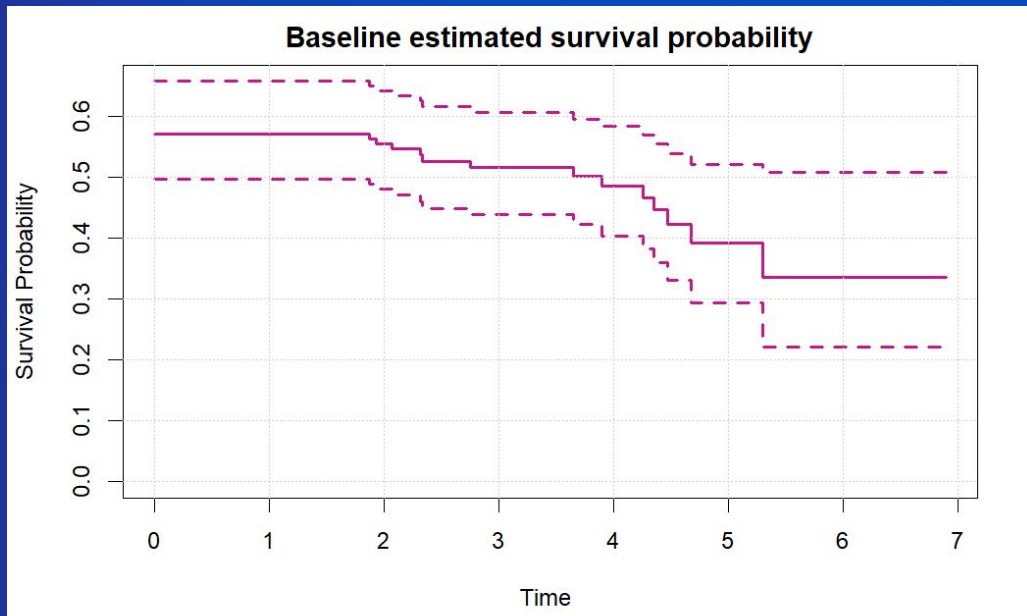


Male
risk factor



COX MODEL WITH VARIABLE AGE

$$h(t) = h_0(t) \exp(\beta_{Age} Age)$$



AGE

coefficient = -0.0377

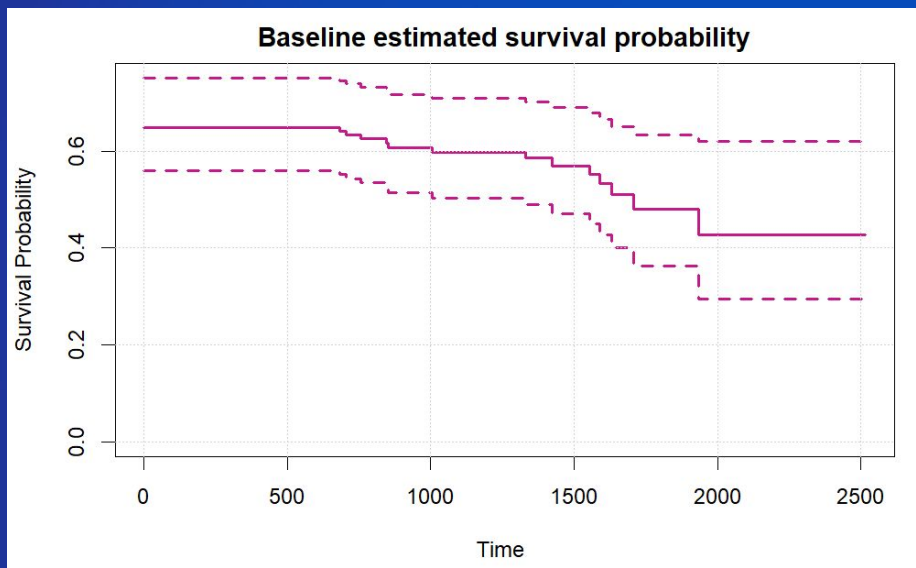
Wald test p-value = 0.0116

Hazard ratio and CI at 95%

<i>lw</i>	<i>fit</i>	<i>up</i>
0.9351	0.9629	0.9916

COX MODEL WITH AGE AND GENDER

$$h(t) = h_0(t) \exp(\beta_{Age}Age + \beta_{Gender}Gender)$$



GLOBAL TEST:

likelihood ratio test, Wald test, Score (logrank) test
p-values = 0.002

AGE

coefficient = -0.03664
Wald test p-value = 0.0155
Hazard ratio and CI at 95%

<i>lw</i>	<i>fit</i>	<i>up</i>
0.9358	0.9640	0.9931

GENDER

coefficient = -0.58
Wald test p-value = 0.0106
Hazard ratio and CI at 95%

<i>lw</i>	<i>fit</i>	<i>up</i>
0.3583	0.5595	0.8737

CONFORMAL CLASSIFICATION

INDUCTIVE CONFORMAL PREDICTION (ICP)

TRANSDUCTIVE CONFORMITY PREDICTION (TCP)

RANDOM
FOREST

CONFORMITY SCORE: $\alpha_i(y) = \frac{\# \text{ trees voting for class } y}{\# \text{ of trees}}$

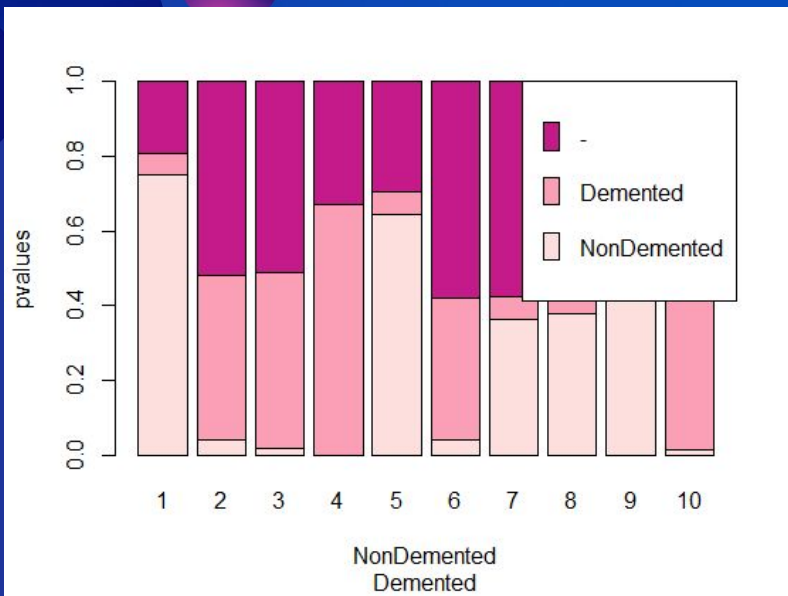


INDUCTIVE CONFORMAL PREDICTION (ICP)

- Proper Training set: $Z_p = z_1, \dots, z_q$ of size $q=216$ (80% of Training Set)
- Calibration set: $Z_c = z_{q+1}, \dots, z_n$ of size $n-q= 271- 216$
- Test set: $Z_t = z_1, \dots, z_m$ of size $m=100$

ICP- PVALUES

$$p_y = \frac{|z_i \in Z_c: y_i = y, \alpha_i(y) < \alpha_{new}(y)| + u_i * |z_i \in Z_c: y_i = y, \alpha_i(y) = \alpha_{new}(y)|}{n_y + 1}$$



$$u_i \sim U[0, 1]$$

EVALUATION OF PERFORMANCES

ERROR RATE: $\frac{1}{m} \sum_{i=1}^m I_{y_i \notin \Gamma_i^\epsilon} = 0.1$

EFFICIENCY: $\frac{1}{m} \sum_{i=1}^m I_{|\Gamma_i^\epsilon| > 1} = 0.5$

OBSERVED
FUZZYNESS: $\frac{1}{m} \sum_{i=1}^m \sum_{y_i \neq y} p_i^y = 0.155$

$$\epsilon = 0.05$$

The background is a solid dark blue. It is decorated with numerous abstract shapes, including circles and organic, bubble-like forms. These shapes are rendered in various shades of blue and purple, with some having a gradient or a bright highlight, giving them a three-dimensional appearance. They are scattered across the frame, with a higher concentration on the right side.

LOGISTIC REGRESSION

FOR ALZHEIMER PREDICTION

TWO DIFFERENT APPROACHES

MMSE as unique covariate

GAM without MMSE

very informative for disease classification
we want to exploit all the information in it

Develop a model to be used in addition to
MMSE test

- Global Polynomial Regression
- Local Likelihood Regression
 - Regression Splines

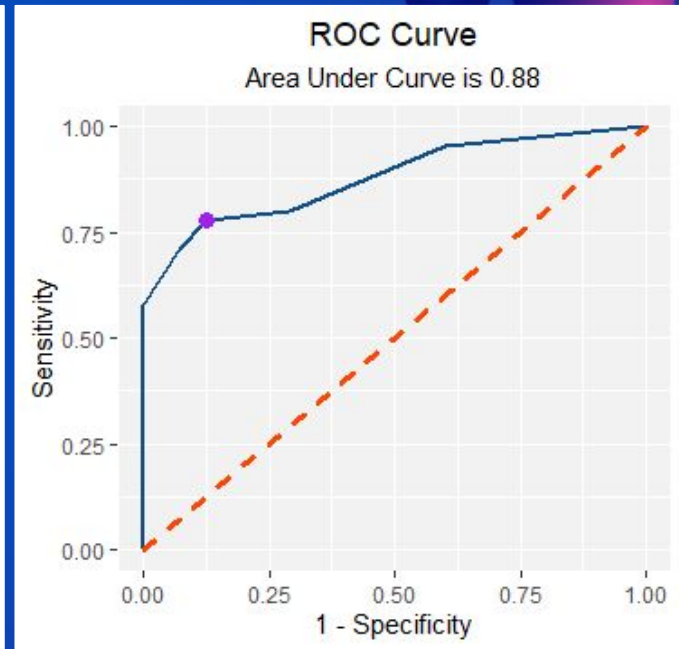
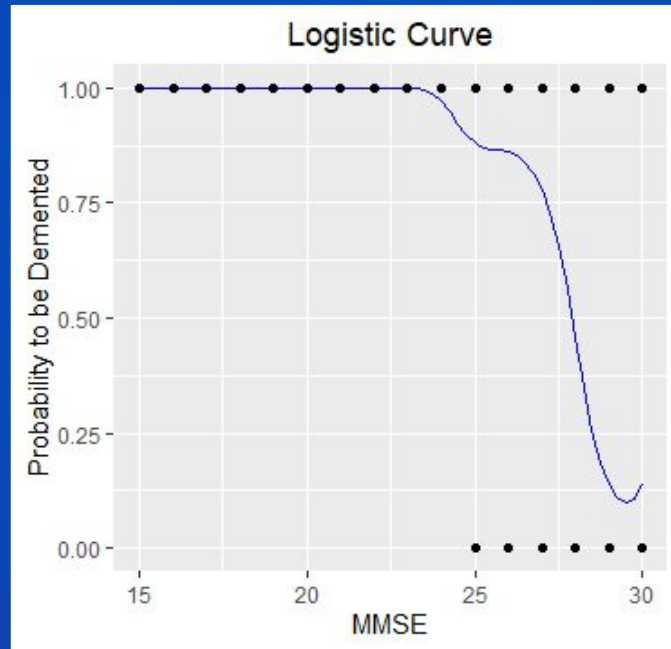
Generalized Additive Model

GLOBAL POLYNOMIALS

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 * MMSE + \beta_2 * MMSE^2 + \beta_3 * MMSE^3 + \beta_4 * MMSE^4$$

- $R^2 = 0.4463$
- Sensitivity = 0.78
- Specificity = 0.87
- accuracy = 0.83
- precision = 0.83
- F1 score = 0.80

No monotonic
behaviour

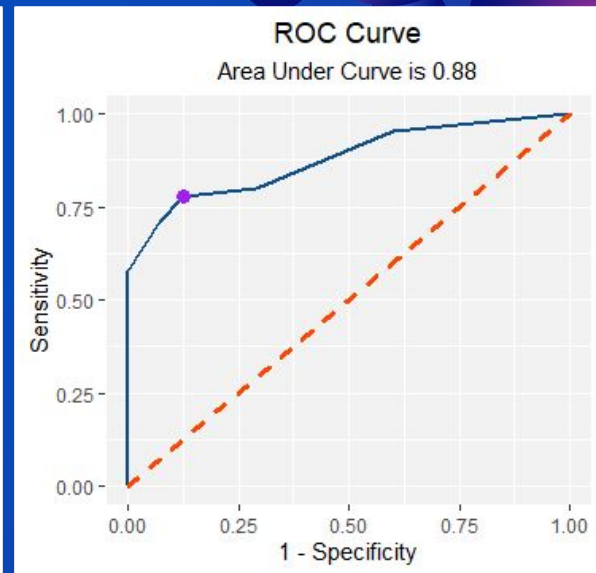
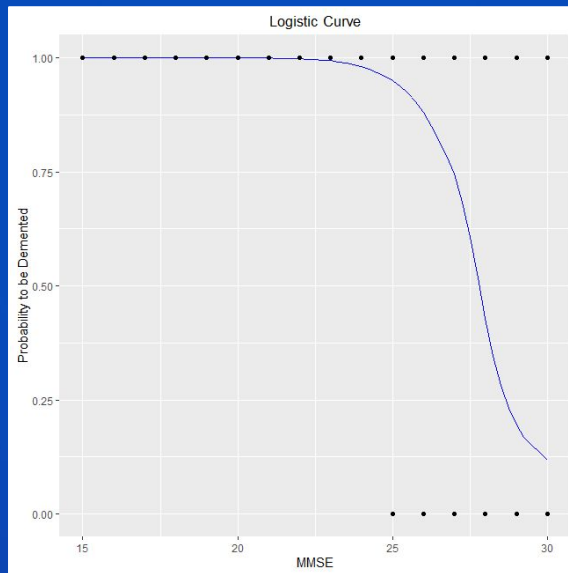


LOCAL LIKELIHOOD

$$\ell_{x,h}(\boldsymbol{\beta}) := \sum_{i=1}^n \ell(Y_i, \eta(MMSE_i - x)) K_h(x - MMSE_i) \quad \eta(x) := \beta_0 + \beta_1 x.$$

$K_h(x)$: gaussian kernel

- $R^2 = 0.4463$
- Sensitivity = 0.78
- Specificity = 0.87
- accuracy = 0.83
- precision = 0.83
- F1 score = 0.80

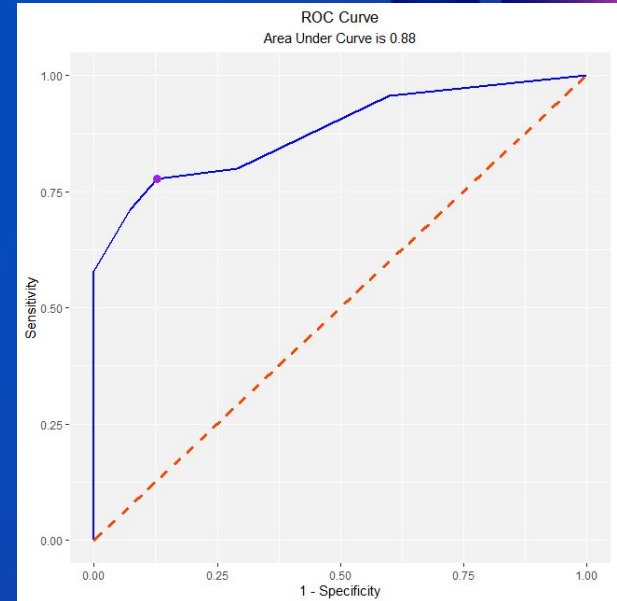
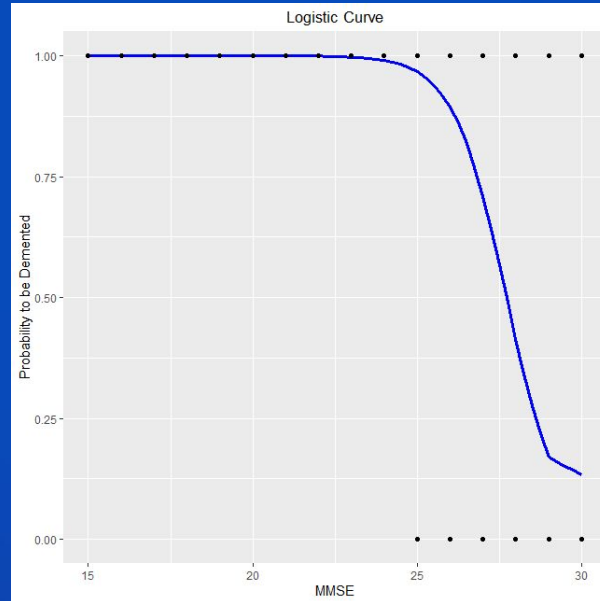


REGRESSION SPLINE

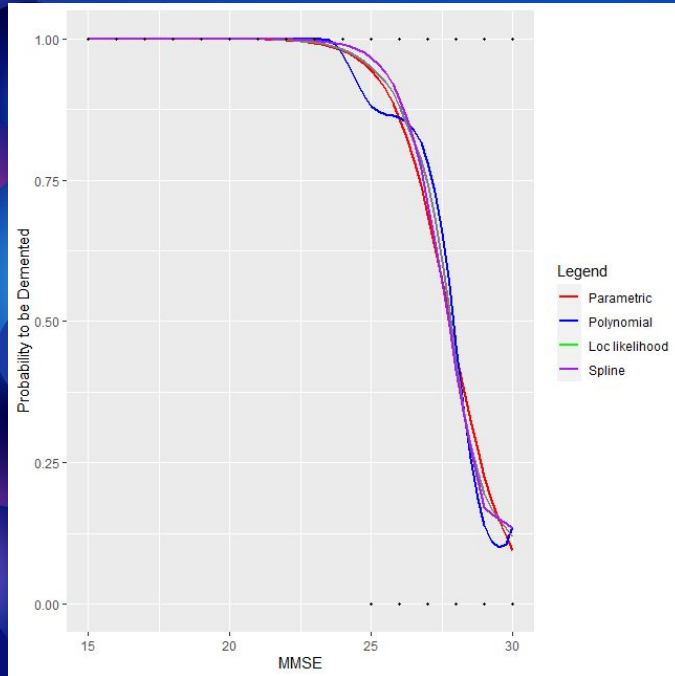
$$\log \frac{p}{1-p} = \beta_0 + \sum_{j=1}^{k+1} \beta_j g_j (MMSE)$$

K = 1 knot at the median

- $R^2 = 0.4377$
- Sensitivity = 0.78
- Specificity = 0.87
- accuracy = 0.83
- precision = 0.83
- F1 score = 0.80



COMPARISON BETWEEN MODELS

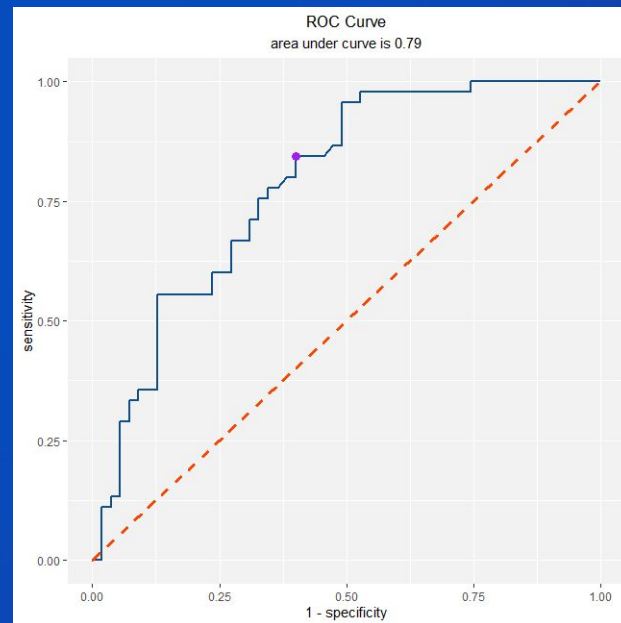


	R^2	AIC	AUC
PARAMETRIC GLM	0.429	216.0	0.879
GLOBAL POLYNOMIALS	0.446	215.8	0.879
LOCAL LIKELIHOOD	—	—	0.879
CUBIC SPLINE	0.4377	213.03	0.854

GENERALIZED ADDITIVE MODEL

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 \text{Sex} + f_2(\text{EDUC}) + f_3(\text{nWBV}) + f_4(\text{nWBV} * \text{Sex}) + f_5(\text{Age}) + f_6(\text{eTIV}) + f_7(\text{eTIV} * \text{Sex})$$

- $R^2 = 0.4377$
- Sensitivity = 0.84
- Specificity = 0.6
- Accuracy = 0.71
- Precision = 0.63
- F1 score = 0.72



ROBUST REGRESSION

The background is a solid dark blue. It is decorated with numerous overlapping circles and organic, bubble-like shapes in various shades of blue and purple. Some shapes have a gradient, appearing lighter on one side, giving them a three-dimensional feel. The shapes are scattered across the frame, with a higher concentration on the right side.

THEORY ABOUT *GLMROB*

We find the robust estimator for beta from the estimating equations:

$$\sum_{i=1}^n \psi(y_i, \mu_i) = 0$$

It is an M-estimator characterized by the score function:

$$\psi(y, \mu) = \nu(y, \mu) \cdot w(x) \cdot \mu' - a(\beta)$$

with
$$a(\beta) = \frac{1}{n} \sum_{i=1}^n E[\nu(y_i, \mu_i)]$$

For binomial models we have:
$$\nu(y_i, \mu_i) = \psi_c(r_i) \frac{1}{V^{1/2}(\mu_i)} \quad r_i = \frac{y_i - \mu_i}{V^{1/2}(\mu_i)}$$

$\psi_c(r)$ is the Huber function defined by r if $|r| < c$ and by $c \cdot \text{sign}(r)$ if $|r| > c$

THREE MODELS

Robust Regression Model with all the covariates:

- with B-splines
- without B-splines

Robust Regression Model with only MMSE:

- with B-splines
- without B-splines

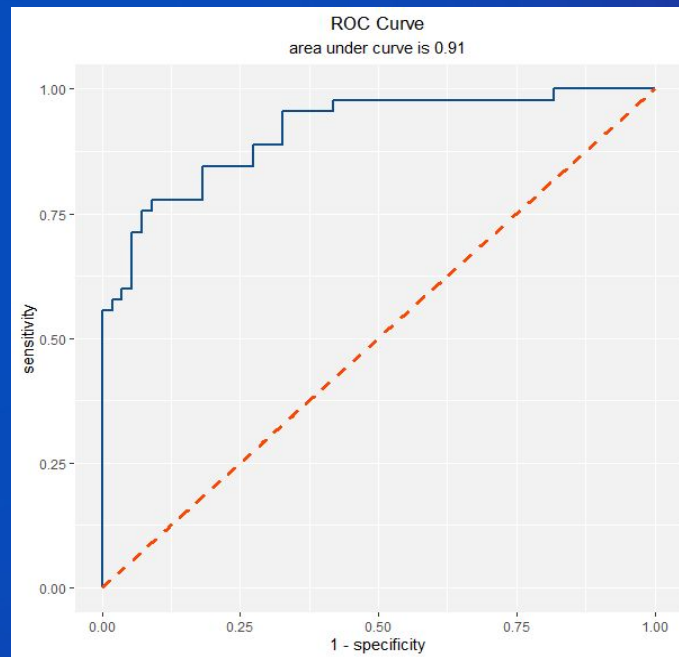
Robust Regression Model without MMSE:

- all the covariates with and without B-splines

COMPLETE MODEL

$$\log\left(\frac{p}{1-p}\right) = \beta_1 \cdot M.F + \beta_2 \cdot EDUC + \beta_3 \cdot nWBV + \beta_4 \cdot Age + \beta_5 \cdot MMSE + \beta_6 \cdot eTIV$$

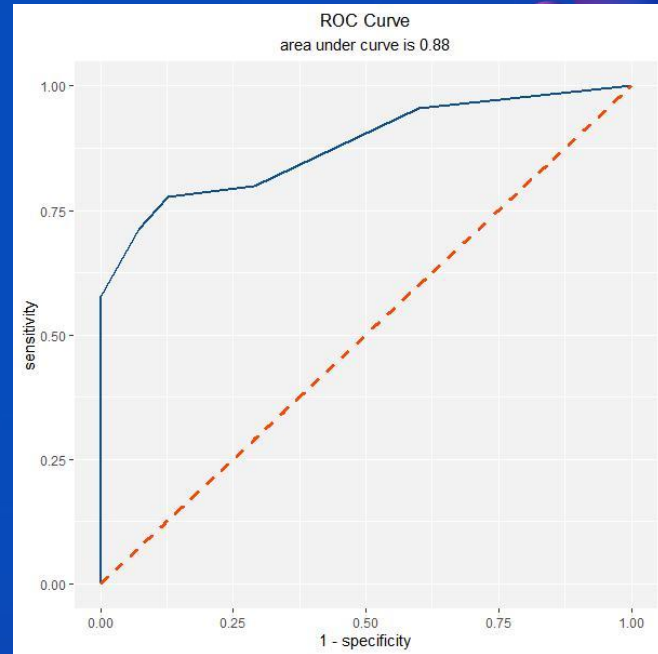
- F1-score=0.7916667
- accuracy= 0.8
- precision=0.745098
- sensitivity=0.875
- specificity=0.733333



MODEL WITH ONLY MMSE

$$\log\left(\frac{p}{1-p}\right) = \beta \cdot MMSE$$

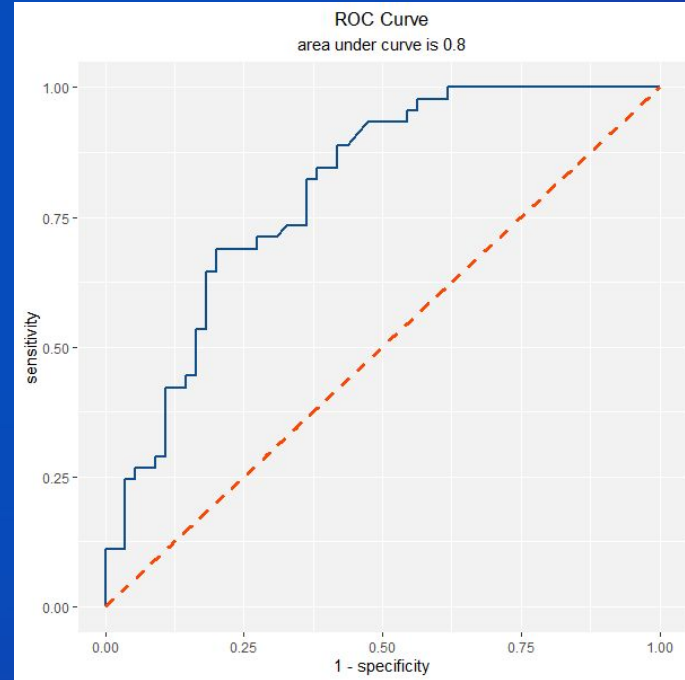
- F1-score=0.80
- accuracy= 0.83
- precision=0.83
- sensitivity=0.78
- specificity=0.87



MODEL WITH ALL PARAMETERS WITHOUT MMSE

$$\log\left(\frac{p}{1-p}\right) = \beta_1 \cdot M + \beta_2 \cdot bs(Age) + \beta_3 \cdot bs(nWBV) + \beta_4 \cdot bs(eTIV)$$

- F1-score=0.73
- accuracy= 0.71
- precision=0.63
- sensitivity=0.73
- specificity=0.71



The background is a deep blue gradient, populated with numerous overlapping circles and organic, fluid shapes in various shades of blue and purple. These shapes vary in size, creating a sense of depth and movement. The overall aesthetic is modern and digital.

CONCLUSIONS

We can say that there are significant differences between demented and nondemented

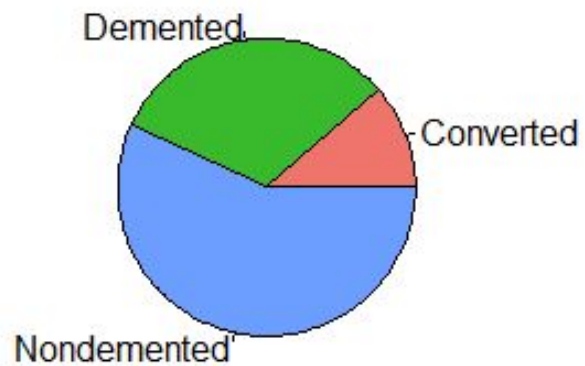
We can predict Dementia with the use of information that are not the Mini Mental State Examination, and so with less medical information

For our analysis the risk to contract dementia is higher in men and it decreases when people get older

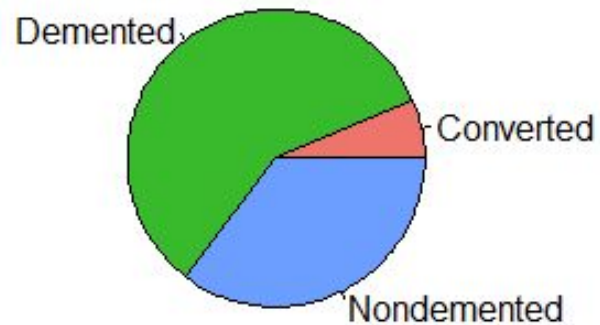
The background is a solid dark blue. Overlaid on this are numerous spheres of varying sizes. Some spheres are a deep blue, while others have a gradient of blue and purple. The spheres are arranged in a way that some appear to be connected by thin, translucent lines, creating a molecular or network-like structure. In the center of the image, there is a horizontal rectangular box with rounded corners and a thin white border. Inside this box, the words "THANK YOU" are written in a white, sans-serif, all-caps font.

THANK YOU

Female pie plot

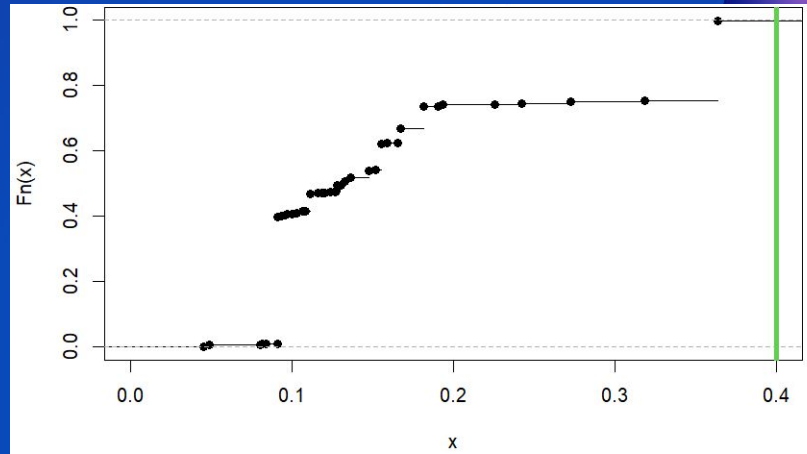


Male pie plot

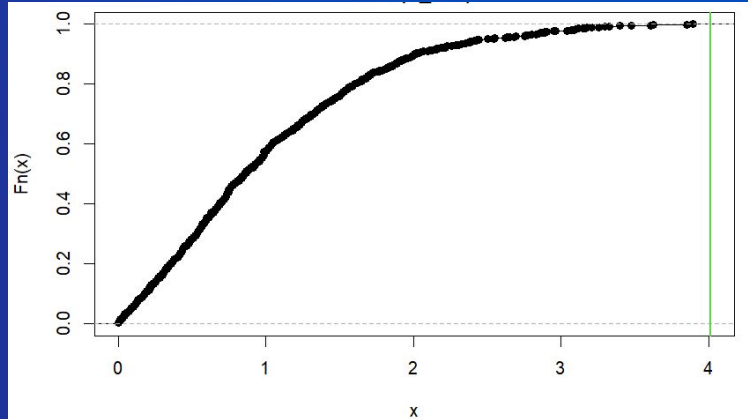


PLOTS TESTS

Permutation test on
grouping



Permutation test on
Age importance



Permutational ANOVA
test on Sex importance

