Spring 2019

Fuzzy data analysis

Practical assignment with Bank dataset

Joni Kettunen

# Introduction

# Data wrangling

# Classification models

* Yleistä
* Ongelman esittely
* Baseline

## Fuzzy K-nn

**Using the whole dataset**

Asd

**Pre-processing data with PCA**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Training set | | | Test set | | |
| Accuracy | Sensitivity | Specificity | Accuracy | Sensitivity | Specificity |
| BASELINE |  |  |  | 0.9419 | 0 | 1 |
| PCA &  Fuzzy K-NN | 0.933 | 0.335 | 0.970 | 0.750 | 0.202 | 0.937 |

K-nn neighbours included: 1

Num of columns from data: 32

**Pre-processing data with FPCA**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training set | | | Test set | | | |
| Accuracy | Sensitivity | Specificity | | Accuracy | Sensitivity | Specificity |
| BASELINE |  |  |  | | 0.9419 | 0 | 1 |
| Fuzzy PCA &  Fuzzy K-NN | 0.933 | 0.318 | 0.971 | | 0.718 | 0.253 | 0.876 |

K-nn neighbours included: 1

Num of columns from data: 26

## Normal K-nn

**Using the whole dataset**

Asd

**Pre-processing data with PCA**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Training set | | | Test set | | |
| Accuracy | Sensitivity | Specificity | Accuracy | Sensitivity | Specificity |
| BASELINE |  |  |  | 0.9419 | 0 | 1 |
| PCA & K-NN | 0.937 | 0.333 | 0.971 | 0.750 | 0.213 | 0.932 |

K-nn neighbours included: 1

Num of columns from data: 28

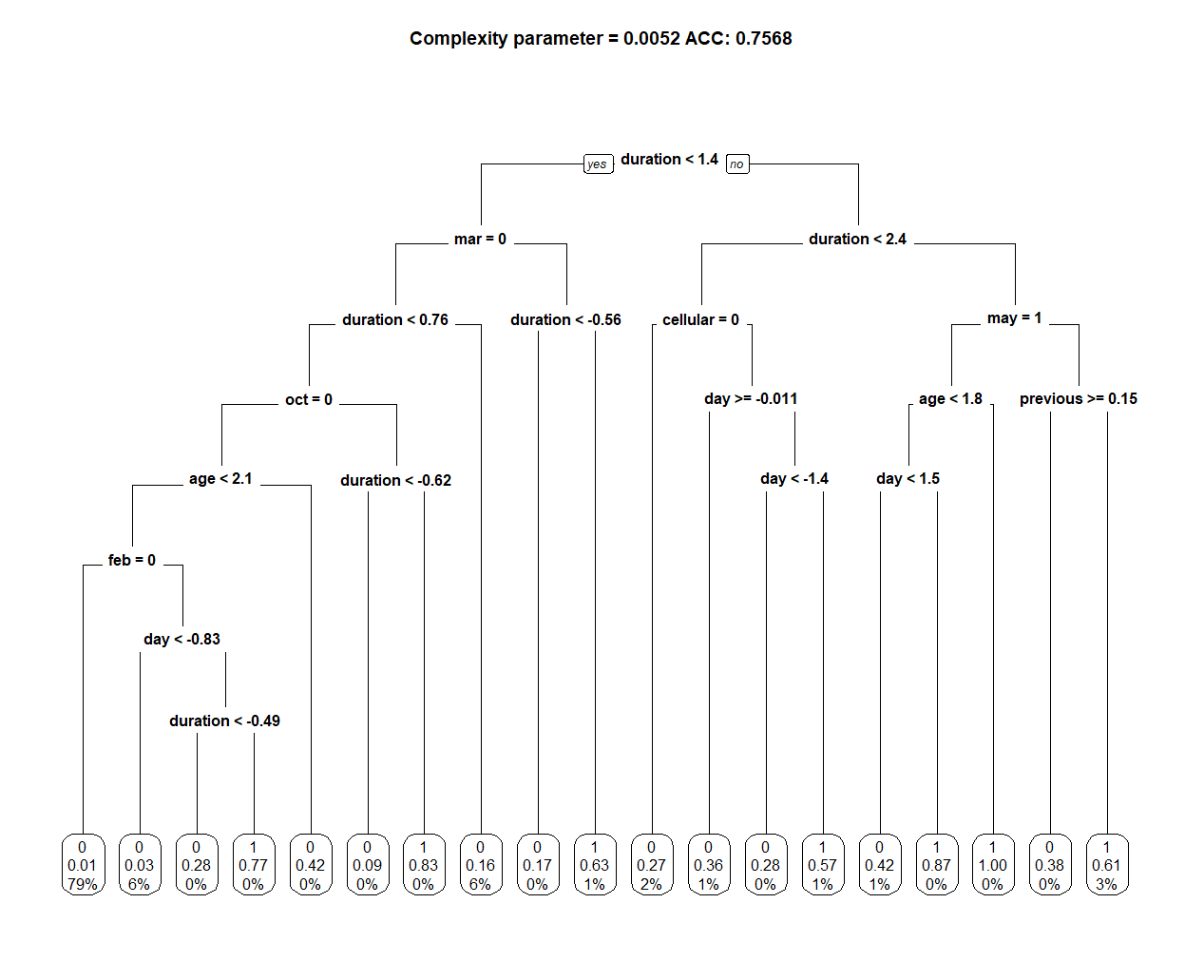
**Pre-processing data with FPCA**

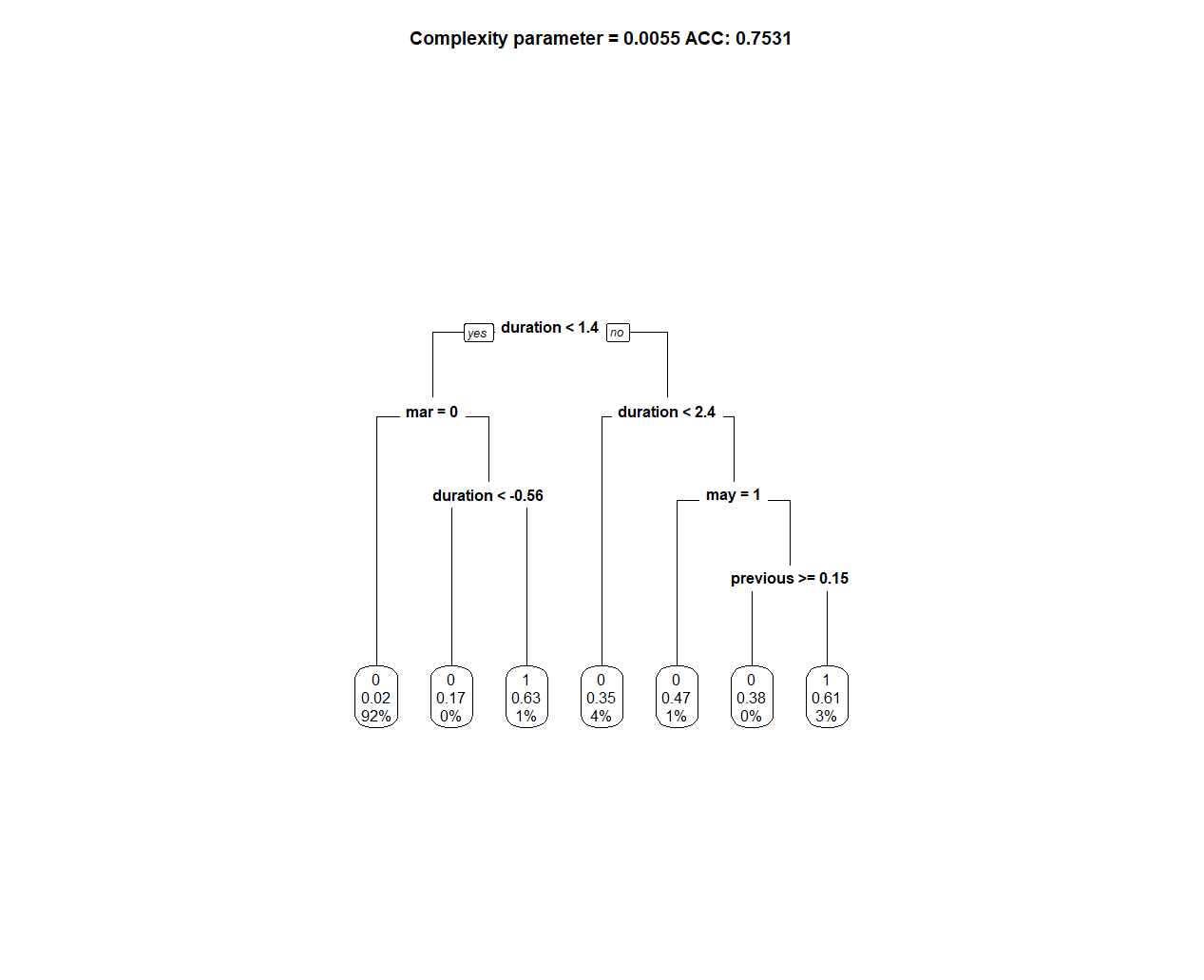
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Training set | | | Test set | | |
| Accuracy | Sensitivity | Specificity | Accuracy | Sensitivity | Specificity |
| BASELINE |  |  |  | 0.9419 | 0 | 1 |
| Fuzzy PCA & K-NN | 0.934 | 0.323 | 0.971 | 0.724 | 0.251 | 0.885 |

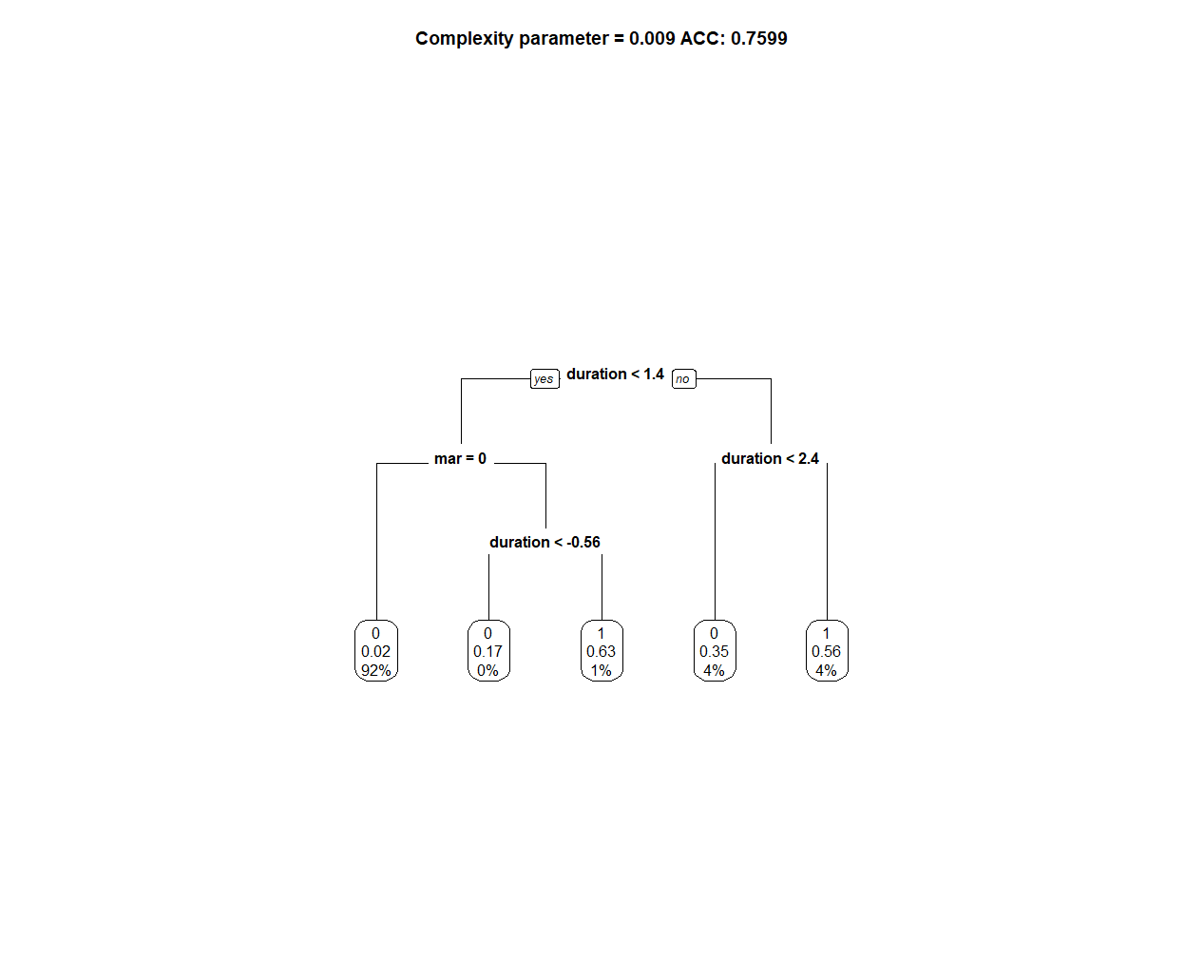
K-nn neighbours included: 1

Num of columns from data: 30

## Decision tree







## Other methods and summary results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training set | | | Test set | | | |
|  | Accuracy | Sensitivity | Specificity | | Accuracy | Sensitivity | Specificity |
| BASELINE |  |  |  | | 0.9419 | 0 | 1 |
| Fuzzy PCA &  Fuzzy K-NN | 0.933 | 0.318 | 0.971 | | 0.718 | 0.253 | 0.876 |
| PCA &  Fuzzy K-NN | 0.933 | 0.335 | 0.970 | | 0.750 | 0.202 | 0.937 |
| Fuzzy K-NN |  |  |  | |  |  |  |
| K-NN |  |  |  | |  |  |  |
| PCA & K-NN | 0.937 | 0.333 | 0.971 | | 0.750 | 0.213 | 0.932 |
| Fuzzy PCA & K-NN | 0.934 | 0.323 | 0.971 | | 0.724 | 0.251 | 0.885 |
| Simple linear regression | 0.946158 | 0.158696 | 0.994767 | | 0.750885 | 0.118005 | 0.966683 |
| Simple linear regression PCA data | 0.946158 | 0.158696 | 0.994767 | | 0.254792 | 1.000000 | 0.000692 |
| Simple linear regression FPCA data | 0.946158 | 0.158696 | 0.994767 | | 0.750516 | 0.116556 | 0.966683 |
| Decision tree classification |  |  |  | | 0.7569301 | 0.0919107 | 0.9836876 |
| Kernel SVM | 0.9456206 | 0.1391304 | 0.9954039 | | 0.7483781 | 0.02464482 | 0.99151557 |
| Logistic Regression | 0.9482748 | 0.3065217 | 0.9878892 | | 0.7677676 | 0.2220934 | 0.9538309 |