Data Science

2020/21

Exam 2

February 5th, 2021
Duration: 2 hours

Student ID: __________ Name: ____

Rules:

- **No** consultation, **but** calculator, is allowed.
- Delivery just the **this** sheet, with your identification and answers inside the grid.
- Withdrawals: 1 hour after starting time. Room entries: up to 30 minutes of starting time.
- Each group counts at most 2 and at least 0 points. Each correct answer counts 0.4 points and each wrong one counts -0.2.

Solution

Data Profiling		
	T	**F**
1		X
2	X	
3	X	
4		X
5		X

Data Preparation		
	T	**F**
1		X
2		X
3	X	
4	X	
5		X

Classifiers Evaluation		
	T	**F**
1		X
2	X	
3	X	
4		X
5	X	

Classification		
	T	**F**
1		X
2		X
3		X
4	X	
5		X

Pattern Mining		
	T	**F**
1	X	
2		X
3		X
4	X	
5	X	

Clustering		
	T	**F**
1		X
2	X	
3	X	
4		X
5	X	

Time Series		
	T	**F**
1	X	
2		X
3	X	
4		X
5	X	

SNA		
	T	**F**
1		X
2	X	
3	X	
4	X	
5		X

Anomaly Detection		
	T	**F**
1	X	
2		X
3	X	
4	X	
5		X

Ethics		
	T	**F**
1	X	
2		X
3	X	
4		X
5		X

Data Description

Consider the problem of predicting if some patient will survive, through the use of a dataset with 165 medical records, described by 50 variables. From these the `class` variable has two possible values `survive` (102) and `die` (63).

The tree on the left was learned through the C4.5 algorithm and the information gain criteria, when applied over 100 of the 165 records available, to learn the target variable `Class`, after applying some preparation techniques.

The tree was printed through `sklearn.tree` package. Each node in the tree shows the variable tested, the number of records satisfying the branch conditions, the number of records from `survive` and `die` classes, respectively, and the label predicted by the tree.

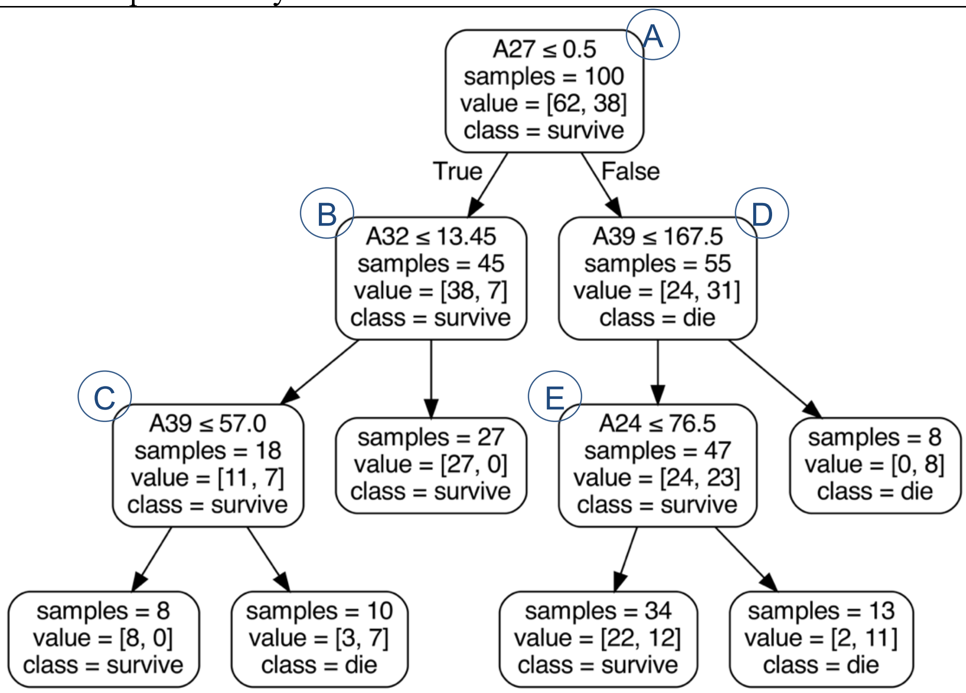

Figure 1 Decision tree trained over 100 records

Suppose we generate 5 new variables computed as follows: A=(A27<=0.5), B=(A32<=13.45), C=(A39<=57), D=(A39<=167.5) and E=(A24<=76.5).

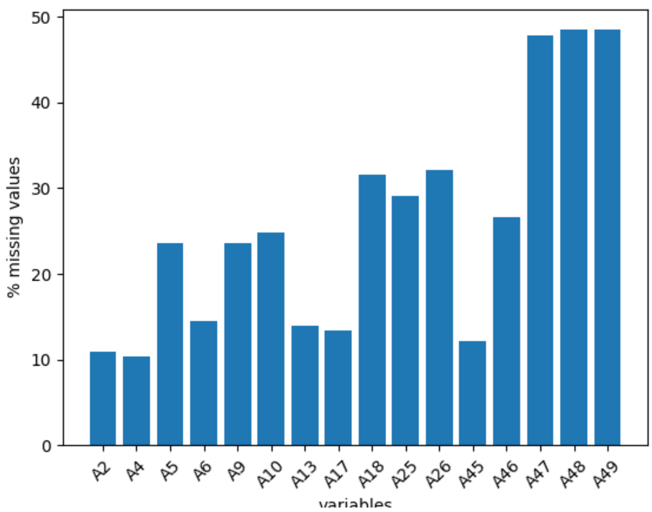

Figure 2 Variables with more than 10% of missing values

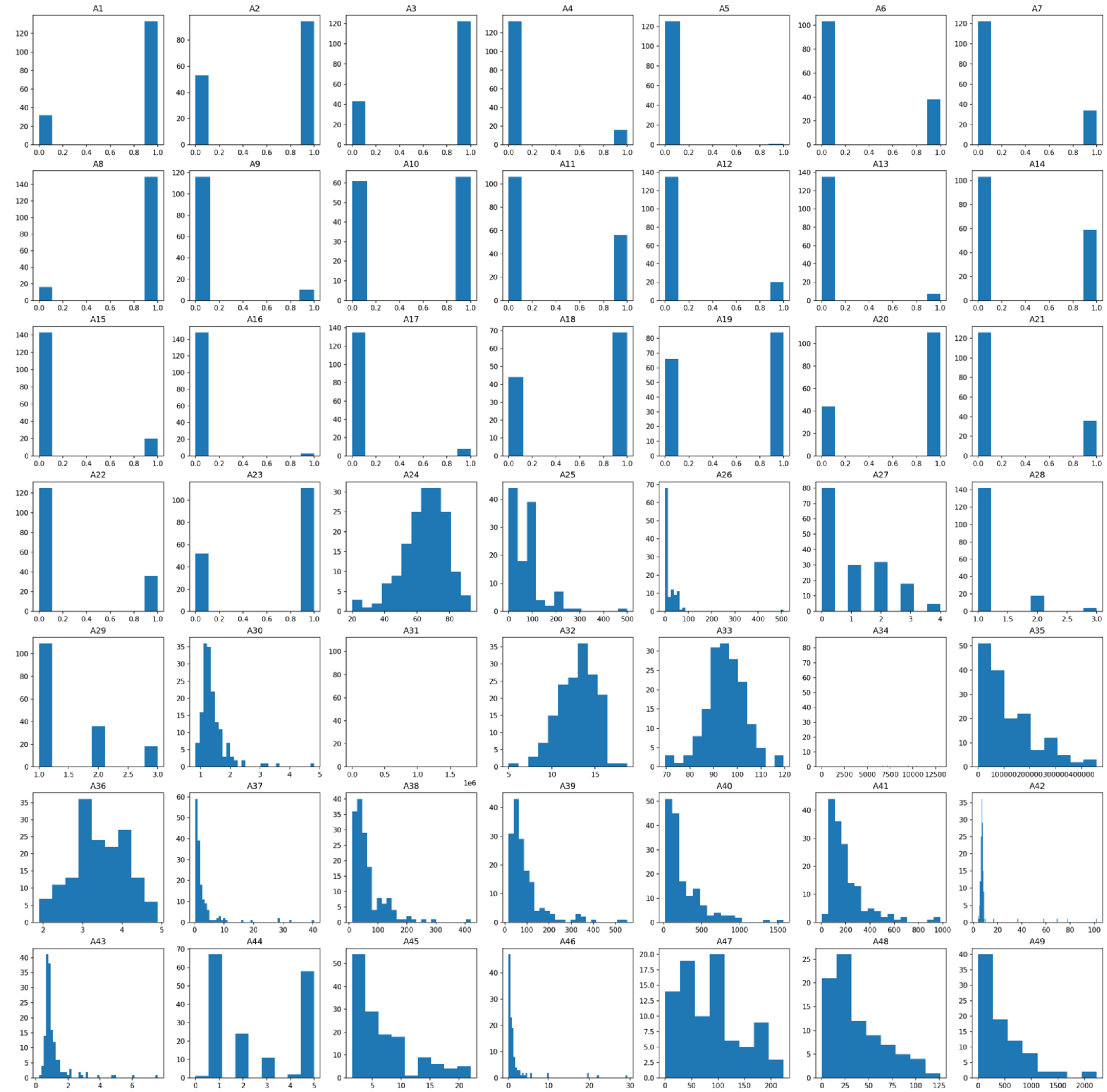

Figure 3 Histograms for all descriptive variables

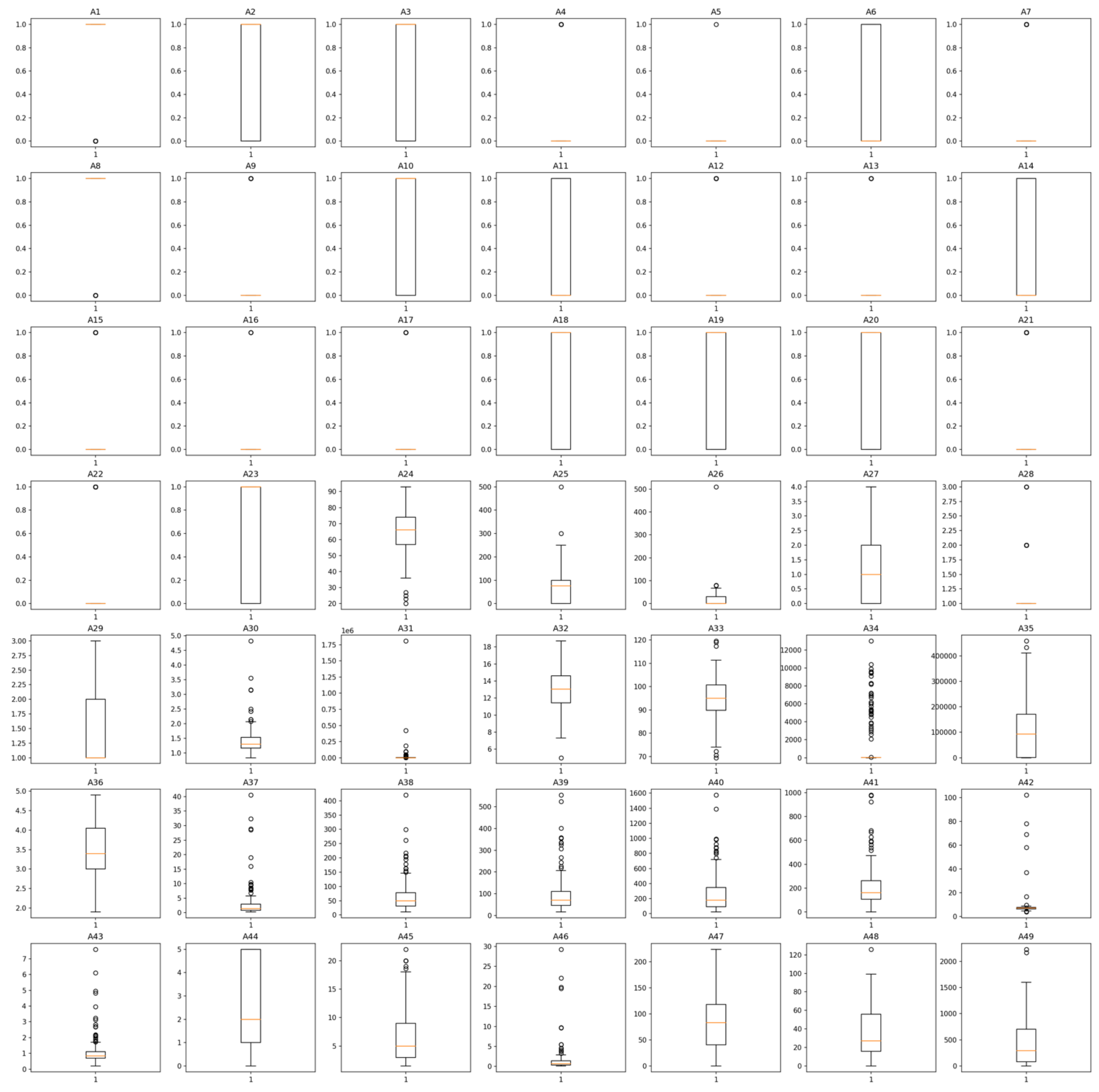

Figure 4 Boxplots for all descriptive variables

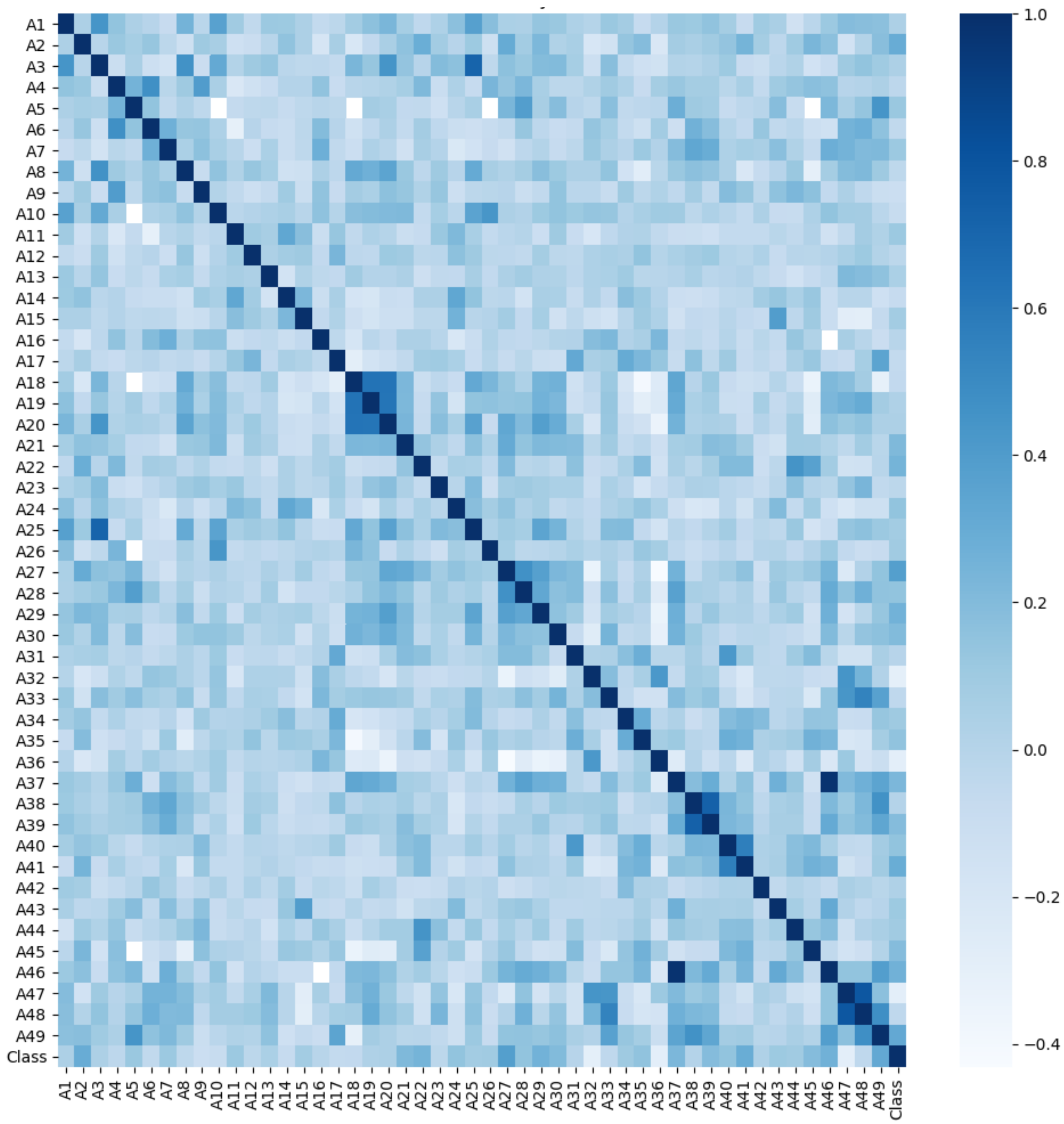

Figure 5 Correlation analysis

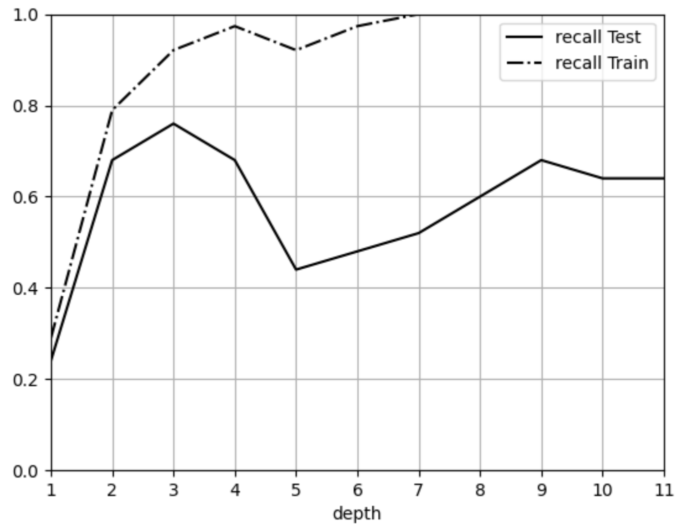

Figure 6 Recall for different decision trees specializations

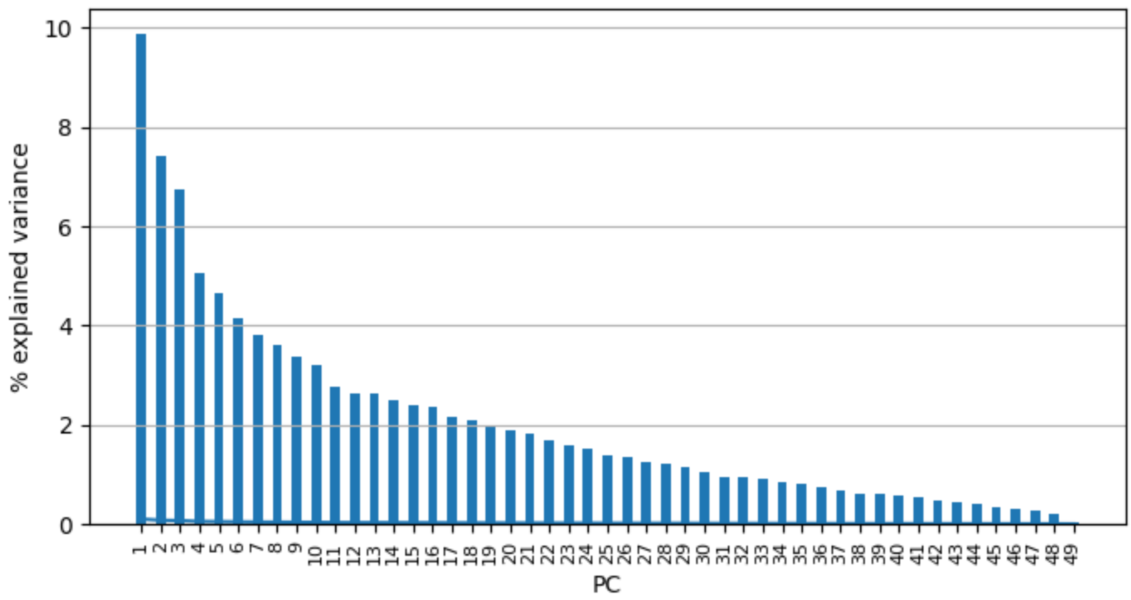

Figure 7 Explained variance for each principal component

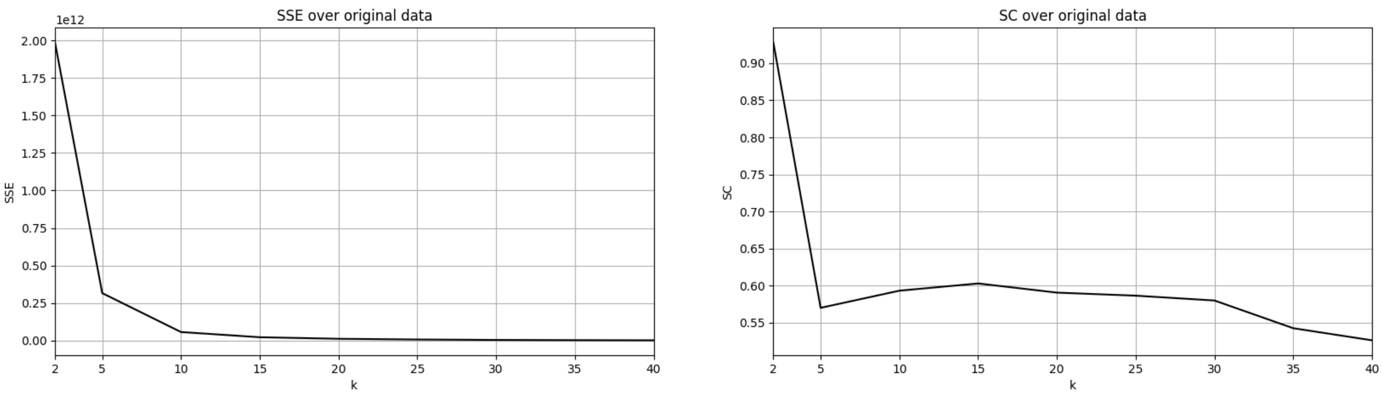

Figure 8 Sum of squared errors (SSE) and silhouette coefficient (SC) over original data along different number of clusters

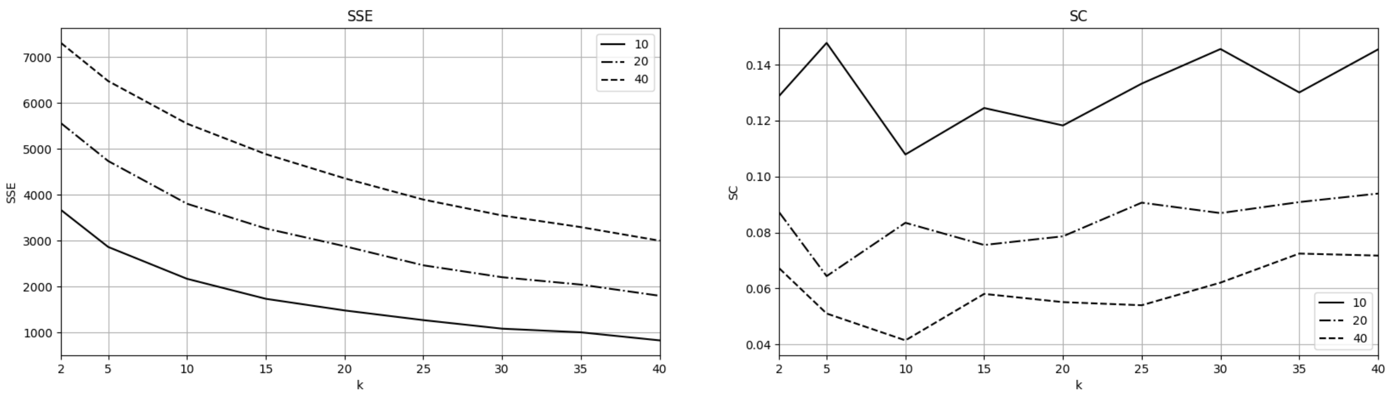

Figure 9 Sum of squared errors (SSE) and silhouette coefficient (SC) after applying PCA with different number of components along different number of clusters

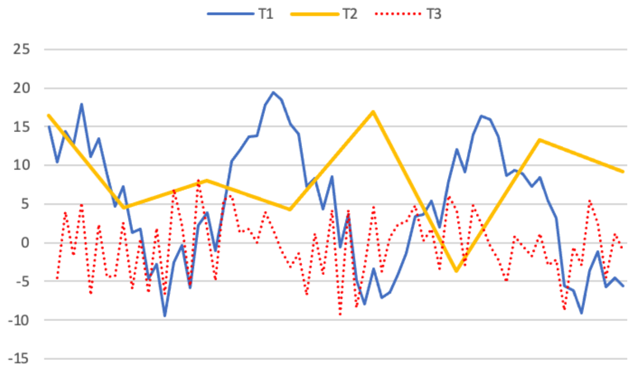

Figure 10 Time series

0.09	0.36	0.09	0.25	1	1.69	0.04	0.04	1.96
0.14	0.01	0.64	1	0.25	3.24	0.09	0.49	3.61
-0.04	3.61	1	0.64	5.29	0	2.25	1.21	0.01
0.16	0.01	1	1.44	0.09	4	0.25	0.81	4.41
0.05	1	0.01	0.01	1.96	0.81	0.36	0.04	1
0.07	0.64	0.01	0.09	1.44	1.21	0.16	0	1.44
0.05	1	0.01	0.01	1.96	0.81	0.36	0.04	1
0.16	0.01	1	1.44	0.09	4	0.25	0.81	4.41
	0.15	0.06	0.04	0.19	-0.04	0.11	0.07	-0.05

Figure 11 Accumulated cost matrix between T1 and T2

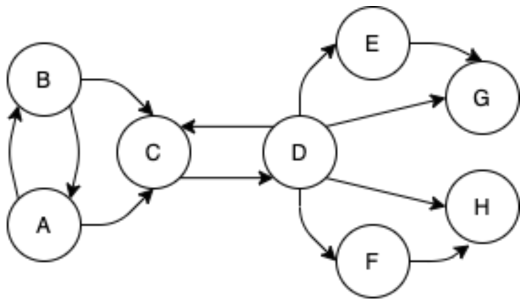

Figure 12 Social network

	A	B	C	D	E	F	G	H
A		0.5	0.5					
B	0.5		0.5					
C				1				
D			0.2		0.2	0.2	0.2	0.2
E							1	
F								1
G	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
H	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

Figure 13 Matrix

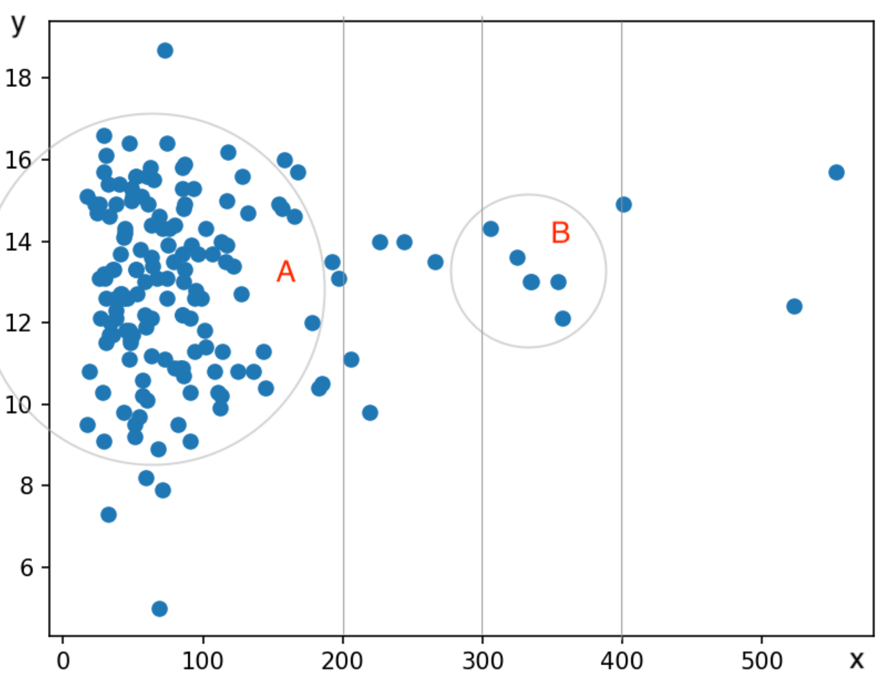

Figure 14 Data for anomaly detection: records in A and B are regular and the others are anomalies

Statements

A. Data Profiling

1. Considering the dataset described in Figure 1: all variables, but the class, should be dealt with as numeric.
2. Considering the dataset described in Figure 1: outliers seem to be a problem in the dataset.
3. Considering the dataset described in Figure 1: variable A27 is one of the most relevant variables.
4. Considering the dataset described in Figure 1: the intrinsic dimensionality of this dataset is 30.
5. Considering the dataset described in Figure 1: the number of existing missing values in this dataset **highly** impairs the learning process.

B. Data preparation

1. Considering the dataset described in Figure 1: If A38 and A39 were redundant then selecting just one of them would obviously increase the accuracy of KNN.
2. Considering the dataset described in Figure 1: Dummification is mandatory in this dataset.
3. Considering the dataset described in Figure 1: Discarding variables A48 and A49 would be better than discarding all the records with missing values for those variables.
4. Considering the dataset described in Figure 1: Using the first 30 principal components would imply an error between 10 and 20%.
5. Considering the dataset described in Figure 1: Feature generation based on the use of variable A16 **wouldn't be** useful, but the use of A5 seems to **be** promising.

C. Classifiers Evaluation

1. The accuracy for the tree in Figure 1 is 62%.
2. As reported in the tree in Figure 1, the number of False Positive is **smaller** than the number of False Negatives.
3. The recall for the tree in Figure 1 is **less than** 75%.
4. The chart on Figure 6 reporting the recall for different trees shows that the model **enters** in overfitting for models with **depth higher than 5**.
5. Consider the tree in Figure 1: a smaller tree would be delivered if we would apply post-pruning, accepting an accuracy reduction of 5%.

D. Classification

1. The performance of a decision tree trained over the binarized dataset would be **smaller** than the one in Figure 1.
2. Considering the binarized dataset and that records just described by A, B and C: KNN would classify the record **(A,B,C)** as **survive**, independently of the k chosen.
3. Considering the binarized dataset, records just described by A, B and C and suppose that the posterior probability for the true value of each variable given `survive` is always a third of the posterior probability given `die`: Naïve Bayes would classify the record **(A,B,C)** as **survive**.
4. A random forest trained over the binarized dataset described by the 5 variables and only considering regular decision trees with a maximum depth=2, would generate **less than 200 different** decision trees. `(Remember that a regular tree is one that tests the same variable on all branches with the same parent).`
5. Considering the dataset described in Figure 1, binarized or not: the descent algorithm would be able to learn a perceptron (a single neuron) with a good performance, since the data is linearly separable.

E. Pattern Mining

1. Consider 10% as the minimum support threshold and the binarized dataset, **(A,B,C,D,E)** is a **not** pattern.
2. Consider 10% as the minimum support threshold and the binarized dataset: knowing **(A,D)** is frequent, then **(A,D,E)** has to be frequent.
3. Consider the binarized dataset: the rule **A⇨B** presents a confidence higher than 80%
4. Consider the binarized dataset: the rule **A⇨B** presents a lift smaller then 2.5.
5. Consider the binarized dataset: the lift for the rule **(A,B)⇨C** is the same as for the rule **C⇨(A,B)**.

F. Clustering

1. Based on the SSE chart (Figure 8) we can say that the partition with 40 clusters is a **good** partition.
2. According to Figure 8 and Figure 9: the partition obtained over the original data is better than the one after applying PCA.
3. According to Figure 9: after applying PCA, the partition with 2 clusters is as good as the one with 40 clusters.
4. According to Figure 9: the number of principal components used to represent the data shows significant improvement in the partitions obtained.
5. The SSE for the original data is larger than for the data after applying PCA, which may be explained by the scale operation required by PCA.

G. Time Series

1. Consider the time series in Figure 10: the time series T1 exhibits a seasonal component.
2. Consider the time series in Figure 10: the series T2 is a piece-wise aggregate approximation of T1.
3. Consider the time series in Figure 10: the series T3 may be a differentiation of T1.
4. Consider the accumulated cost matrix between two series in Figure 11: the dynamic time warping path between them is (1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8).
5. Consider T1=[1.5, 0.6, 0.4, 1.9, -0.4, 1.1, 0.7, -0.5] and T2=[1.6, 0.5, 0.7, 0.5, 1.6, -0.4, 1.4, 0.9]). If T2 were the prediction of T1 according some regression model, its MAE would be between less than 1.5.

H. Social Network Analysis

1. Consider the social network presented in Figure 12: node C is more central than node D.
2. Consider the social network presented in Figure 12: node C is more prestigious than node D.
3. Consider the social network presented in Figure 12: the smallest path from node A to node H is 3 edges.
4. Consider the social network presented in Figure 12: the matrix in Figure 13 is the transformation applied in the context of the PageRank algorithm, to make the graph adjacency matrix into a stochastic one.
5. Consider the social network presented in Figure 12: if a **new edge** would be added from G to H, then **D**'s PageRank score would **kept unchanged**.

I. Anomaly Detection

1. Based on Figure 14: records with $x \geq 400$ are point anomalies.
2. Based on Figure 14: the records with $200 \leq x \leq 300$ are collective anomalies.
3. Based on Figure 14: clustering might be able to identify all the anomalies, using K-means and Euclidean distance.
4. Based on Figure 14: LOF would identify records in A, but it would be hard to identify the ones in B without considering the records with **$200 \leq x \leq 300$** also as regular.
5. Based on Figure 14: a non-parametric statistical-based approach (based on histograms for example) would be able to identify records in B as regular and records with **$200 \leq x \leq 300$** as anomalies.

J. Ethical Concerns

1. According to the GDPR, the collection and processing of student results from their school is legal, despite students consent.
2. According to the GDPR, a school may transfer student results to a third party, whenever for students benefit.
3. Privacy protection is one of the fundamental principles for practice ethical data sharing.
4. A school responsible for issuing studies certification, collect health data about their students, which is **adequate** according to the GDPR.
5. As a data scientist, your opinion matters.

Good work!!!