Technical University of Denmark

02450 Introduction to Machine Learning and Data Mining

Data: Feature extraction, and visualization

Authors:	***Student numbers:***
Filippo Bentivoglio	s210299
Jonas Løvenhardt Henriksen	s195457

Contents

Preface

This report is the first of two reports in a study, conducted during the Autumn of 2021 as part of the course "02450 - Introduction to Machine Learning & Data Mining" offered at the Technical University of Denmark (DTU).

The objective of this report is to apply methods learned during the course on a data set, giving a basic understanding of the data prior to further analysis. This further analysis is the objective of the second report. All students have contributed to the project equally.

Section	**Filippo s210299**	**Jonas s195457**
1. Description of data set	60%	40%
2. Explanation of data attributes	40%	60%
3. Data visualizations	60%	40%
4. Discussion	40%	60%
5. Exercise	60%	40%
6. Code	40%	60%

Table 1: Students contribution

1 Description of data set

1.1 What the data is about

The data used for this study is a forest fire data from the Montesinho natural park (Tras-os-montes northeast region of Portugal). The data used in the experiments was collected from January 2000 to December 2003. The problem of interest of this study is to predict forest fires (wildfires) based on some meteorological condition measures and the Fire Weather Index (FWI), which is a system, designed in the 1970s in Canada, for rating fire danger used for prevention and to support fire management decision.

The data set consists of a 518 x 13 matrix, containing measurements of four components of the FWI system and several weather observation for 517 wildfires registered during the period of interest.

1.2 Data reference

The data set was obtained putting together the Montesinho natural park' s database, regarding the date, spatial location, some components of the FWI system and the total burned area, and a database collected by the Bragança Polytechnic Institute, containing several weather conditions. The reader may access the data following this link: `https://archive.ics.uci.edu/ml/datasets/Forest+Fires` [1].

1.3 Previous analysis of the data

In [1], the considered dataset has been used for predicting the burned area of forest fires using Data Mining approaches, e.g. Support Vector Machines and Random Forests. Before fitting the models, some preprocessing has been applied: *month* and *day*, nominal attributes, were transformed into a 1-of-C encoding. All attributes were standardized to a zero mean and one standard deviation. To infer about the inpact of the input variables, four distinct feature selection setups were tested for each DM algorithms. The best configuration uses a SVM and four meteorological input (temperature, relative humidity, rain and wind) to predict small fires. However, it does not perform well when it comes larger fires.

1.4 Context of problem

As mentioned before, the aim of the FWI system is rating forest fire, in particular predict the burned area (size) by forest fires, based on the remaining attributes of the data set. By using every available attribute, the aim of this project is to determine a set of Principal Components (PCs), which have a sufficiently covering level of variance explained.

Talking about what can be done in the future with this data set, there are two options available. First, create a *classification* model in order to predict to which class a new fire belongs to (in term of dimension) having grouped the observation in n class based on the dimension of the burned area. Second, predict the dimension, using a *regression* model.

To handle the data it was needed to remove the first row, containing the attributes' names, and transform the attribute values `Month` and `Day` respectively decoded with index [1-12] and [1-7]. Furthermore, the main attributes (burned area size) is a continuous and interval attribute but we decide

to transform it in a binary variable (0 or 1) depending on the size of the burned area (0 smaller than a certain value and 1 higher).

2 Explanation of data attributes

2.1 Description of data attributes

The data set attributes are presented in Table 2.

Attribute	Short Description	Attribute type
X	x-axis coordinate (from 1 to 9)	discrete, nominal
Y	x-axis coordinate (from 1 to 9)	discrete, nominal
month	Month of the year	discrete, interval
day	Day of the week	discrete, interval
FFMC	Fine Fuel Moisture Code	continuous, ratio
DMC	Duff Moisture Code	continuous, ratio
DC	Drought Code	continuous, ratio
ISI	Initial Spread Index	continuous, ratio
temp	outside temperature (in $^{\circ}C$)	continuous, interval
RH	ouside ralative humidity (in $\%$)	continuous, ratio
wind	outside wind speed (in km/h)	continuous, ratio
rain	Outside rain (in m^2)	continuous, ratio
area (original)	Total burned area (in ha)	continuous, interval
area (transformed)	Size burned area	binary, nominal

Table 2: Attributes of the data set

Here follows an explanation of the attributes.

- *X*, *Y*, *month* and *date*: denote the spatial and temporal condition of the forest fires recorded. *X* and *Y* stand for the spatial location within a 9 x 9 grid by which the Montesinho natural park has been divided for the analysis.
- *FFMC*: denotes the moisture content surface litter and influences ignition and fire spread.
- *DMC* and *DC*: represent the moisture content of shallow and deep organic layers, which affect fire intensity.
- *ISI*: is a score that correlates with fire velocity spread.
- *temp*: instant temperature record when fire was detected.
- *RH*: instant relative humidity record when fire was detected.
- *wind*: instant wind record when fire was detected.
- *rain*: accumulated precipitation within the previous 30 minutes before fire was detected.
- *area (original)*: dimension of the burned area. It is not a *ratio* attribute, because the value 0 does not mean absence of fire, but it means that the burned area is lower than 100 m^2.
- *area (transformed)*: transformed in binary variable depending on the burned area size: if lower than $100m^2$ is equal to 0, otherwise 1.

2.2 Data issues

The data-set does not have any missing values and no visible corruptions

2.3 Summary statistics

To get an overview of the data set, brief summary statistics of the attributes were made.

	mean	**std**	**min**	**Q1**	**median**	**Q3**	**max**
X	4.67	2.31	1.00	3.00	4.00	7.00	9.00
Y	4.30	1.23	2.00	4.00	4.00	5.00	9.00
month	7.48	2.27	1.00	7.00	8.00	9.00	12.00
day	4.26	2.07	1.00	2.00	5.00	6.00	7.00
FFMC	90.64	5.51	18.70	90.20	91.60	92.90	96.20
DMC	110.87	63.98	1.10	68.60	108.30	142.40	291.30
DC	574.94	247.83	7.90	437.70	664.20	713.90	860.60
ISI	9.02	4.56	0.00	6.50	8.40	10.80	56.10
temp	18.89	5.80	2.20	15.50	19.30	22.80	33.30
RH	44.29	16.30	15.00	33.00	42.00	53.00	100.00
wind	4.02	1.79	0.40	2.70	4.00	4.90	9.40
rain	0.02	0.30	0.00	0.00	0.00	0.00	6.40
area	12.85	63.59	0.00	0.00	0.52	6.57	1090.84

Table 3: Summary statistics of attributes

In the data set chosen for this report, it is clear that the scales of the values are markedly different. As illustrated Table 3, *DC* values range approximately from 8 to 860, and the standard deviation (std) has a value of 247.83, while *wind* range is 0.40-9.40 and has std 1.79.

Overall, the 4 parameters related to the FWI system (*FFMC, DMC, DC, ISI*) and *RH* have a wider range and higher standard deviation compared to the others. Therefore, based on this initial and quick observation, with a great degree of certainty, we can say that we expect to need not only to center the data (subtracting the mean) as the PCA algorithm requires, but also to standardize our data (dividing each attribute by its standard deviation). Otherwise, the first principal component will be extremely driven by the *DC* attributes, which has the highest variance, in order to account for as much of the variance as possible in the data.

It can be noticed, regarding the output variable *area*, it is positive skewed and the majority of the samples have zero value, denoting that the burned area is small (lower than $100m^2$). To reduce the skewness and improve symmetry, in a future work, it could be beneficial to apply the logarithm function $y = ln(x + 1)$ to the variable.

3 Data visualizations

3.1 Outliers

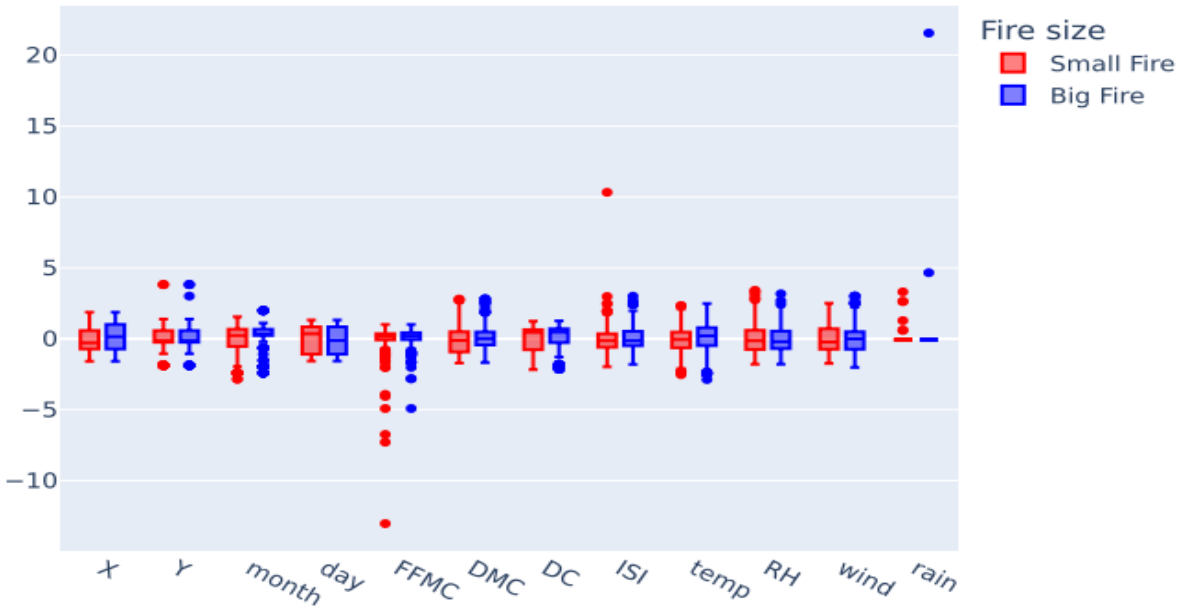


Figure 1: Box plot on the standardised data set.

As shown in Table 3, there is large variation in scales between attributes that are continuous and discrete. This can be a setback because so much variation in data (scales) can be hard to visualize and may baffle the machine learning model. To overcome this setback the data has been standardized by subtracting the mean and dividing by the standard deviation.

The box plot is a crucial tool for understanding whether or not our data set presents outliers. For a comprehensive understanding, the box plots are grouped by the target variable *Fire size*. In this specific data set, several outliers can be found across the attributes. However, if we think at the meaning of each attribute we can come up with a plausible explanation. For instance *rain*, we can consider the outliers shown as extreme case where it rained before the fire, which could be a nonsense since the rain should prevent the fire. Similar consideration can be done for both the other weather conditions and the FWI attributes, where extreme and uncommon condition has been recorded. Furthermore, even regarding the temporal attributes we can find a logical explanation. Looking at the *month* attribute we can see that most of the outliers are below 0, meaning that several forest fires has been recorded during the first months of the year (since the data are centered) which can be considered unlikely since during that months the cold weather should help preventing fires from arising.

For the reason explained above, no one of the observation has been discarded because removing outliers without clear reasons may affect conclusions deducted from the data.

3.2 Distribution

In this section we analyze the features to find if the observations follow a normal distribution. We will highlight some of the most interesting findings from this analysis. To get a visual aid for for the distribution densities we plot the data in a histogram of each feature, see Figure 2.

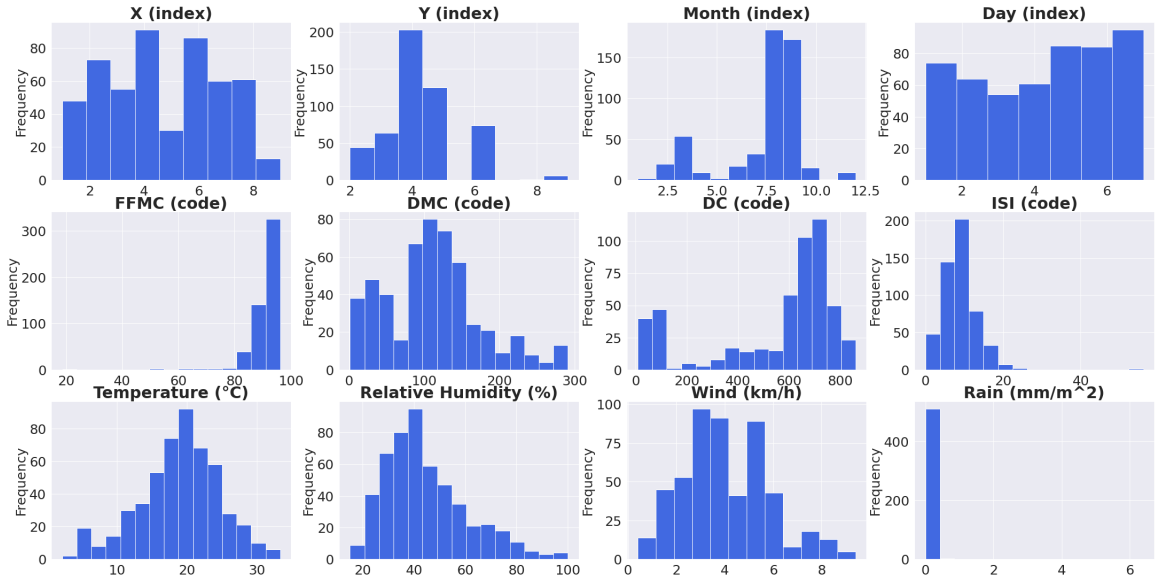


Figure 2: Histograms.

After reviewing the histograms, we can see that they tend not to follow a normal distribution. We for example see the *DC index* is left skewed, presenting also a second lower peak for low values. As well as the *FFMC* code. Instead, concerning *DMC*, *ISI*, *Temperature*, *RH* and *Wind*, despite a bit a skewness, it seems feasible to be approximated with a normal distribution. Regarding the *Rain*, as discussed in 3.1, there are only a few instances where a fire is preceded by rain, which is why the bulk of the observations have a value of 0.

We can also consider that the temporal attributes (*month* and *day*) are nominal, consequently we can not tell if they follow a normal distribution. There is an interesting observation to be made with the *month* feature. By viewing the histogram it is apparent that most of the density belongs to the 9th and 10th *month*.

Another interesting observation is the geographical features (X and Y). We see that these do not appear to follow a normal distribution, X having to peaks of densities and Y seemingly having a tail on the right side. We see from the distributions that most fires occur at the 4 and 5 Y-coordinate and the X-coordinate having a more even spread.

3.3 Correlation

For having an effective machine learning model, the non-target attributes should give a good description of the output variable and be as most as possible non-correlated to each other.

The correlation matrix is useful in order to assess the correlation between the attributes. It is a symmetric matrix where each value, that describes the relationship between two variables x and y, is

calculated by the following equation:

$$\widehat{cor}[x, y] = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\hat{\sigma}_x \hat{\sigma}_y}$$

where i is the observation number, N is the total number of observations, $\hat{\mu}_x$ and $\hat{\mu}_y$ are the empirical means of x and y respectively, and $\hat{\sigma_x}$ and $\hat{\sigma_y}$ are the empirical standard deviation of x and y.

From Figure 3, it is possible to see that the most correlated attributes are *month* & *DC* and *DMC* & *DC*, with scores 0.87 and 0.68. Furthermore, the 4 FWI system's variables and *temp* result to be the attributes with the highest correlation coefficients between each other.

It is worth to mention that *month* and *DC*, even if only with little amounts, are the two attributes that explain the most about the target attribute *area* with a correlation of 0.13 and 0.097 respectively.

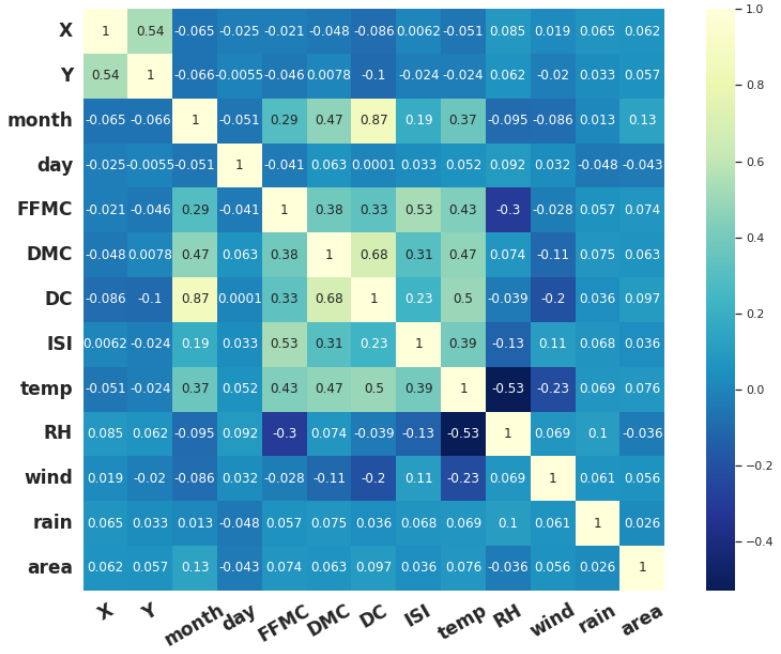


Figure 3: Correlation matrix.

3.4 Feasibility of modeling aim

To asses whether the primary machine learning modelling aim of this study appears to be feasible, based on the previous graphs, the following consideration can be made:

- According to the box plot (Figure 1) and summary statistics (Table 3), no observation needed to be removed or altered.
- Few problems can arise if we look at the histograms plot (Figure 2), since, for instance, the *DMC* seems to follow a bi-normal distribution and the majority of the attributes does not have a really clear normal distribution.
- In the correlation matrix (Figure 3), the coefficients between attribute-attribute pairs are non-zero and mostly low. Furthermore, there is not single attribute that lead the correlation with the target attribute (*area*), as well as not single zero-value coefficient between the target and the other attributes.

Said so, the data may need some manipulation to get useful results, since the attributes for large parts are not normal distributed. Hence, based on the above consideration, the modelling aim seems to be possible.

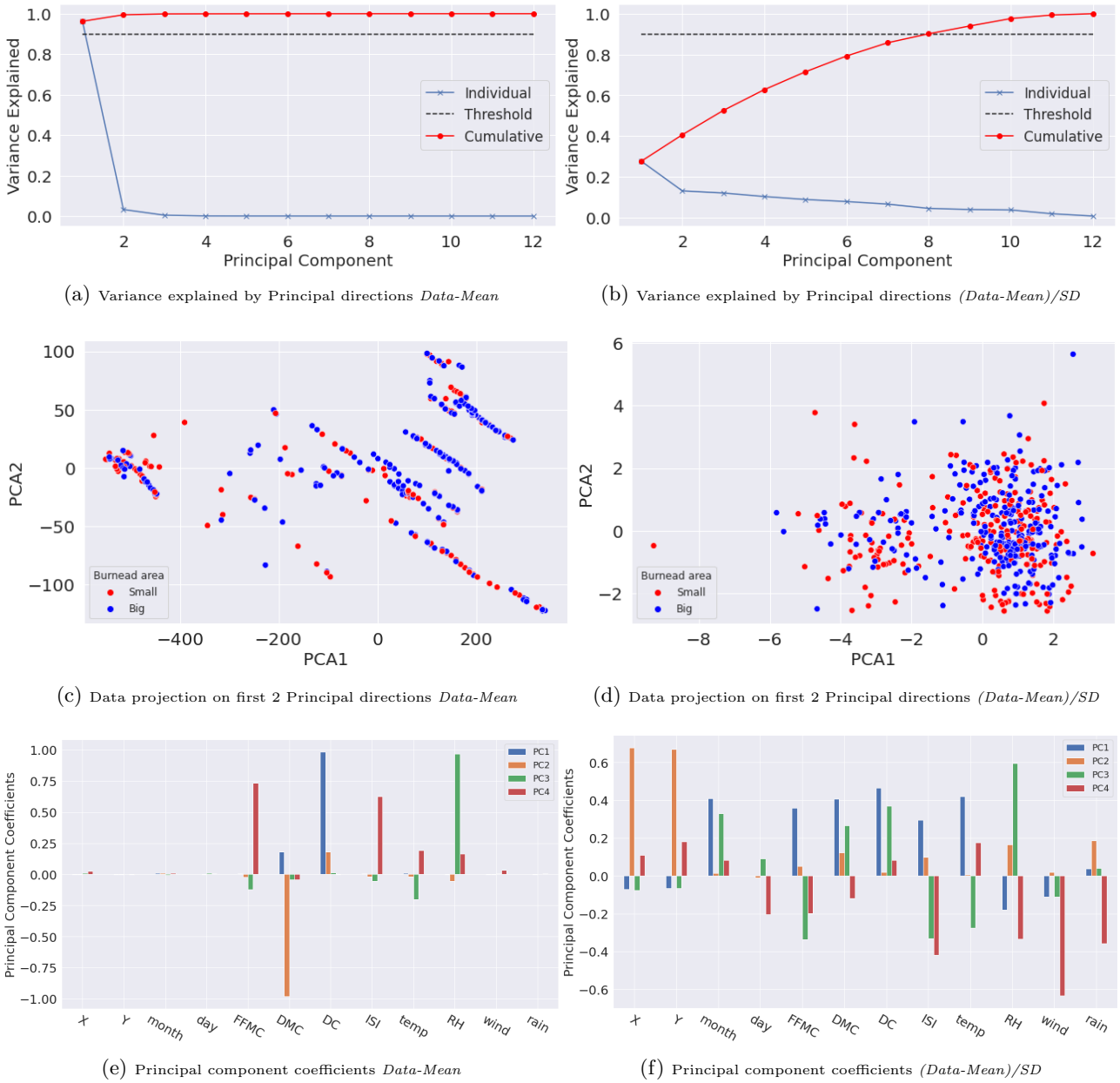
3.5 Principal component analysis

The goal of the Principal Components Analysis is to find a linear lower-dimensional representation of a high-dimensional data set with an great level of variance explained. Then the transformed data set will be used in the machine learning model.

The PCA algorithm requires to perform first linear transformation of the data by subtracting the mean of each attribute from the original value, then, if the data have different range, which is our case, dividing each attribute by its standard deviation in order to have all the attributes on the same scale. Finally, the Singular Value Decomposition is applied to the transformed data set to compute the eigenvectors that maximize the variance.

In the data set chosen, the attributes are in different scales (Table 3) so for the final analysis the standardization need to be applied to the data before applying PCA. However, to get a better understanding of why this step is needed, both the results obtained on the non-standardize and standardize data are shown. In both the case, a threshold of 90% of variance explained has been chosen.

Looking at the pane 4a, it is possible to deduce that the first principal component alone accounts for more variance than the adopted limit. Referring to pane 4c, it is clear that the first principal component is extremely positively correlated to the attributes DC (which is the attribute with the widest range and standard deviation), meaning that to obtain a large projection on the first principal component direction the observation should have a large positive DC value. Hence, in this scenario, adopting a model with only the first PC is enough to have the cumulative variance higher than the threshold. Nevertheless, looking at 4c, where the observations have been plotted on the first two principal directions, the distribution results distorted since it is not enough to discriminate between *small* and *big* burned area, and probably the second principal component captures only noise.



(a) Variance explained by Principal directions *Data-Mean*

(b) Variance explained by Principal directions *(Data-Mean)/SD*

(c) Data projection on first 2 Principal directions *Data-Mean*

(d) Data projection on first 2 Principal directions *(Data-Mean)/SD*

(e) Principal component coefficients *Data-Mean*

(f) Principal component coefficients *(Data-Mean)/SD*

Figure 4: Plots from Principal Component Analysis: PCA without standardization (left panes) & PCA including standardization (right panes).

Moving to the right panes, after the standardisation, PC1-PC9 are needed in order for cumulative variance explained to exceed the threshold (pane 4b). Moreover, the first principal component is not anymore solely correlated with the *DC* attributes, but it is more equally distributed between the four FWI's index, the *temp* and *month*. In all the case, there is a positive correlation meaning that observation showing high positive value for these parameters, will be re-projected in large positive value along the first principal direction (similarly for negative values). Really interesting and probably useful for further analysis, it is the second principal component which is extremely related to the (x,y) location of the recorded fire inside the grid in which the Montesinho park has been splitted into. Finally, it's worth to mention, the 4th principal component is mainly negatively related to the weather condition (*Rh*, *wind* and *rain*) and the *ISI* index. The projection on the first 2 principal directions (pane 4d) is not really improved, showing that to obtain a good classification of the forest fires several attributes need to be taken into account.

4 Discussion - Leanings about the data

The data set consists of 12 attributes and one target value (*area*); only two of them are nominal (*month* and *day*) which have been converted into their correspondent number. The PC analysis has shown which problem can arise if the different data's scales and standard deviation are not taken into consideration, otherwise the results obtained by machine learning model could be compromised and misled.

The aim of this study is to predict whether or not the dimension of the burned area, following a forest fire, is *small* or *big*. Hence, in a future work, there are two options available. First, create a *classification* model in order to predict to which class a new fire belongs to (in term of dimension) having grouped the observation in n class based on the dimension of the burned area. Second, predict the dimension, using a *regression* model.

As suggested in [1], the results can be improved selecting only a subset of the input attributes (for instance, only the weather condition). In subsection 3.3, we have seen that DC and *month* are highly correlated, then, it could be useful to discard one of them to reduce the dimension of the data set.

5 Appendix: exam problems

5.1 Question 1

Answer: A.

- *Time of day* is nominal because the values refer to a specific time interval and it does not make any sense to compare time interval of the same leght.
- *Traffic lights* and *Running over* are ratio because the value 0 has in both case a physical meaning: any traffic light has been broken and there were not any run over accident respectively.
- *Congestion level* is ordinal because we can apply the <,> operands to establish the level of congestion.

5.2 Question 2

Answer: A.

The *Max norm distanced*$_{p=\infty}$ is defined as:

$$d_{p=\infty} = max\{|x_1 - y_1|, .., |x_M - y_M|\}$$

which in our case is:

$$d_{p=\infty} = max\{7, 0, 2, 0, 0, 0, 0\} = 7$$

5.3 Question 3

Answer: A.

The variance explained by the first 4 principal components is calculated as follow:

$$VarExpPC1-4 = \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.867$$

5.4 Question 4

Answer D.

Time of day is the only attribute with negative coefficient on the principal direction $v_2 = [-0.5, 0.23, 0.23, 0.09, 0.8]^T$ while *Broken Truck*, *Accident victim* and *Defects* have positive ones. This means that an observation to have a positive or large projection on v_2 need to have low value for *Time of day* and large ones for the remaining attributes.

5.5 Question 5

Answer A.

In a text document, the Jaccard similarity can be calculate by the following formula:

$$J(s_1, s_2) = \frac{s_1 \cap s_2}{s_1 \cup s_2}$$

which is the intersection of the two documents, so the number of words they have in common, divided by the total number of words. It gives:

$$J(s_1, s_2) = \frac{[bag, words]}{[the, bag, of, words, representation, becomes, less, parsimonious, if, we, do, not, stem]} =$$

$$= \frac{2}{13} = 0.153846$$

References

[1] [Cortez and Morais, 2007] P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9