External Climate Data Extraction Using the Forward Feature Selection Method in the Context of Occupational Safety

Felipe G. Silva(✉), Inês Sena, Laires A. Lima, Florbela P. Fernandes, Maria F. Pacheco, Clara B. Vaz, José Lima, and Ana I. Pereira

Research Center in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
{gimenez,ines.sena,laires.lima,fflor,pacheco,clvaz,jllima,apereira}@ipb.pt

Abstract. Global climate changes and the increase in average temperatures are some of the major contemporary problems that have not been considered in the context of external factors to increase accident risk. Studies that include climate information as a safety parameter in machine learning models designed to predict the occurrence of accidents are not usual. This study aims to create a dataset with the most relevant climatic elements, to get better predictions. The results will be applied in future studies to correlate with the accident history in a retail sector company to understand its impact on accident risk. The information was collected from the National Oceanic and Atmospheric Administration (NOAA) climate database and computed by a wrapper method to ensure the selection of the most features. The main goal is to retain all the features in the dataset without causing significant negative impacts on the prediction score.

Keywords: Weather conditions · Data extraction · Occupational hazard prediction · Machine learning · Data mining

1 Introduction

Workplace safety is a relevant and multifaceted concern for workers, companies, and policymakers like ministers, legislators, and government officials [5] and it is becoming increasingly important around the world.

The need of ensuring minimum working conditions stems from Greek and Roman civilizations [10]. Ensuring good labour conditions has the potential to increase productivity and consequently improve the operational performance of organizations [3].

Safety is integrated empirically into work performance, as the work is dependent on the production factors. Implementing occupational safety measurements and health policies is essential to ensure a healthy and safe work environment, contributing to reduce the work accidents and occupational diseases, to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13378, pp. 3–14, 2022.
https://doi.org/10.1007/978-3-031-10562-3_1

increase the competitiveness of companies, and to improve the sociability, processes, productivity, and quality of products or services, and to adopt innovative practices [3].

Occupational accidents and diseases have multiple impacts on the company operations and their costs, workers, and society in general [6]. These accidents affect significantly the quality of life of workers and the economy of the companies. The loss in production capacity is clear, in terms of days off, compensation, and pensions payable by the companies [6].

Occupational accidents have been the subject of several studies over the years, developing theories that try to explain, prevent and reduce the accidents. As such, work-related accidents have been the subject of several studies, with the development of theories that try to explain, prevent and reduce them. Initially, the actions most used by the various sectors in the reduction of work accidents were focused on the detailed investigation of the event that caused the accident and on the implementation of preventive actions, such as the reduction of excessive workload [5], changing processes and procedures in carrying out tasks, improving and evolving machines and equipment, providing workers with information and training on OSH [6]. However, several companies have started to adopt intelligent technologies that have already been implemented to improve other factors and which can be used to increase safety in the workplace, such as robotics, augmented reality, wearables, and predictive analytics [4].

An example of some areas of implementation is a predictive analysis based on Artificial Intelligence to forecast work hazards, namely, in energy, construction, infrastructures, agro-industries among others.

The studies that use predictive analysis generally apply mathematical models, statistical strategies, and computational techniques to predict the accident. Generally, the input parameters are based on the characterization of the accident, such as the day of the incident, incident event, type of injury, incident type, among others.

Some studies have identified global non-organizational variables that may influence safety and health at work, namely climatic conditions. However, there are few works using the variables and scarce strategies have been developed to explore the association between occupational accidents and climate conditions [2, 12].

Thus, this work aims to collect historical information about weather conditions and detect the most relevant elements to correlate in the future with the historical accident data of a company in the retail sector to understand its impact on the risk of accidents. This study is carried out to validate the importance of using this type of data as an input parameter in a forecasting model under development – in a parallel investigation – to minimize the risk of accidents in a Portuguese company in the retail sector.

This paper is divided into five sections. Section 1 describes the framework of the study carried out. Section 2 presents a brief literature review about the relevance of the climate data in accident prediction and the usual procedure used to extract this kind of data. The adopted procedure to collect and process data

is described in Sect. 3. In Sect. 4 the results obtained are analyzed. Finally, the study is concluded and future work is presented in Sect. 5.

2 Literature Review

The weather plays a significant role in everyday life and, from the consumer's point of view, is considered an influential factor in consumer buying behavior and in many aspects of decision-making in the retail sector [19]. The literature review presented in this section encompasses studies on accident prediction considering some climatic factors, along with data extraction techniques and methods for data pre-processing, namely feature selection methods.

2.1 Accident Prediction Models

Despite the relevance of this subject, studies regarding the modeling and prediction of accidents in the workplace considering climate data is scarce. In [12], a study was performed to determine which criteria should be adopted in terms of universal or industry-specific safety climate measures. The authors performed a meta-analytic comparison of their relationships with various of safety-related outcomes, such as safety behavior, risk perceptions, accidents, and injuries, among others. Industry-specific safety climate measures were found to exhibit better predictive power in predicting safety behavior and risk perceptions than universal safety climate measures.

Aleksic et al. [1] use temperature and air pressure as possible external variables to explore their impact on work capacity. A considerable number of studies discuss the dependence of working capacities on temperature. Based on existing research, the authors assume that air pressure can have a significant impact on the occurrence of accidents [1].

The prediction of accidents is based on a good analysis of the root causes, dealing with the consequences, and making the right decisions to prevent the occurrence of the same or similar accident.

Regardless of what is the observed and researched factor for the occurrence of the accident, the objective of the previous mentioned works, including this research, is to increase the level of safety by determining and analysing the factors that tend to reduce the risk of occurrence of the accident.

2.2 Data Extraction

Web data extraction is a complex and time-consuming activity since the web is a vast source of structured or unstructured information in different formats. The tools for collecting data from the web are often based on interacting with the original (website) and extracting data through various methods, such as machine learning, logic, and natural language processing. In the case of an HTML source, the content to be extracted can range from elements on the page to structured

hidden data. The extracted data can be post-processed, converted to different formats, and stored for later use [18].

The main challenges involving data capture are concerned with the complexity of data handling, being necessary to provide a high degree of automation to minimise human effort and maintain accurate performance. When extraction techniques are applied to extensive databases such as climate monitoring, capture tools need to be able to process a large volume of information in relatively short periods. On the other hand, as it involves temporal data which vary over time, changes can occur suddenly, so one must take into account the flexibility and structural changes in the data acquisition system [9]. Due to the general public's high consultation of climate data, several sources of weather data are available online and offer servers with accessible application programming interfaces (API). Despite that, studies regarding the influence of weather data on the prediction of accidents are scarce.

2.3 Data Pre-processing

Given the extensive data history and numerous real-time measurements, it is advisable to use computational techniques such as machine learning and data mining for processing high-dimensional data [22].

Feature selection, for example, has become a widely used technique to reduce the size of complex databases. It proposes to choose a subset of relevant features based on criteria of greater relevance, resulting in lower computational cost and better interpretation of the data [22]. The advantages are the improvement of machine learning performance, reduction of the database dimension by optimising the initial dataset for only relevant content, enabling the use of simpler models [20].

Feature selection algorithms can be categorized into supervised, unsupervised, and semi-supervised. Supervised selection evaluates the relevance of features based on labeled data. However, a better choice requires more labeled data and, consequently, time-consuming [7,8,11]. In contrast, unsupervised resource selection is based on a less constrained search problem but it may result in several less accurate and valid subsets [16]. For both cases, it is possible to observe that high-dimensional data with small labeled samples results in a considerable hypothesis space with few restrictions. The combination of the two methods results in semi-supervised feature selection, in which labeled and unlabeled data are combined to estimate the relevance of features [21]. Among the methods used for scanning and capturing climate data, we can highlight the supervised selection algorithms: filter models, wrapper models, embedded models, and hybrid models [7,8,11], which are summarized hereafter.

Filter methods are generally applied in data pre-processing steps. The selection depends on the general characteristics of the training data to select the features, regardless of the estimator, so the relevant variables are chosen based on statistical correlation tests (*e.g.* Pearson's correlation, ANOVA, LDA, and χ^2 test) with the response variable. It is a faster option and applicable to cases where the number of features is extensive [23].

Wrapper methods involve building subsets of features and training models. Based on subsets and inferences from previous models, new features can be added for a better approach. Despite being a good alternative as a forecasting method, it is computationally expensive. Some common examples of wrapper methods are forward selection, backward elimination, and recursive feature elimination [23].

Embedded methods combine aspects of filter and wrapper methods by classifying data and applying learning algorithms to quantify the weights of subsets. Lasso, ridge regression, regularized trees, memetic algorithm, and random multinomial logit are popular examples of embedded methods [23].

Finally, hybrid methods incorporate several types of feature selection into the same process. They are based on combining subsets of independent features, providing a better approximation while integrating external data into the selection process [23].

3 Methodology

Multiple climate factors can be related to accidents, and common examples can be identified every day such as people falling due to a wet floor, or changes in blood pressure of workers on hot days. Thus, climatic factors may be related to some occupational accidents that happen in work environments, it is important to include them in the predictive models to improve their accuracy, and consequently helping to define corrective actions to mitigate accidents.

For this study, it was necessary to get access to a climate dataset by exploring some strategies currently used to collect information and to search for available resources on the internet to produce a historical climate database. As main requirements, climate data resources should have free access and contain information from previous years, so that they can later be linked to the dataset of past accident occurrences.

The National Oceanic and Atmospheric Administration (NOAA) database [17] is used to collect the climate data since it contains historical data from different weather stations around the world. Through the API service available on the NOAA website, it was possible to search for climate information from 2017 until the present.

For greater convenience in obtaining the data, a script was developed using Python programming language to interact with the service provided by NOAA, enabling the automated collection of data recorded from January 1 of 2017 until September 1 of 2021. According to the location of the retailing stores, some specific geographical locations were selected to create the climate datasets, which are Bragança, Gaia, Felgueiras, Matosinhos, Sobreda and Guimarães, as shown in Fig. 1.

Since these climate data are derived from real records, some reading errors can occur due to failures of physical sensors installed in each specific weather station. To eliminate these errors, a pre-processing was applied to the data to prepare them for use.

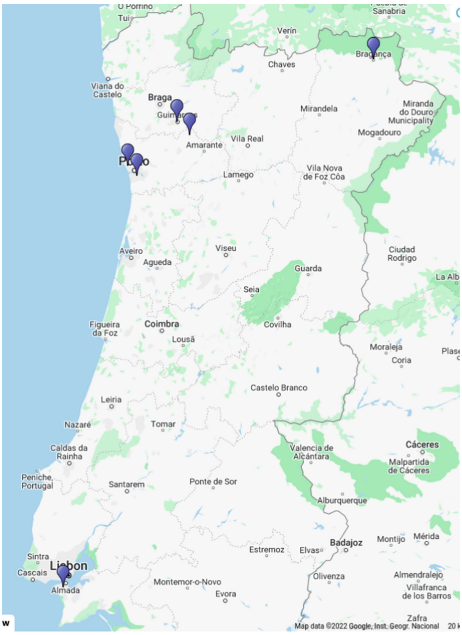

Fig. 1. Selected locations for climate studies.

In order to pre-process the data, record lines with missing information were removed. This missing data occur due to the weather stations that do not emit complete information on specific days or times, resulting in the total or partial lack of data. Due to the high number of records collected, that deletion did not result in shortage of information.

The date and time data were pre-processed to obtain new features to improve the representativeness of information. According to the work shifts of the retailing stores, four periods per day: dawn, morning, afternoon and night, were used to categorize the time of each record. The day and month were used to identify the seasons of the year in which a record was made. The year, month and day were used to identify the day of the week in which a new record was made.

In learning algorithms, excess of features can add noise that decreases the prediction success rate, an effect known as the curse of dimensionality [15]. Thus, the study of strategies for dimensionality reduction was applied to detect and select the information that proved to be most relevant for the generation of a predictive model. This data reduction also brings benefits in terms of space used for data storage, processing time consumed to generate the model and the time required to perform a prediction.

To ensure the selection of the relevant features, a wrapper method was used by implementing the forward feature selection. This strategy searches the features subgroups that achieve better accuracy, using the accuracy as selection

criteria after performing the prediction tests. Figure 2 represents the implementation applied in this study.

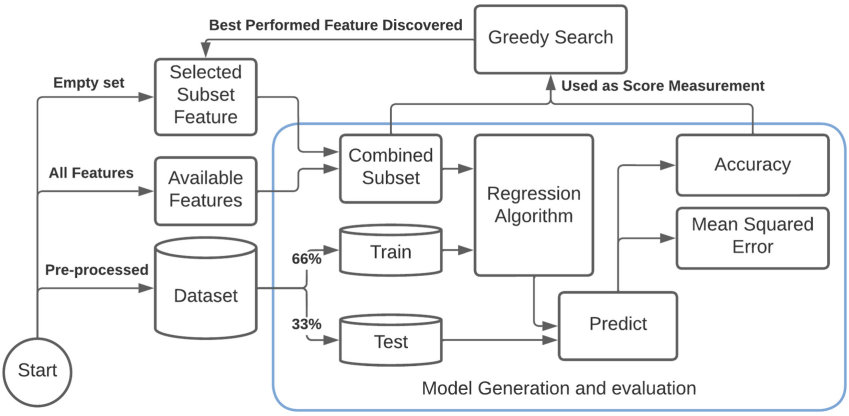

Fig. 2. Forward feature selection method.

The information about the air temperature was used as a target to perform the predictions. The air temperature was chosen because is used in other study [1] and presented a satisfactory accuracy with Linear Regression approach. To execute the tests, the linear regression algorithm provided by SKLearn library [14] was adopted. At each iteration of the algorithm, the behavior of the model was analized when adding a new feature, to detect if the model was getting better or worst.

The wrapper methods find a group of variables that achieve good performance during the prediction, being the results directly affected by the data chosen for training and testing, and also by the learning algorithm used. To reduce this effect, the method was repeated a hundred times for each database to detect the variables that perform better on average, by determining a ranking score that measures the importance of features in each database.

4 Results

Forward Feature Selection initiates with an empty subset and follows two steps: searching for a subset of features to be used; generating a score for the selected features. After these basic steps, the process is repeated until a stop criteria is satisfied [15]. To test all subsets found by the greedy algorithm, the stop criteria was ignored, so the algorithm stops when the available feature set is empty. Using the greedy algorithm as strategy for performing the search and considering the maximum number of input features, $\frac{(N+1)N}{2}$ models were evaluated, where N is the number of input features. In this study, the N is 20, evaluating 210 models during a unique test.

Each test was repeated 100 times for each location. Thus, for all the 6 geographical locations selected, 21000 models were evaluated, totalizing 126000 models, processed in 36 min and 13 s approximately. If the number of features increases, the regression algorithm can increase drastically the time to process the information.

By using 20 features, the total number of possibility sets is 2^{20} [13]. Then, the goal is to discover a good feature subset because finding the best implies testing all the possibilities, which can be impractical. The algorithm evaluates approximately 12% of all possibilities of subsets using the greedy search.

Since, the climate conditions can vary in different locations, each location was tested separately, resulting in the selection of the features subset that most contribute to achieve good regression scores for each specific location. The variables selected for each location are shown in Fig. 3 in which 6 features were defined for each location since the major groups did not present relevant accuracy improvements.

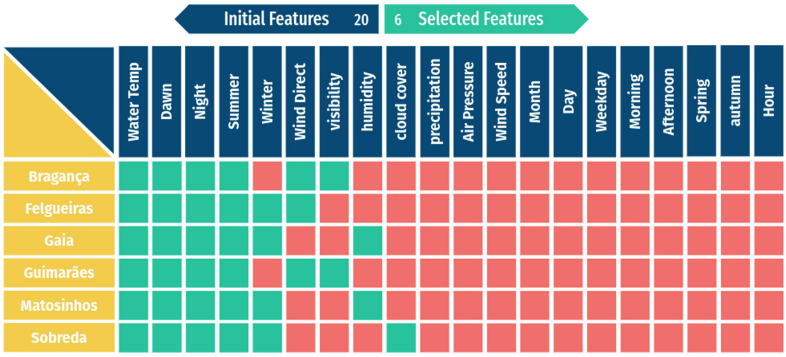

Fig. 3. Selected features for each location

After analyzing the number of features that had improved the accuracy, the variables to be considered were selected. Figure 4 shows the relationship between the number of variables in the best-discovered set and the improvement obtained in the regression accuracy. It is possible to observe that there is a maximum limit of improvement for each location since this score stabilizes from a given size of subset as it is shown in the Fig. 4.

To get the error value, was used 33% of dataset to predict the target. The errors were calculated at the better subsets of each size through the Mean Squared Error (MSE). Figure 5 shows the relationship between the size of the subset and the obtained errors.

Observing the average behavior of the variables detected for each location, of the twenty extracted initially, nine variables will be considered:

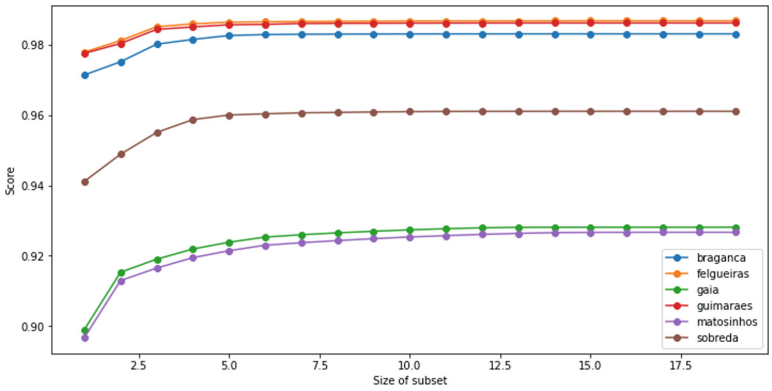

Fig. 4. Relationship between the number of characteristics and the score obtained in the regression

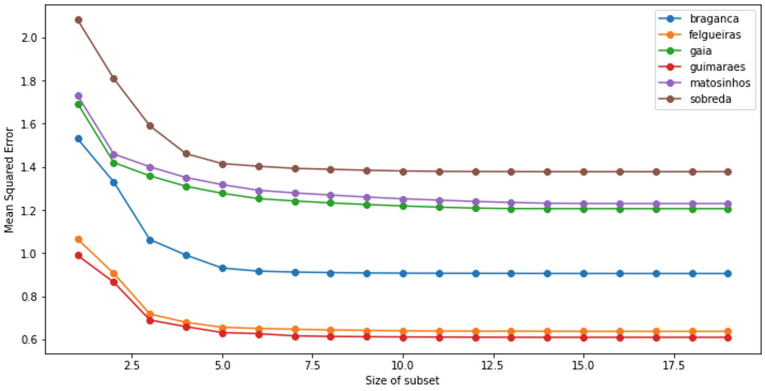

Fig. 5. Relationship between the number of features and the mean square error (MSE) obtained

1. cloud;
2. humidity;
3. visibility;
4. water temperature;
5. wind direction;
6. night;
7. dawn;
8. winter;
9. summer.

As the tests were carried out, the curse of dimensionality effect was detected, which enables to eliminate 14 features from each dataset without causing negative impact on the prediction. This approach enables to obtain an aproximation for the best group of features for learning, because greedy search algorithms can

converge to local maximums. The studied locations showed a consensus on four characteristics to be maintained and eleven to be eliminated. The remaining five features were still helpful to improve the assertiveness of the model, standing out in specific places. The four commom characteristics are: water temperature, night, dawn and summer.

5 Conclusions and Future Work

In the scenario used, the Forward Feature Selection approach proved to be advantageous over the statistical methods, such as Pearson's method. When analyzing the obtained results, only by direct correlation with the target feature, it was difficult to define which thresholds were good enough to include or exclude each input feature of the model. The Forward Feature Selection strategy enables to understand what changes were caused by accepting certain features into the subset.

The method also allowed to test predictions, because multiple models are generated and tested at each iteration. In this way, optimizing the model through the selection of features became a simpler and automated task. Each test performed by the algorithm brought feedback on the learning assertiveness that was being obtained.

Despite the existing advantages of the Forward Feature Selection, the use of the algorithm still has a major weakness, which is its computational effort. Although not all possible combinations of features to generate the models are evaluated, its execution time can make it unfeasible, since it creates a considerable quantity of predictive models at each iteration. The selection of the learning algorithm and the number of features directly impact its performance.

By changing the learning algorithm, there is the possibility of increasing the score obtained in the regression and decreasing the mean square error. So far, tests with regression strategies that consume greater computational resources have not been performed, due to the time that would be spent to obtain conclusions.

In the future, we intend to use faster methods that combine filtering and wrapper strategies, such as approaches with embedded feature selection, due to the need to use more demanding learning algorithms and the use of larger datasets and diversified data. Other feature selection techniques will also be applied to analyze the relationship and impact that climate variables have on accident records. In addition to climate information, it is also intended to collect more external information to work place, such as holidays and events, to detect other possible factors that can impact on the occurrence of occupational accidents.

Acknowledgments. This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the Project Scope UIDB/05757/2020 and NORTE-01-0247-FEDER-072598 iSafety: Intelligent system for occupational safety and well-being in the retail sector.

References

1. Aleksic, D., Markovic, M., Vasilijevic, M., Stojic, G., Pavlovic, N., Tanackov, I.: Analysis of impact of meteorological conditions on human factors in estimating the risk of railway accidents. Transport **33**, 57–69 (2018). https://doi.org/10.3846/16484142.2017.1332684
2. Alruqi, W.M., Hallowell, M.R., Techera, U.: Safety climate dimensions and their relationship to construction safety performance: a meta-analytic review. Saf. Sci. **109**, 165–173 (2018). https://doi.org/10.1016/j.ssci.2018.05.019
3. Antão, P., Calderón, M., Puig, M., Michail, A., Wooldridge, C., Darbra, M.R.: Identification of occupational health, safety, security (ohhs) and enviromental performance indicators in port areas. Saf. Sci. **85**, 266–275 (2016)
4. Blanchard, D.: A smarter way to safety (2019). https://www.ehstoday.com/safety-technology/article/21920103/a-smarter-way-to-safety, (last accessed 10 January 2022)
5. Cioni, M., Sabioli, M.: A survey on semi-supervised feature selection methods. Work Employ Soc. **30**, 858–875 (2016)
6. Commission, E.: Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions (2012). https://www.eea.europa.eu/policy-documents/communication-from-the-commission-to-1, (last accessed 20 January 2022)
7. Dy, J.G., Brodley, C.E.: Feature selection for unsupervised learning. J. Mach. Learn. Res. **5**, 845–889 (2004)
8. Ferrara, E., Meo, P.D., Fiumara, G., Baumgartner, R.: Web data extraction, applications and techniques: a survey. Knowl. Based Syst. **70**, 301–323 (2014)
9. Ferreira, A.J., Figueiredo, M.A.: Efficient feature selection filters for high-dimensional data. Pattern Recog. Lett. **33**, 1794–1804 (2012). https://doi.org/10.1016/j.patrec.2012.05.019
10. Freitas, L.: Manual de segurança e saúde do trabalho. Sílabo (2016)
11. Irmak, U., Suel, T.: Interactive wrapper generation with minimal user effort. In: Proceedings of the 15th international conference on World Wide Web - WWW 2006, p. 553. ACM Press (2006). https://doi.org/10.1145/1135777.1135859
12. Jiang, L., Lavaysse, L.M., Probst, T.M.: Safety climate and safety outcomes: a meta-analytic comparison of universal vs. industry-specific safety climate predictive validity. Work Stress **33**, 189–214 (2019)
13. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997). https://doi.org/10.1016/s0004-3702(97)00043-x, https://www.sciencedirect.com/science/article/pii/S000437029700043X?pes=vor
14. Scikit Learn.: scikit-learn: Machine Learning in python (2022). https://scikit-learn.org/stable/index.html. Accessed 4 Apr 2022
15. Li, J., et al.: Feature selection: a data perspective. ACM Comput. Surv. 50, 1–45 (2017). https://doi.org/10.1145/3136625, https://arxiv.org/abs/1601.07996
16. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman and Hall/CRC. October 2007
17. NOAA: National oceanic and atmospheric administration, National Oceanic and Atmospheric Administration
18. Omondi, A., Lukandu, I.A., Wanyembi, G.W.: A monte carlo-based search strategy for dimensionality reduction in performance tuning parameters. J. AI Data Min. **8**, 471–480 (2020)

19. Rose, N., Dolega, L.: It's the weather: quantifying the impact of weather on retail sales. Appli. Spat. Anal. Policy **15**, 189–214 (2022). https://doi.org/10.1007/s12061-021-09397-0
20. Sheikhpour, R., Sarram, M.A., Gharaghani, S., Chahooki, M.A.Z.: A survey on semi-supervised feature selection methods. Pattern Recogn. **64**, 141–158 (2017)
21. Song, L., Smola, A., Gretton, A., Borgwardt, K.M., Bedo, J.: Supervised feature selection via dependence estimation. In: Proceedings of the 24th International Conference On Machine Learning - ICML 2007, pp. 823–830. ACM Press (2007)
22. Sánchez-Maroño, N., Alonso-Betanzos, A., Tombilla-Sanromán, M.: Filter methods for feature selection - a comparative study. Intell. Data Eng. Auto. Learn. IDEAL **2007**, 178–187 (2007)
23. Weston, J., Elisseff, A., Schoelkopf, B., Tipping, M.: Use of the zero-norm with linear models and kernel methods andré elisseeff bernhard schölkopf mike tipping. J. Mach. Learn. Res. **3**, 1439–1461 (2003)