Task 1 - Data mining for Energy Forecasting

Introduction

Energy forecasting plays a crucial role in the efficient operation and planning of power systems. Accurate forecasts are important tools used in the power industry for predicting the amount of electricity needed to meet consumer demand. They help utilities to optimize energy supply and demand, reduce costs, and improve reliability.

Since the 1980's, electrical power systems have become increasingly complex and have generated large amounts of data. This complexity is due to the continuous advancements of information and communication technologies, and has been further exacerbated by factors such as market deregulation, the integration of renewable energy sources, as well as the extended use of SCADAs and WAMs for monitoring [1]. As renewable energy resources (RES) become more prevalent in power systems, the need for accurate energy forecasting increases. Thus, it is of vital importance for the Grid operators and decision-makers to have information about the amount of power RES will generate in the coming hours and days. [2]. As power generation shifts from traditional coal-fired plants to greener sources, such as renewable energy systems, the power grid has become more complex. This increase in complexity requires the integration of advanced information and communication technologies (ICTs) into the grid, known as the Smart Grid. As a result, accurate forecasting of demand and supply is now an essential component of these necessary ICTs [3].

In the same manner, renewable energy sources like wind and photovoltaic power plants generate electricity based on the weather. So, by nature, the power they generate is unpredictable and volatile. This creates difficulties in balancing electricity demand and supply. Hence, accurate forecasting of renewable energy output and electrical load is crucial for planning and scheduling the grid. Forecasts provide insight into future volatile renewable power generation and electrical load, which helps to ensure a stable and reliable supply of electricity to consumers [3].

Forecast information is critical not only for electrical grids but also for natural gas and district heating grids since the stability and good performance of these grids also rely on accurate load forecasts. In addition, excess renewable electricity can be converted into gas or heat ('Power to Gas' and 'Power to Heat' technologies), which can be fed to gas and district heating grids, respectively. This integration makes it even more important to have precise volatile renewable power forecasts because energy utility companies, particularly those operating in competitive energy markets, rely heavily on load forecasts to make informed decisions, including electricity generation, purchasing, and infrastructure planning [3].

In the same context as the evolution of power grid into smart grid, the traditional city has been transformed into the smart city. Cities are like dynamic living organisms and they constantly evolve. So, the smart city is an innovation of the physical city with high

integration of advanced monitoring, sensing, communication, and control technologies, aiming to provide real-time, interactive, and intelligent services to citizens. A city is a complex system to operate, and new methods are required to manage it and use the massive amounts of data it generates. City administrations can gain knowledge that is hidden in large-scale data to provide better urban governance and management by applying Information and Communication Technologies (ICT) solutions. For instance, ICT solutions can enable better transport planning, efficient water management, new energy efficiency strategies, improved waste management, and effective risk management policies for the city users. Moreover, other important aspects of urban life, such as public health, air quality, and pollution, and public security, can also benefit from these ICT solutions [4].

Taking into account the evolution of the energy sector in general, data acquirement is the most critical part in all aspects of the energy forecasting field. Data can be tapped from city-wide sites, such as power grid status, transportation grid status, vehicular networks, locations of emergency service providers, and the size of crowds in locations throughout the region [4].

In order to cope with the overwhelming amount of information, Data Mining (DM) has emerged as a knowledge discovery approach that uses convenient and versatile methods to perform effective data processing, find patterns, meet correlations and make knowledge inference [1]. However, the acquired data are highly noisy and redundant in most cases thus making the systematic use of Data Mining (DM) and Machine Learning (ML) techniques all the more important since they can facilitate processing by extracting only relevant information. Compared to traditional processing methods, ML techniques provide some distinct advantages in the extraction and release of big data services. Moreover, advanced techniques like, Deep Learning (DL) and Reinforcement Learning (RL) can achieve high data rate and precision [4].

Several applications in different areas of the power systems field have arisen within the DM framework. Moreover, in the last years the pursue for Smart Grids, Advanced Metering Infrastructure (AMI), novel sensing and ICT technology, smart cities, the Electric Power Internet of Things (EPIoT), electrified transportation, energy blockchain among others, have significantly challenged the paradigms of the power industry regarding the knowledge extraction from continuously growing databases [1].

The ever-increasing number of energy-related forecasting tools and services available is reflective of the importance of energy forecasts. However, the sheer volume of data mining tools available makes understanding the differences among them a complicated task [3]. To address this challenge, the present survey attempts to briefly describe the state of the art as far as Data Mining tools and techniques for energy forecasting are concerned.

Background - Methods

In the second half of the 1990s, the electrical sector underwent an in depth restructuring and renewable energy generation was integrated into the system leading to an increase in the size and complexity of electrical infrastructures and databases and thus making crucial the enhancement of AI approaches in order to systematize knowledge acquisition. This

prompted researchers in power systems to turn to Data Mining (DM), also known as Knowledge Discovery from Data. As a result, DM proved to be highly useful in power systems to handle their vast scale nature (huge datasets and vast amount of state variables), complex statistical analysis (combining deterministic and stochastic events), mixture of discrete and analog information, variable temporal character (from milliseconds to years), effective visualization requirements, need of rapid decision-making and highly-uncertain data handling. A timeline of the DM unfolding can be seen in Fig. 1 [1].

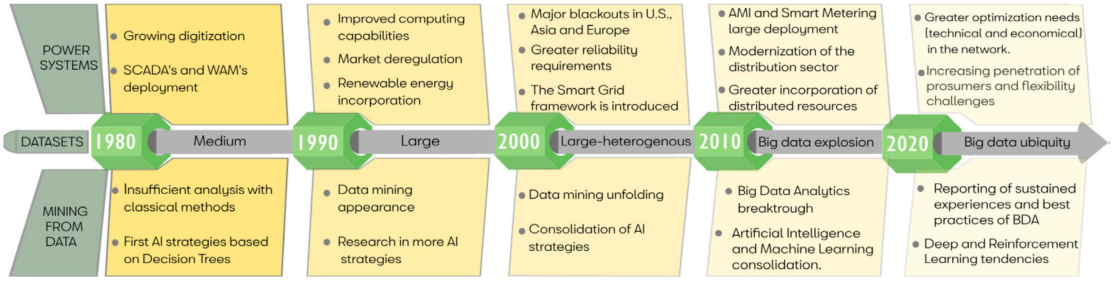


Fig 1. A timeline of data mining in power systems [1].

DM was conceived as to achieve three major goals: description, prediction and prescription [1].

For description, DM techniques are used to gain an understanding of systems by revealing patterns and relationships in datasets. In this case, segmentation, clustering and dimensionality-reduction tools such as k-means, density-based or Principal Components Analysis (PCA) are normally employed for this purpose with unsupervised manners [1].

Subsequently, regarding prediction, the mining tools carry out an induction process on data to go beyond classification, regression or estimation and then infer unknown values future variables under study. For instance, in the first years of this century, advanced prediction approaches have become increasingly important due to major blackouts and the smart grid framework introduction which needed enhanced ancillary services and higher reliability standards. This prediction stage is commonly conducted with ML supervised approaches and nowadays also with cutting-edge Reinforcement Learning and novel Artificial Neural Networks (ANN) mechanisms rebranded as Deep Learning [1].

On the other hand, to perform prescriptive analytics and when needed, wisely influence the future outcomes with suitable course of actions, the previous techniques are leveraged with the inclusion of rich broad-context structured and unstructured Big Data so that it is ensured that the best possible deductions have been determined and can be applied in a self-healing closed-loop manner. This is not simple or possible to fully automate and therefore can need different degrees of human intervention and validation, specifically for critical duties which is for instance the case of dynamic security assessment in power systems. All the previous perspectives have been employed by the power systems community in a large variety of ambits going from system security and stability, passing through expansion planning, and arriving to monitoring and visualization issues [1].

In the last two decades, the DM process and its different methods have been largely employed in the power systems scope. Nevertheless, ANNs with ML purposes have been

preferred due to their noticeable capabilities to learn from training, classify patterns and perform feature extraction. After this strategy, DTs, fuzzy systems, statistical analysis (SA) and rough set theory have also been consistently employed in that order [1].

Forecasting models can be separated into three different types:

- White-box models: These models use known relations, expert knowledge, and so on (e.g., physical models for volatile renewable power forecasting) to define the relation between the utilized inputs and the future of the time series of interest.
- Black-box models: These models infer the relation between used inputs and future time series values through the application of data mining techniques on available data.
- Gray-box models: These models are a combination of white and black-box models [3].

Figure 2 depicts the types of time series forecasting models [3]. As Mikut and Reischl state data mining tools play a crucial part in the forecasting of energy consumption [5].

Furthermore, soft-computing techniques have brought new developments in artificial intelligence-based forecasting models, which outperform traditional physical methods and statistical approaches due to their data-mining and feature-extracting abilities. These models frequently use support vector machines, artificial neural networks, extreme learning machines, and adaptive fuzzy neuron networks to handle the nonlinear relationship between input and output via error minimization [6].

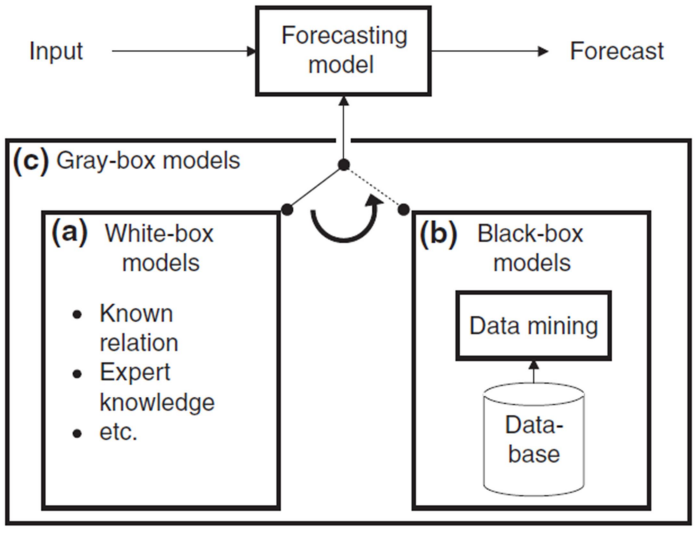


Fig 2. Different forecasting model types [3].

Moreover, Deep learning, as a promising branch of machine learning, has attracted much attention in recent years due to three major attributes, i.e., unsupervised feature learning, strong generalization capability and big-data training, compared to shallow models. It is naturally a kind of alternative of shallow models and has been extensively implemented in pattern recognition, image processing, fault detection, classification and forecasting tasks [6].

Over the past few decades, different methods for forecasting energy production, distribution and consumption had been utilized. To predict energy consumption, researchers have employed several techniques which include both traditional methods such as regression, time series, statistical methods and soft computing techniques such as Artificial Neural Networks (ANNs), Support Vector Machine (SVM), fuzzy logic, and Grey prediction [2].

Regarding the smart city concept, the production of analytics can lead to advanced insights, a better understanding of city phenomena, and supports the design of evidence-based urban strategies and innovation. Searching for useful patterns and correlations in the public-service facilities of developed cities using a DM approach has gradually become a significant area of research. The extracted patterns can be used to plan layouts or arrange new facilities in cities. Advancements of big DM technologies can support, explore and discover environmental and societal changes, including how people go about their life, behavior, and preferences; social trends, and public opinion [4].

In this context, DM and ML are vital technologies for data-centric applications for smart cities, thus several techniques have been employed for this manner such as Bag of Words (BoW) and Principal Component Analysis (PCA) for data preprocessing as well as numerous machine learning algorithms for classification, including Decision Trees, Random Forest, Support Vector Machine (SVM), Bayesian Network (BN), K Nearest Neighbors (KNN), Artificial Neural Network (ANN), Deep Learning (ANNs with many hidden perceptron layers) and Reinforcement learning (RL). Moreover, the emerging Deep Reinforcement Learning (DRL) can be considered as a promising technology, which takes a long-term goal into account and can generate optimal control actions to time-variant dynamic systems. As far as forecasting is concerned, models based on time series forecasting analysis are used like the exponential smoothing state-space model (ETS) and the Auto-Regressive Integrated Moving Average (ARIMA). Furthermore, regression analysis is also a prominent part of forecasting with Linear Regression (LinR) and Logistic regression (LogR) being the most popular methods as well as Support Vector Regression (SVR), which complements the linear regression method in case of data with a large number of features. In the case of unsupervised learning K-means and the more advanced Self-Organizing Map (SOM) method are used. Finally, Association Rules (AR) is one of the most popular methods within the context of extracting relationships among items hidden within data sets, as it has been used in several smart city applications [4].

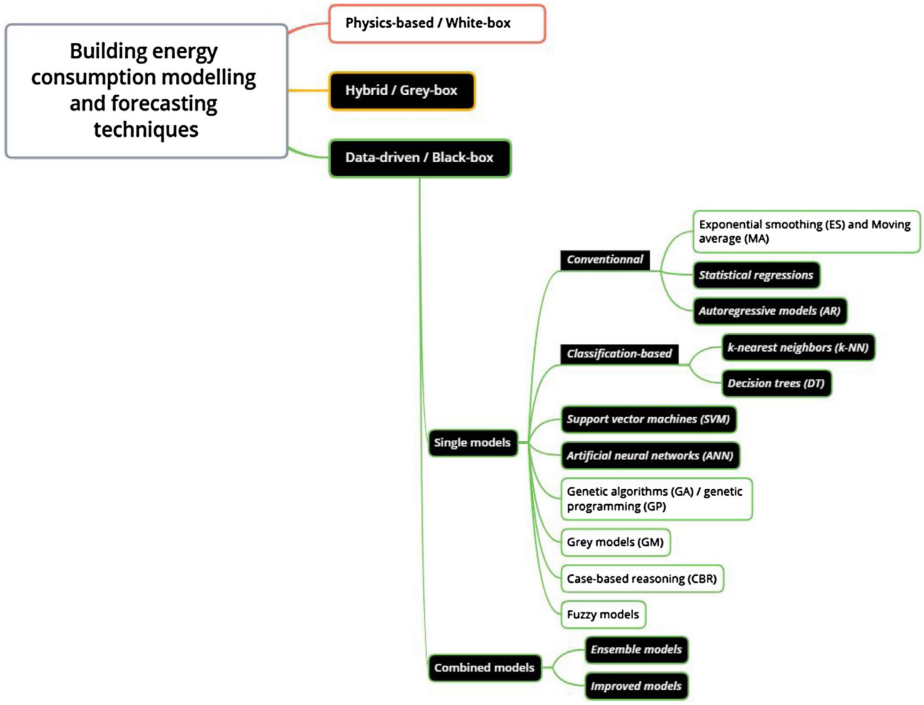


Fig. 3 Summary classification of building energy consumption modeling and forecasting methods [7].

Building energy consumption forecasting plays an equally essential role in the general energy forecasting framework. Among the three main approaches in building energy consumption modeling and forecasting (BECMF) – physics-based, data-driven and hybrid models, data-driven techniques are considered to be the most suitable for integrating buildings into smart environments. The most prevalent techniques for building energy consumption forecasting use single and combined models. Single models, include methods such as autoregressive models and statistical regressions, as well as classification-based methods like k-nearest neighbors (k-NN) and decision trees (DT), support vector machines (SVM), and artificial neural networks (ANN). On the other hand, combined modeling framework either combines several single algorithms together (ensemble models) or combines them with optimization methods (improved models). The classification of building energy consumption modeling and forecasting methods is shown in Fig. 3 [7].

In BECMF studies the aim of data-driven techniques is to create models that examine the relationships between a combination of inputs and outputs under a specific process. The model is then used to forecast building energy consumption or power load demand. Model outputs are always known since they are the target of the whole study. However, there are several strategies concerning the utilization of input variables and the extraction of features from an input dataset to train data-driven models. There are two primary approaches for selecting input variables in BECMF: supervised and unsupervised learning. Moreover, other tasks exist such as reinforcement and transfer learning [7].

Finally, it is important to mention once more that DM solutions have proven successful in situations where the use of conventional model-driven approaches, based on physics-based methods, have been unable to alone handle the influx of new information, the novel technical challenges and limitations of previous tactics. In this respect, Engineers have been

proposing data-driven solutions to take advantage of the newly generated data. These solutions offer new modeling alternatives that view data as actionable and testable knowledge, rather than as isolated silos of information. Under this trend, mapping rules and relationships are derived by extracting patterns and knowledge from historical measurements or simulations. These data-driven alternatives have proven to be highly effective, providing strong decision-making capabilities. Be that as it may, model-driven and data-driven techniques should be combined in a systematic manner to complement each other's benefits and drawbacks in a productive way [1].

Applications – Case Studies

First of all, it is worth emphasizing that data mining is not a lineal process but an iterative one. Besides, there is no matching of specific techniques for a given application as the literature reveals. Even after a method has been selected, it is common to recurrently adjust the selected strategy or even use other complementary DM tools to gain different perspectives on the data and enhance results [1]. Under this perspective, several case studies are presented from a variety of fields concerning energy forecasting in general.

For instance, in energy and power applications, anomaly detection emerges as an important aspect in fields like electric load forecasting and energy production forecasting. So, Luo et al. implemented a model-based anomaly detection method for very short-term load forecasting. The method includes two components, an underlying model i.e. dynamic regression model (DRM) and an adaptive anomaly threshold [2]. Additionally, Zhao et al. proposed a fault detection method based on pattern recognition for chiller, which transformed the fault detection problem into a data description problem. However, a major limitation of these methods is that the feasibility depends on the accuracy of feature labels and the effectiveness of classification methods [8]. Respectively, Lei et al recognize the potential in using data mining technology for anomaly detection since great amounts of energy consumption data are collected by building energy monitoring platforms (BEMS). However, due to the fact that data recording equipment and transmission channel are affected by malfunctions and weather, as well as the incomplete energy consumption monitoring technology platform, the detection of abnormal energy consumption data is not always accurate. To reduce resource loss, it is necessary to detect anomalies from the data and fix the faults behind to realize energy-saving. So, the authors presented a dynamic anomaly detection algorithm that integrates unsupervised clustering algorithm with a supervised algorithm to establish a semi-supervised matching mechanism, which avoids the influence of error label and improves the efficiency of anomaly detection. Moreover, a particle swarm optimization (PSO) is used to optimize the unsupervised clustering algorithm. The complete anomaly detection process is shown in Fig. 4 [9].

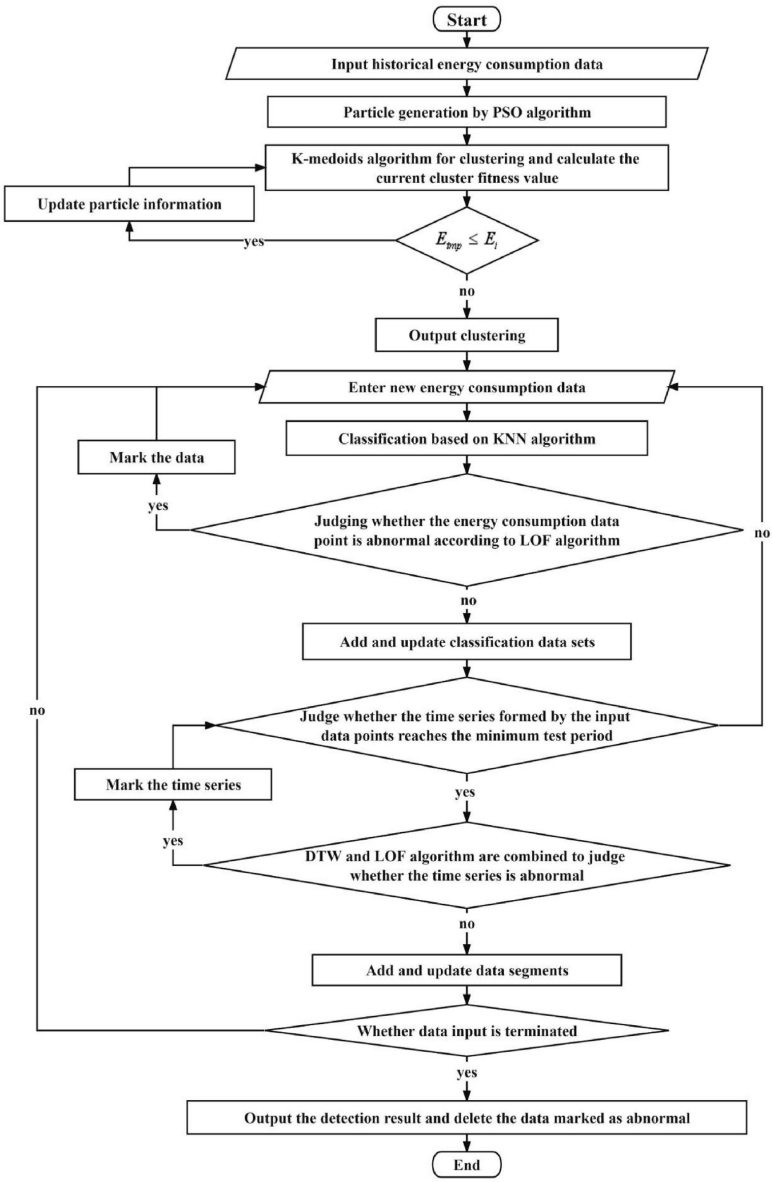


Fig. 4 Overall flow chart of anomaly detection [9].

As far as long term load forecasting is concerned, Kazemzadeh et al proposed a hybrid method based on time series or ARIMA method, ANN method, and the developed PSO-SVR forecasting method. A major advantage of the proposed hybrid method is the combination of analytic time series method and the data mining approach which can handle the nonlinearity and seasonal trends in input samples. Since the ARIMA model as the major forecasting tool is very sensitive to the noise and seasonal trends in time series, the authors proposed, a data mining method based on the SVR algorithm to forecast the long term load and energy demands [10].

In the case of renewable energy, various algorithms have been reported in the literature to provide accurate predictions for the next few minutes to the next few days. A hybrid of mixed data sampling regression and back propagation neural network was developed to perform real-time forecasting of carbon prices in Shenzhen, resulting in a better performance [11]. Considering the instability and randomness of PV power output, Li et al.

proposed the combination of Particle Swarm Optimization (PSO) and Deep Belief Network (DBN) for short-term forecast of PV power generation. Moreover, by using PSO to optimize the stochastic parameters of DBN and training the similar days screened by the gray correlation method as the training samples of the model, they were able to build a PV power generation forecasting model, which surpasses the traditional DBN neural network with higher prediction accuracy [12]. Zhang et al. compared multiple MLM models by analyzing historical PV power generation data and established ANN and nonlinear fitting prediction models based on the effects of multiple meteorological factors on PV output power, which was helpful to reduce prediction errors [13].

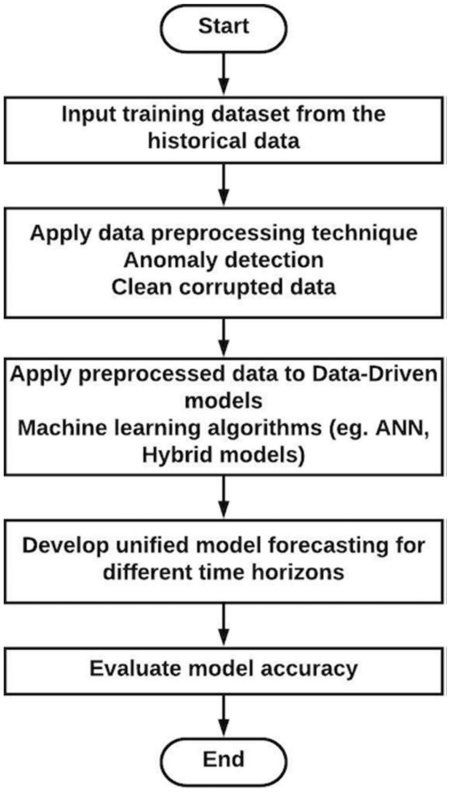


Fig. 5 Flowchart of forecasting process based on predictive data mining techniques [2].

As stated previously, there is a strong dependency between renewable energy and weather. So, Wang et al proposed a weather classification model for day ahead photovoltaic power forecasting based on generative adversarial networks and convolutional neural network, and it was found that weather classification plays a decisive role in determining the most efficient photovoltaic power forecasting model [6]. On the other hand, E. Sharma proposed a weather-free forecasting model created using a database containing relevant information about past produced power data and data mining techniques. The idea of using a weather-free data-driven model is first to alleviate the dependence on weather data which, in some scenarios is difficult to obtain and second to reduce the computational effort. Figure 5 presents the flowchart of the proposed forecasting process [2].

Deng et al. compared ANN, Support Vector Regression (SVR), Random Forest (RF) and Gradient Boosting (GB) for forecasting Energy Use Intensity (EUI) of US commercial office buildings and the individual energy end-uses of HVAC, lighting, and plug loads, based on

Commercial Building Energy Consumption Survey (CBECS) microdata. Based on their results, SVR and RF showed better accuracy in BEPF comparing to ANN and GB [14]. Ascione et al. compared two families of feed forward MLP ANNs on the outcomes of E+ simulations as targets for networks' training and testing. The two ANNs addressed an existing commercial building stock, and a renovated stock with energy retrofit measures (ERMs), and both showed regression coefficients above 0.95 [15]. Ma and Cheng used Multiple Linear Regression (MLR), ANN, and SVR, as well as a GIS-integrated data mining framework for estimating the Annual EUI values of 3640 residential buildings in New York City. They also compared their regression method to feature selection and algorithm optimization, based on 216 prepared features. Their results showed higher accuracy of the SVR-based method, comparing to the others [16]. Capozzoli et al. compared (MLR) and Classification and Regression Tree (CART) performance in predicting annual heating and cooling demands of educational buildings, while both approaches provided a good level of accuracy. Their results showed that heating energy consumption of the selected school buildings was mostly influenced by the gross heated volume, heat transfer surfaces, boiler size, and thermal transmittance of their windows [17].

Other studies categorized buildings' functionality, characteristics, and consumption patterns, in order to efficiently forecast urban building energy performance. An et al. used clustering and statistical analyses to represent AC use patterns for over 300 residential buildings in Zhengzhou, China, using building Key Performance Indicators (KPI) [18]. Tardioli et al. used RF, k-means Cluster Analysis (CA), and Principal Component Analysis PCA, as well as data preprocessing and predictive modeling, in order to identify representative buildings and building groups in an urban dataset including residential and commercial buildings [19]. Papadopoulos et al. implemented an unsupervised learning method with k-means clustering to optimally cluster energy consumption time series of a mixed set of commercial and residential buildings. They also have used supervised learning methods to understand how building characteristics vary between clusters, which resulted in identifying buildings with similar temporal energy performance patterns and their shared characteristics [20]. Sokol et al. used BN to develop a new methodology in defining different arch-types in urban building energy models. They defined unknown or uncertain parameters in archetype descriptions as probability distributions and used measured energy data to update these distributions by Bayesian calibration [21]. Ma et al. used k-means clustering to forecast daily heating load profiles of higher educational buildings, identifying typical daily load profiles and classifying buildings based on these profiles. Pearson correlation coefficient was used to determine a dissimilarity measure to classify the daily load profiles according to variation similarity instead of magnitude similarity [22]. Yang et al. used k-means and Dynamic Time Warping (DTW) clustering to cluster building energy consumption patterns of 10 educational buildings. Their results showed higher accuracy levels for k-means versus DTW clustering [23]. Hsu compared a cluster-wise regression to common two-stage algorithms that use K-means and model-based clustering with linear regression, in order to predict annual urban building energy consumption. Their results showed that K-means method gives more stable clusters when the correct number of clusters is chosen [17].

Last but not least, electricity customer characterization is an important sector in energy forecasting since liberalization of electricity markets has resulted in the emergence of new

players, increasing the competitiveness in the markets, standing those who can provide better services for better prices. As a result, the knowledge of energy consumers' profile has been an important tool to help players to make decisions in the electrical sectors. In this respect, Ramos et al. proposed a characterization framework based on clustering and classification to achieve the characterization of low voltage electricity customers. The purpose of their study is to find a tariff that can better represent the grid utilization, by increasing energy prices when the energy consumption is high and reducing the price when the consumption is low. This gives the economic sign for customers to reduce energy consumption when the price is high, trying to improve the grid utilization. The authors studied nine clustering algorithms including agglomerative hierarchical clustering, K-means, Diana, Partitioning Around Medoids, Clara, Fanny, Self-Organizing Maps (SOM), Model-based clustering and SOTA and concluded that the SOM algorithm proved to be the best, compared with K-means and K-medoids algorithm [24].

Conclusions

This literature survey attempted to introduce the concepts of data mining as far as energy forecasting is concerned. Subsequently, a minor background as well as several methods and approaches that are used in energy forecasting and its various applications have been presented. Last but not least, several case studies have been showcased separated by application in the general energy forecasting framework.

Taking the aforementioned information into account, the field of data mining for energy forecasting has grown in importance and complexity due to the increasing need for accurate energy predictions, driven by market deregulation, the incorporation of renewable energy sources, and the extended insertion of SCADAs and WAMs. In this respect, to take advantage of the vast amounts of newly generated data, data-driven solutions have emerged, that treat data as actionable knowledge rather than isolated silos of information. By incorporating robust tools and techniques, such solutions leverage the potential of Big Data to extract valuable patterns, correlations, and insights. This approach enables knowledge discovery from the overwhelming flood of heterogeneous information. Therefore, new modeling alternatives are being developed that rely on the incorporation of such data-driven solutions making the role of the different DM and ML approaches increasingly noticeable [1].

In addition, the Data Mining field is no longer exclusively available to large corporations. Nowadays, it is common practice for practitioners and engineers to access free, ready-to-use libraries and tools from popular programming languages and software. As a result, Big Data Analytics (BDA) has emerged as the backbone framework in power networks, resulting in the first sustained deployments and experiences with real-world applications that have already yielded successful outcomes [1].

Despite the significant progress made in the field of data mining for energy forecasting, there are still several challenges that need to be addressed. One of the major challenges is the availability of data. In many cases, data is scarce or of poor quality, which limits the accuracy of forecasting. Another challenge is the complexity of energy systems, which requires the integration of multiple data sources and the use of advanced algorithms.

Be that as it may, data mining methods have proved to be effective in energy forecasting, and have played a significant role in improving the accuracy of predictions. The case studies presented in this paper provide evidence of the success of various techniques, and demonstrate their potential to address current and future challenges in the field, with the use of methods and approaches ranging from traditional techniques to hybrid methods. However, there is still much work to be done, and the development of new algorithms and techniques, as well as the integration of advanced technologies such as machine learning and artificial intelligence, will be essential in the ongoing evolution of data mining for energy forecasting. Overall, data mining methods are poised to continue making significant contributions to the field of energy forecasting, and have the potential to transform the way we think about and manage energy consumption.

Bibliography

[1] Xavier Dominguez a, Alvaro Prado, Pablo Arboleya, Vladimir Terzija, 2023. Evolution of knowledge mining from data in power systems: The Big Data Analytics breakthrough. Electric Power Systems Research 218

[2] Ekanki Sharma, 2018. Energy forecasting based on predictive data mining techniques in smart energy grids. 7th DACH+ Conference on Energy Informatics

[3] Jorge A. González Ordiano, Simon Waczowicz, Veit Hagenmeyer, Ralf Mikut, 2018. Energy forecasting tools and services. WIREs Data Mining Knowl Discov 2018, 8:e1235. doi: 10.1002/widm.1235

[4] Anestis Kousis, Christos Tjortjis, 2021. Data Mining Algorithms for Smart Cities - A Bibliometric Analysis. Algorithms 2021, 14, 242. https://doi.org/10.3390/a14080242

[5] Ralf Mikut, Markus Reischl, 2011. Data mining tools. WIREs Data Mining Knowl Discov 2011 1 431-443 DOI: 10.1002/widm.24

[6] Huaizhi Wanga, Zhenxing Leia, Xian Zhangb, Bin Zhouc, Jianchun Penga, 2019. A review of deep learning for renewable energy forecasting. Energy Conversion and Management 198 (2019) 111799.

[7] Mathieu Bourdeaua, Xiao qiang Zhaia, Elyes Nefzaouib, Xiaofeng Guob, Patrice Chatellierd, 2019. Modeling and forecasting building energy consumption - A review of data-driven techniques. Sustainable Cities and Society 48 (2019) 101533.

[8] Yang Zhao, Shengwei Wang, Fu Xiao, 2012. Pattern recognition-based chillers fault detection method using Support Vector Data Description (SVDD). Applied Energy 112 (2013) 1041–1048.

[9] Lei Lei, Bing Wu, Xin Fang, Li Chen, Hao Wu, Wei Liu, 2022. A dynamic anomaly detection method of building energy consumption based on data mining technology. Energy 263 (2023) 125575.

[10] Mohammad-Rasool Kazemzadeh , Ali Amjadian , Turaj Amraee, 2020. A hybrid data mining driven algorithm for long term electric peak load and energy demand forecasting. Energy 204 (2020) 117948.

[11] Meng Han, Lili Ding, Xin Zhao, Wanglin Kang, 2019. Forecasting carbon prices in the Shenzhen market, China: The role of mixed-frequency factors. Energy 171 (2019) 69-76.

[12] Wenyong Zhang, Qingwei Li, Qifeng He, 2022. Application of machine learning methods in photovoltaic output power prediction - A review. Journal of Renewable and Sustainable Energy 1 March 2022; 14 (2): 022701. https://doi.org/10.1063/5.0082629.

[13] Zhang, Jy., Wang, Z., Zhang, Xh, 2020. Research on photovoltaic output power short term prediction method based on machine learning. Energy Syst (2020). https://doi.org/10.1007/s12667-020-00386-9.

[14] Hengfang Denga, David Fannona, Matthew J. Eckelmana, 2018. Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy and Buildings 163 (2018) 34–43.

[15] Fabrizio Ascione, Nicola Bianco, Claudio De Stasio, Gerardo Maria Mauro, Giuseppe Peter Vanoli, 2017. Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. Energy 118 (2017) 999-1017.

[16] Jun Ma, Jack C.P. Cheng, 2016. Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology. Applied Energy 183 (2016) 182–192.

[17] Soheil Fathi, Ravi Srinivasan, Andriel Fenner, Sahand Fathi, 2020. Machine learning applications in urban building energy performance forecasting - A systematic review. Renewable and Sustainable Energy Reviews 133 (2020) 110287.

[18] Jingjing An, Da Yan, Tianzhen Hong, 2018. Clustering and statistical analyses of air-conditioning intensity and use patterns in residential buildings. Energy & Buildings 174 (2018) 214–227

[19] Giovanni Tardiolia, Ruth Kerrigana, Mike Oatesa, James O'Donnellb, Donal P. Finn, 2018. Identification of representative buildings and building groups in urban datasets using a novel preprocessing, classification, clustering and predictive modelling approach. Building and Environment 140 (2018) 90–106.

[20] Sokratis Papadopoulos, Bartosz Bonczak, Constantine E. Kontokosta, 2018. Pattern recognition in building energy performance over time using energy benchmarking data. Applied Energy 221 (2018) 576–586.

[21] Sokol Juliaa, Cerezo Davila Carlos, Reinhart Christoph F., 2017. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. Energy and Buildings 134 (2017) 11–24.

[22] Zhenjun Ma, Rui Yan, Natasa Nord, 2017. A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. Energy 134 (2017) 90-102.

[23] Junjing Yang, Chao Ning, Chirag Deb, Fan Zhang, David Cheong, Siew Eang Lee, Chandra Sekhar, Kwok Wai Thamk, 2017. Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy and Buildings 146 (2017) 27–37.

[24] Sérgio Ramos, João Soares, Samuel S. Cembranel, Inês Tavares, Z. Foroozandeh, Zita Vale, Rubipiara Fernandes, 2021. Data mining techniques for electricity customer characterization. Procedia Computer Science 186 (2021) 475–488.

Task 2 - World Energy Dataset 1980-2019

Data Loading: We began by loading our dataset, which contains information about CO_2 emissions, energy consumption, production, GDP, population, energy intensity per capita and energy intensity by GDP for different countries from 1980 to 2019.

Data Preprocessing: First, we checked the data for duplicates to avoid biasing the performance of the analysis and machine learning model by giving more weight to duplicate information. We handled missing values in the dataset. We dropped rows with missing values since they could potentially distort our analysis and predictions. We applied Outlier Detection and Removal Using Z-Score method. The Z-score is a measure of how many standard deviations an element has from the mean. It is a common method to detect and remove outliers in a dataset. In our data set every data point that has a z-score greater than or less than a threshold -4 and +4 can be considered an outlier. In our case we did not remove outliers because they were few and we judged them to be important for the performance of the models.

Feature Engineering (Continent column): Adding a 'Continent' column to our dataset became helpful for our analysis and model because there are regional patterns in our data. Countries within the same continent might have similar CO_2 emissions, energy consumption, GDP, or population growth rates due to geographical, economical, or policy similarities.

The above data preprocessing and feature engineering process was so crucial in our data analysis that greatly affects the quality of our insights or predictions.

Exploratory Data Analysis (EDA): We visualized the data to understand the patterns and relationships in it. We used seaborn, matplotlib and plotly libraries for creating various plots like scatter plots, line charts and bar plots. We also calculated the correlation between different variables. Thereinafter the initial plots to better understand our data set are described:

1. A pairplot that provides a high-level overview and first-hand inspection of the relationships and distributions worldwide. For more detailed and precise statistics, we will need to conduct further statistical tests or use more specialized visualizations.
2. A line chart visualization that provides a clear view of the CO_2 emissions trend for each continent over time. It helps to understand how the CO_2 emissions have changed year-by-year on different continents. All continents have shown an increase in CO_2 emissions over the years, with Asia showing the highest increase in terms of absolute numbers. This trend is a significant concern given the role of CO_2 in climate change.
3. A stacked bar plot showing the total CO_2 emissions for each continent, grouped by different energy types. The resulting plot provides a visual representation of the total CO_2 emissions by energy type and continent, allowing you to compare the emissions across different energy types and see the contributions of each continent to the overall emissions. Asia is leading in terms of total CO_2 emissions from all

energy types, with the highest contribution coming from coal. On the other hand, Oceania emits the least amount of CO_2 among the given continents. The data shows a prevalent dependence on fossil fuels (coal, natural gas, and petroleum) across all continents, with no recorded CO_2 emissions from nuclear energy or renewables and other energy sources.

4. A bar chart that displays the country with the maximum CO_2 emissions for each energy type. The plot provides a visual comparison of the countries with the highest CO_2 emissions for different energy types, allowing you to identify the dominant contributors to CO_2 emissions in each energy category. In summary, the United States has the highest total emissions among the countries listed, with significant contributions from both petroleum and other liquids as well as natural gas. China's emissions, on the other hand, are primarily from coal, making it the country with the highest CO_2 emissions from this energy source.
5. A heatmap that provides a visual summary of the correlation structure among the numeric features in the DataFrame, helping to identify potential associations and dependencies between the variables.

Here are some key observations from the data:

- Energy Consumption and Energy Production: There is a very high positive correlation (0.993731), which suggests that as energy consumption increases, energy production also tends to increase. This makes intuitive sense, as higher consumption often necessitates higher production.
- GDP and Population: There is a high positive correlation (0.946185) between GDP and population. This typically means that as a country's population increases, its GDP also tends to increase, likely due to increased economic activity with more individuals.
- CO_2 Emission and Energy Consumption/Production: CO_2 emissions show a very high correlation with both energy consumption (0.982769) and energy production (0.975663). This suggests that as energy consumption and production increase, CO_2 emissions also tend to rise, likely due to the fact that much of our energy still comes from carbon-intensive sources, since renewable energy has not got yet the capacity to fully replace legacy technologies.

Data Mining: We performed various data mining methods to extract useful information from our dataset including (Regression, Classification, Clustering and Association rules).

1) We performed a random forest regression analysis based on selected features to provide information about the importance of each feature in CO_2 emissions. These importances represent the relative contribution of each feature to the prediction task as determined by the trained Random Forest Regressor model. It indicates that "Energy production" 41.5% and "Energy consumption" 56.1% are the most important features for predicting CO_2 emissions, while "Population" and "GDP" have comparatively less influence.

2) We use a data mining workflow for classification using a decision tree and random forest classifier. It filters the dataset, performs data preprocessing, trains the classifier, predicts CO_2 emission categories, evaluates the model's performance, and provides visualizations of feature importances and confusion matrix.

Decision Tree Classifier:

Accuracy: The accuracy of the model is 0.9551, indicating that it correctly predicts the CO_2 emission category for approximately 95.51% of the instances in the test set.

Classification Report: The classification report provides a detailed assessment of the model's performance for each CO2 emission category.

Metrics: For the 'high' category, the precision, recall, and F1-score are all around 0.98, indicating high performance.

For the 'low' category, the precision is 0.90, the recall is 0.96, and the F1-score is 0.93. These metrics suggest that the model performs well but with slightly lower precision compared to the 'high' category.

For the 'medium' category, the precision is 0.77, the recall is 0.67, and the F1-score is 0.72. These metrics suggest that the model has relatively lower performance for the 'medium' category.

The macro avg F1-score is 0.88, indicating overall good performance. The weighted avg F1-score is 0.95, which takes into account class imbalance and provides an average F1-score weighted by support.

Random Forest Classifier:

Accuracy: The accuracy of the model is 0.97, indicating that it correctly predicts the CO_2 emission category for approximately 97% of the instances in the test set.

Classification Report: The classification report shows high precision, recall, and F1-score values for all categories. The metrics are consistently above 0.94 for precision, recall, and F1-score for each category, indicating strong performance.

The macro avg F1-score is 0.97, indicating excellent overall performance. The weighted avg F1-score is 0.97, taking into account class imbalance, and providing an average F1-score weighted by support.

In summary, both models perform well, but the Random Forest Classifier demonstrates slightly better overall performance compared to the Decision Tree Classifier, as reflected by higher accuracy and F1-scores.

3) We apply a data mining workflow for association rule mining using the Apriori algorithm. This model performs association rule mining on a dataset with categorical variables. It filters the dataset, discretizes the CO_2 emission column, encodes transactions, finds frequent itemsets using the Apriori algorithm, generates association rules, filters the rules based on desired relationships, and prints the filtered rules.

In summary, the results of association rule mining using the Apriori algorithm provide insights into the relationships between different energy types and the CO_2 emission category of "high." Here's a brief summary of the findings:

Energy type= petroleum_n_other_liquids -> CO2_emission=high:

The presence of the energy type "petroleum_n_other_liquids" in a transaction indicates a high likelihood (confidence of 78.33%) of also having a "high" CO_2 emission category.

This association has a strong positive relationship (lift of 1.93), meaning that the occurrence of "petroleum_n_other_liquids" is significantly related to a "high" CO_2 emission category.

Energy type=natural_gas -> $_{CO2}$_emission=high:

The presence of the energy type "natural_gas" in a transaction suggests a moderate likelihood (confidence of 46.07%) of having a "high" CO2 emission category.

This association has a positive relationship (lift of 1.14), indicating that "natural_gas" and a "high" CO_2 emission category are somewhat related.

These findings suggest that there are notable associations between specific energy types and the CO_2 emission category of "high." The presence of "petroleum_n_other_liquids" demonstrates a strong positive relationship with a "high" CO_2 emission category, while the presence of "natural_gas" indicates a moderate association. These associations can provide insights into the impact of different energy types on CO_2 emissions and can inform decision-making in energy and environmental policies.

4) Clustering analysis to group continents based on their total CO_2 emissions. This model clusters continents based on their total CO_2 emissions using the KMeans algorithm. It groups the data by continent and calculates the sum of CO_2 emissions for each continent. Then, it applies KMeans clustering to assign continents to one of the three clusters based on their CO_2 emission levels. Based on the results, the continents are grouped into three clusters:

   Cluster 0: This cluster consists of Africa, Oceania, and South America. These continents have relatively lower total CO_2 emissions compared to the other clusters.

   Cluster 1: This cluster includes Europe and North America. These continents have higher total CO_2 emissions compared to Cluster 0 but lower compared to Cluster 2.

   Cluster 2: This cluster consists of Asia. Asia has the highest total CO_2 emissions among all continents. This information can be useful for understanding global emissions trends, identifying regions with similar emission profiles, and informing policy decisions related to environmental sustainability and climate change.

5) We perform regression analysis using the XGBoost algorithm to predict CO_2 emissions based on energy consumption, energy production, GDP, and population data. This model trains an XGBoost regression model to predict CO_2 emissions and evaluates its performance using mean squared error. It then uses the trained model to predict CO_2 emissions for the year 2020 for each continent and the top 11 countries with the highest CO2 emissions. For this model we did not standardize the data because tree-based models like XGBoost are often considered scale invariant, meaning that they can handle features on different scales. This is because these models are based on a series of binary splits, so the actual scale of the features does not necessarily impact the model's performance.

The use of XGBoost for prediction in this coursework was the most appropriate because the dataset contains continuous values. Given that the target variable in this case is continuous (CO_2 emissions), using XGBoost as a regression model is a suitable choice. XGBoost is designed to handle regression problems by optimizing the model to minimize the difference between predicted and actual continuous values. It provides flexibility, strong predictive performance, and the ability to capture complex relationships in the data, making it well-

suited for this prediction task. To optimize our parameters we used GridSearchCV from scikit-learn, where the XGBoost model is passed as the estimator, the parameter grid is specified, and cross-validation is applied with 5 folds. The scoring metric is set to negative mean squared error (neg_mean_squared_error) to optimize for lower values.

```
Best Parameters:
{'learning_rate': 0.1, 'max_depth': 9, 'n_estimators': 500}
```

```
Improving MSE:197096.75906577002 to MSE: 156742.2987018007
```

Applying best parameters to our xgboost model we had much better performance reducing MSE and improving our predictions.

The most challenge part of this coursework was to select a related data set and understand it. Furthermore, the **Data Preprocessing** procedure was really confusing trying to use the best method to fill all the NaN values. Firstly we applied interpolate method:

df = data.groupby('Country').apply(lambda group: group.interpolate(method='linear', limit_direction='both')

The interpolate() method uses various interpolation techniques to fill the missing values. The default interpolation method is linear interpolation, which fills the missing values with a linearly interpolated value based on the neighboring data points. However, in the continuation of the coursework we found that the results were not as correct as using dropna() method. We realized that the most important factor in properly analyzing data and applying machine learning methods effectively is having correct and clean data. Finally, standardizing our data helped us to improve model performance, make algorithm convergence quicker and make the model more interpretable.

*All the code and plots are uploaded in an html file and Notebook ipynb.

Task 3

We used the Information Gain algorithm in order to construct the decision tree.

Income has the highest gain so it has been chosen as root node. In the case of the first internal node for income: low, age has the highest gain and subsequently a 2nd internal node is constructed (income: low and age > 40) since credit rate has the highest gain in this case.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \qquad Gain(A) = Info(D) - Info_A(D)$$

Root - Income

age	p_i	n_i	$I(p_i,n_i)$
<=30	2	3	0.971
31…40	2	2	1
>40	4	1	0.722

Info(D)	0.985
$Info_{age}(D)$	0.890
Gain(age)	0.095

income	p_i	n_i	$I(p_i,n_i)$
low	2	2	1.000
medium	5	1	0.650
high	3	1	0.811

$Info_{income}(D)$	0.796
Gain(income)	0.189

student	p_i	n_i	$I(p_i,n_i)$
yes	4	3	0.985
no	4	3	0.985

$Info_{student}(D)$	0.985
Gain(student)	0.000

credit_rt	p_i	n_i	$I(p_i,n_i)$
fair	3	5	0.954
excellent	5	1	0.650

$Info_{credit\_rt}(D)$	0.824
Gain(credit_rt)	0.161

	age	income	student	credit_rating	buys_comp
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	yes
3	31…40	high	no	fair	no
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31…40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	no
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31…40	medium	no	excellent	yes
13	31…40	high	yes	fair	no
14	>40	medium	no	excellent	yes

1st Internal Node - Income: low

age	p_i	n_i	$I(p_i,n_i)$
<=30	0	1	0
31…40	1	0	0
>40	1	1	1

Info(D)	1
$Info_{age}(D)$	0.5
Gain(age)	0.5

credit_rt	p_i	n_i	$I(p_i,n_i)$
fair	1	1	1
excellent	1	1	1

$Info_{credit\_rt}(D)$	1.000
Gain(credit_rt)	0

student	p_i	n_i	$I(p_i,n_i)$
yes	2	2	1
no	0	0	0

$Info_{student}(D)$	1
Gain(student)	0

resulting in two leaf nodes: fair -> buys comp: no and excellent -> buys comp: yes

1st Internal Node - Income: low				
	age	student	credit_rating	buys_comp
5	>40	yes	fair	yes
6	>40	yes	excellent	no
7	31…40	yes	excellent	yes
9	<=30	yes	fair	no

2nd Internal Node - Income: low, age > 40

Info(D)	1

student	p_i	n_i	$I(p_i,n_i)$
yes	1	1	1
no	0	0	0

$Info_{student}(D)$	1
Gain(student)	0

credit_rt	p_i	n_i	$I(p_i,n_i)$
fair	1	0	0
excellent	0	1	0

$Info_{credit\_rt}(D)$	0
Gain(credit_rt)	1

Leaf node age <=30 -> buys comp: no

Leaf node age 31…40 -> buys comp: yes

2nd Internal Node - Income: low, age > 40			
	student	credit_rating	buys_comp
5	yes	fair	yes
6	yes	excellent	no

Similarly, in the case of income: medium, the 1st internal node is split according to age due to the highest gain, but the 2nd split is happening depending on student. Finally, in the case of income: high, the 1st internal node is split according to credit_rt, leading then in two leaf nodes without having a 2nd split.

1st Internal Node - Income: medium

age	p_i	n_i	$I(p_i,n_i)$
<=30	1	1	1
31…40	1	0	0
>40	3	0	0

Info(D)	0.65
$Info_{age}(D)$	0.333
Gain(age)	0.317

credit_rt	p_i	n_i	$I(p_i,n_i)$
fair	2	1	0.918
excellent	3	0	0

$Info_{credit\_rt}(D)$	0.459
Gain(credit_rt)	0.191

student	p_i	n_i	$I(p_i,n_i)$
yes	2	0	0
no	3	1	0.811

$Info_{student}(D)$	0.541
Gain(student)	0.109

resulting in two leaf nodes:
student: yes -> buys comp: yes
student: no -> buys comp: no

Leaf node age 31…40 -> buys comp: yes

Leaf node age > 40 -> buys comp: yes

1st Internal Node - Income: medium

	age	student	credit_rating	buys_comp
4	>40	no	fair	yes
8	<=30	no	fair	no
10	>40	yes	fair	yes
11	<=30	yes	excellent	yes
12	31…40	no	excellent	yes
14	>40	no	excellent	yes

2nd Internal Node - Income: medium, age <= 30

	Info(D)	1

student	p_i	n_i	$I(p_i,n_i)$
yes	1	0	0
no	0	1	0

$Info_{student}(D)$	0
Gain(student)	1

credit_rt	p_i	n_i	$I(p_i,n_i)$
fair	0	1	0
excellent	1	0	0

$Info_{credit\_rt}(D)$	0
Gain(credit_rt)	1

We could also choose to split according to credit_rt but due to small age it makes more sense to split depending on student

2nd Internal Node - Income: medium, age <= 30

	student	credit_rating	buys_comp
8	no	fair	no
11	yes	excellent	yes

Internal Node - Income: high

age	p_i	n_i	$I(p_i,n_i)$
<=30	1	1	1
31…40	0	2	0
>40	0	0	0

Info(D)	0.811
$Info_{age}(D)$	0.500
Gain(age)	0.311

credit_rt	p_i	n_i	$I(p_i,n_i)$
fair	0	3	0
excellent	1	0	0

$Info_{credit\_rt}(D)$	0.000
Gain(credit_rt)	0.811

student	p_i	n_i	$I(p_i,n_i)$
yes	0	1	0
no	1	2	0.918

$Info_{student}(D)$	0.689
Gain(student)	0.123

2nd Internal Node
Income: high, credit_rt: fair -> buys comp: no
Income: high, credit_rt: excellent -> buys comp: yes

1st Internal Node - Income: high

	age	student	credit_rating	buys_comp
1	<=30	no	fair	no
2	<=30	no	excellent	yes
3	31…40	no	fair	no
13	31…40	yes	fair	no

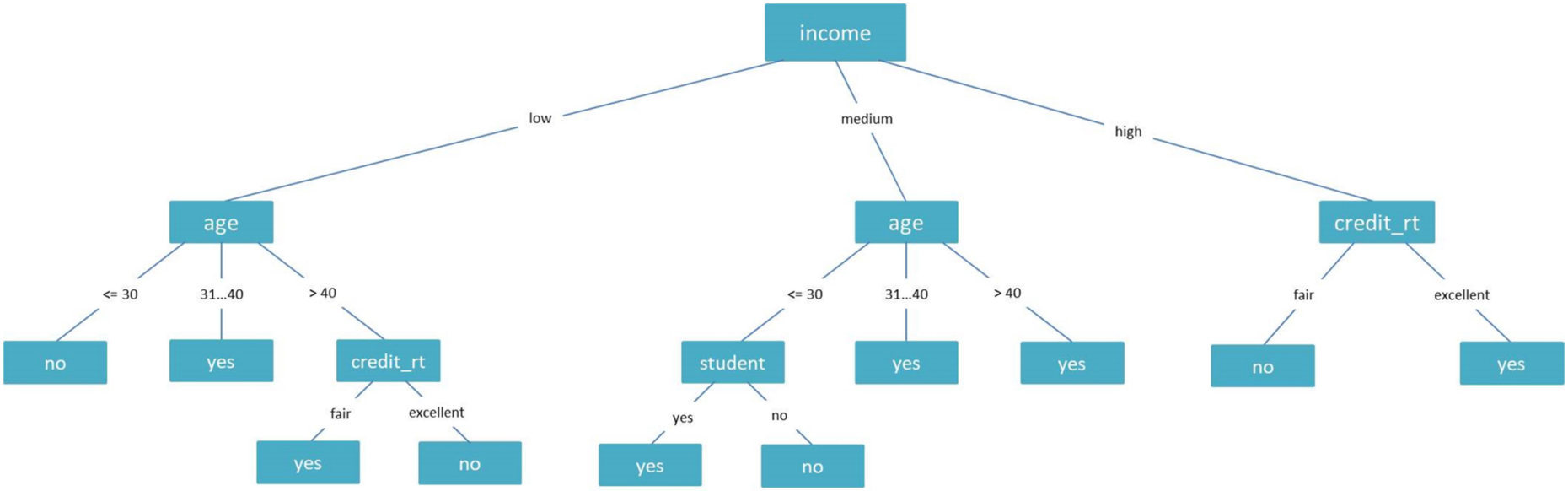
2nd Internal Node - Income: high, credit_rt:fair

	age	student	buys_comp
1	<=30	no	no
3	31…40	no	no
13	31…40	yes	no

2nd Internal Node - Income: high, credit_rt:exce

	age	student	buys_comp
2	<=30	no	yes

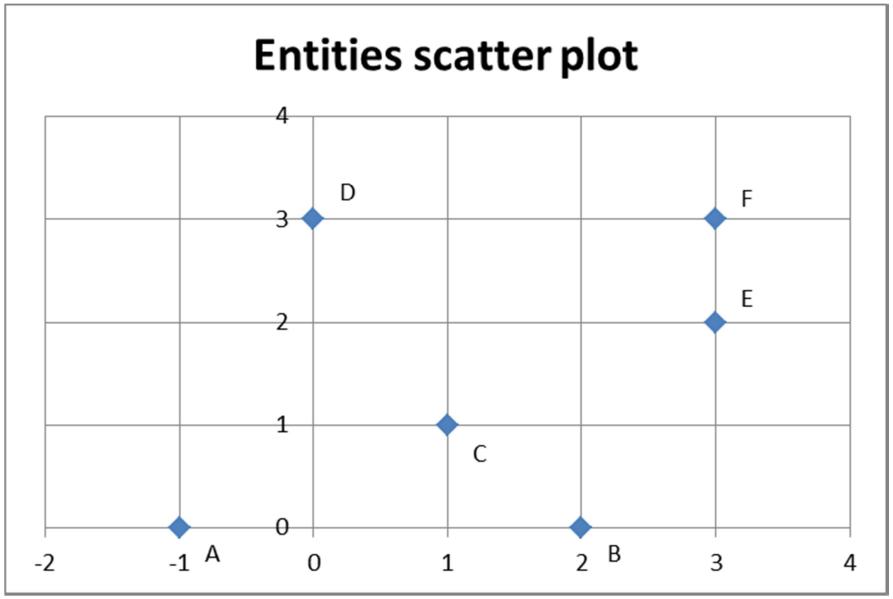
The decision tree is visually depicted below.



Task 4

a) The scatter plot of the given entities

	x	y
A	-1	0
B	2	0
C	1	1
D	0	3
E	3	2
F	3	3



Results of agglomerative hierarchical clustering using single link (smallest distance among entities), complete link (largest distance among entities) and average link (average distance among entities). You can see also the respective excel files for the calculations.

Hierarchical clustering

Single Link

	A	B	C	D	E	F
A	0					
B	3	0				
C	2.236	1.414	0			
D	3.162	3.606	2.236	0		
E	4.472	2.236	2.236	3.162	0	
F	5	3.162	2.828	3	1	0

	A	B	C	D	EF
A	0				
B	3	0			
C	2.236	1.414	0		
D	3.162	3.606	2.236	0	
EF	4.472	2.236	2.236	3	0

	A	BC	D	EF
A	0			
BC	2.236	0		
D	3.162	2.236	0	
EF	4.472	2.236	3	0

	ABC	D	EF
ABC	0		
D	2.236	0	
EF	2.236	3	0

	ABCD	EF
ABCD	0	
EF	2.236	0

Complete Link

	A	B	C	D	E	F
A	0					
B	3	0				
C	2.236	1.414	0			
D	3.162	3.606	2.236	0		
E	4.472	2.236	2.236	3.162	0	
F	5	3.162	2.828	3	1	0

	A	B	C	D	EF
A	0				
B	3	0			
C	2.236	1.414	0		
D	3.162	3.606	2.236	0	
EF	5	3.162	2.828	3.162	0

	A	BC	D	EF
A	0			
BC	3	0		
D	3.162	3.606	0	
EF	5	3.162	3.162	0

	ABC	D	EF
ABC	0		
D	3.606	0	
EF	5	3.162	0

	ABC	DEF
ABC	0	
DEF	5	0

Average Link

	A	B	C	D	E	F
A	0					
B	3	0				
C	2.236	1.414	0			
D	3.162	3.606	2.236	0		
E	4.472	2.236	2.236	3.162	0	
F	5	3.162	2.828	3	1	0

	A	B	C	D	EF
A	0				
B	3	0			
C	2.236	1.414	0		
D	3.162	3.606	2.236	0	
EF	4.736	2.699	2.532	3.081	0

	A	BC	D	EF
A	0			
BC	2.618	0		
D	3.162	2.921	0	
EF	4.736	2.616	3.081	0

	A	BCEF	D
A	0		
BCEF	3.677	0	
D	3.162	3.001	0

	A	BCDEF
A	0	
BCDEF	3.574	0

We observe that in all three methods entity E is clustered with entity F in the first step and entity B with entity C in the next one. Starting with the third step, in the case of single link there are entities with equal minimum distances with each other. So we chose to cluster entity A with BC. However, this is not the case with complete and average link since we had only one minimum distance in every step. Finally, regarding average link, we can see that entities D and A are clustered last, as a result of being generally further from the other entities.

b) In K means we clustered the entities using two options, because point D has the same distance (3) from both centroids C1 and C2. So, we calculated the two options shown below. In the first option we assigned point D in cluster C1 and in the second option we assigned D in cluster C2.

K - means (K = 2), 2 iterations - 1st option

1st iteration (C1 = 0,0 / C2 = 3, 3)

	Distance to C1	Distance to C2	Cluster
A	1	5	C1
B	2	3.162	C1
C	1.414	2.828	C1
D	3	3	C2
E	3.606	1	C2
F	4.243	0	C2

new centroids

C1*	0.67	0.33
C2*	2	2.67

2nd iteration (C1* = 0.67, 0.33 / C2* = 2, 2.67)

	Distance to C1*	Distance to C2*	Cluster
A	1.702	4.016	C1*
B	1.37	2.67	C1*
C	0.747	1.947	C1*
D	2.753	2.027	C2*
E	2.867	1.204	C2*
F	3.544	1.053	C2*

K - means (K = 2), 2 iterations - 2nd option

1st iteration (C1 = 0,0 / C2 = 3, 3)

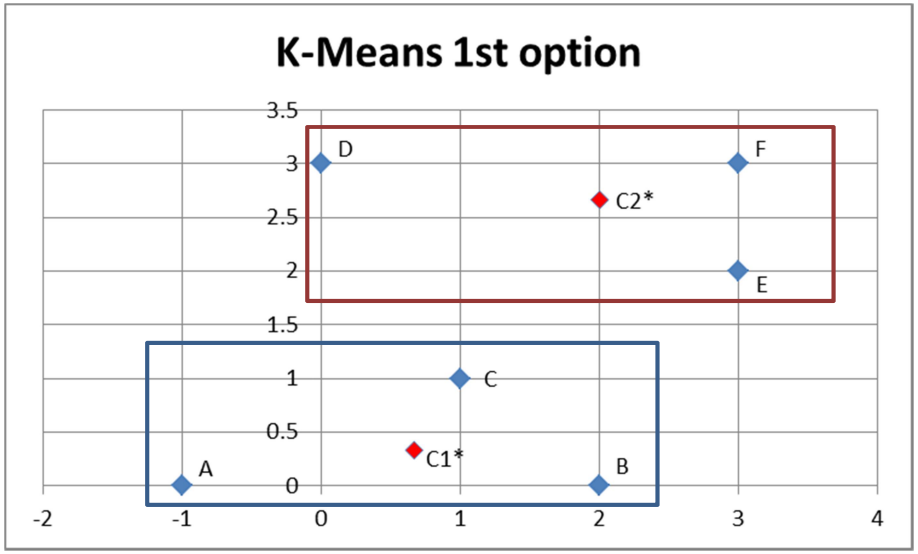
	Distance to C1	Distance to C2	Cluster
A	1	5	C1
B	2	3.162	C1
C	1.414	2.828	C1
D	3	3	C1
E	3.606	1	C2
F	4.243	0	C2

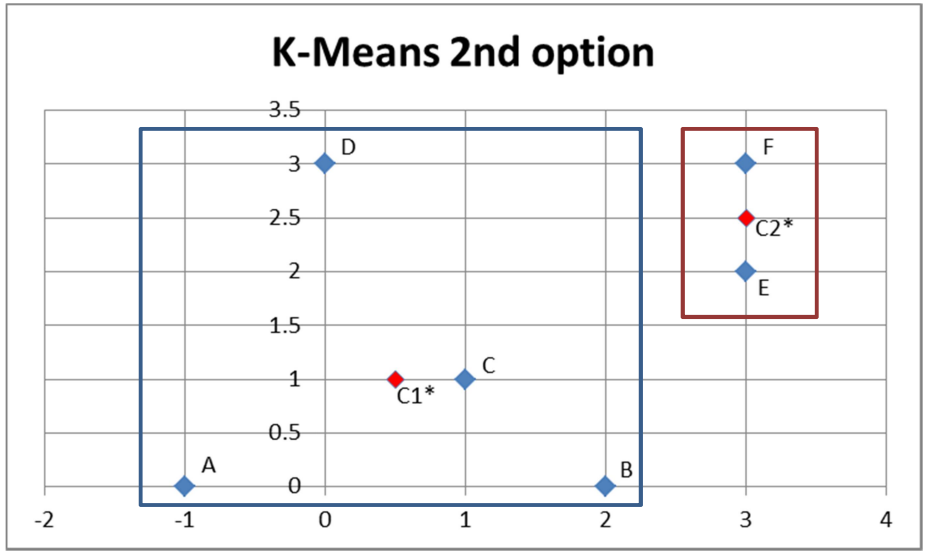
new centroids

C1*	0.5	1
C2*	3	2.5

2nd iteration (C1* = 0.67, 0.33 / C2* = 2, 2.67)

	Distance to C1*	Distance to C2*	Cluster
A	1.803	4.717	C1*
B	1.803	2.693	C1*
C	0.5	2.5	C1*
D	2.062	3.041	C1*
E	2.693	0.5	C2*
F	3.202	0.5	C2*





In the case of K-Means method the algorithm stopped only after 2 iterations, since no changes in the clusters of the entities have been made. We could assume that the position of the centroids helped in the faster completion of the algorithm. Taking into acoount the distribution of the entities we can conclude that the K-means method is much faster than hierarchical methods.

Task 5

In order to produce association rules we used the Apriori algorithm. The results are shown below.

Tid	Items
10	A, B, C
20	A, C, D, F
30	B, C, D, E
40	B, C, E
50	C, E, F
60	C, E, F
70	A, B, C, F
80	A, B
90	B, C, E
100	B, F

(support=25% and confidence=75%)

Step 1

Support (minsup = 25%)

Item	Count	Support
A	4	40%
B	7	70%
C	8	80%
D	2	20%
E	5	50%
F	5	50%

Step 2

Support (minsup = 25%)

A, B	3	30%
A, C	3	30%
A, F	2	20%
B, C	5	50%
B, E	3	30%
B, F	2	20%
C, E	5	50%
C, F	4	40%
E, F	2	20%

Step 3

Support (minsup = 25%)

A, B, C	2	20%
B, C, E	3	30%

Confidence (minconf = 75%)	
B --> C, E	43%
C --> B, E	38%
E --> B, C	60%
B,C --> E	60%
B, E --> C	100%
C, E --> B	60%

	s	c
B, E --> C	30%	100%

support=25% and confidence=50%

Step 1

Support (minsup = 25%)

Item	Count	Support
A	4	40%
B	7	70%
C	8	80%
D	2	20%
E	5	50%
F	5	50%

Step 2

Support (minsup = 25%)

A, B	3	30%
A, C	3	30%
A, F	2	20%
B, C	5	50%
B, E	3	30%
B, F	2	20%
C, E	5	50%
C, F	4	40%
E, F	2	20%

Step 3

Support (minsup = 25%)

A, B, C	2	20%
B, C, E	3	30%

Confidence (minconf = 50%)	
B --> C, E	43%
C --> B, E	38%
E --> B, C	60%
B,C --> E	60%
B, E --> C	100%
C, E --> B	60%

	s	c
E --> B, C	30%	60%
B,C --> E	30%	60%
B, E --> C	30%	100%
C, E --> B	30%	60%

(support=35% and confidence=50%)

Step 1

Support (minsup = 35%)

Item	Count	Support
A	4	40%
B	7	70%
C	8	80%
D	2	20%
E	5	50%
F	5	50%

Step 2

Support (minsup = 35%)

A, B	3	30%
A, C	3	30%
A, F	2	20%
B, C	5	50%
B, E	3	30%
B, F	2	20%
C, E	5	50%
C, F	4	40%
E, F	2	20%

Step 3

Support (minsup = 35%)

B, C, E	3	30%

Confidence (minconf = 75%)	
B --> C	71%
C --> B	63%
C --> E	63%
C --> F	50%
E --> C	100%
F --> C	80%

	s	c
B --> C	50%	71%
C --> B	50%	63%
C --> E	50%	63%
C --> F	40%	50%
E --> C	50%	100%
F --> C	40%	80%

As we change the support and confidence thresholds, the number of association rules varies.

So for support = 25% and confidence = 75%, we found one association rule (B, E --> C). This configuration results in fewer, but more reliable rules due to the high confidence threshold.

For support = 25% and confidence = 50%, we found four association rules. Lowering the confidence threshold leads to more rules being generated. However, the trade-off is that these rules are not as reliable as those with higher confidence levels.

For support = 35% and confidence = 50%, we found six association rules (B --> C, C --> B, C --> E, C --> F, E --> C, F --> CIncreasing the support threshold leads to itemsets that are more frequent. However, in this specific case duw to the higher support threshold the rules that are generated are simpler compared to the other two cases. So, with a higher support threshold, we focus on more frequent itemsets, ensuring that the rules are more relevant to a larger portion of the transactions.