Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 3134–3143

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Application of Machine Learning in medical data analysis illustrated with an example of association rules.

Beata Butryn[a], Iwona Chomiak-Orsa[a], Krzysztof Hauke[a], Maciej Pondel[a]*, Agnieszka Siennicka[b]

[a]*Wroclaw University of Economics and Business, Komandorska 118/120, 53-345 Wrocław, Poland*
[b]*Wroclaw Medical University, Wybrzeże L. Pasteura 1, 50-367 Wrocław, Poland*

Abstract

The authors propose a method for medical data analysis aiming to determine risk groups. The method uses association rules generation and the analysis of a selected target feature. The method is presented with an example of patient analyses in regard to the possibility of suffering a stroke. The method is universal and can be used also for other medical dataset analyses leading to risk group identification.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of KES International.

Keywords: artificial intelligence, mart medicine, machine learning in medicine, tailored intervention, risk identification, association rules

1. Introduction

The development of technology has contributed to the collection of data resources. Technologies supporting management processes through the use of multi-criteria analysis algorithms are becoming tools used in all business areas. Healthcare is also such an area. Collecting information resources about past diseases as well as searching for

* Corresponding author. Phone: +48 502 133 360.
E-mail address: maciej.pondel@ue.wroc.pl

This research was supported by the European Union's Horizon 2020 grant HeartBIT_4.0 - Application of innovative Medical Data Science technologies for heart diseases (No 857446).

The project is financed by the Ministry of Science and Higher Education in Poland under the programme "Regional Initiative of Excellence" 2019 - 2022 project number 015/RID/2018/19 total funding amount 10 721 040,00 PLN,

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of KES International.
10.1016/j.procs.2021.09.086

diagnostic tools is an adequate field of application for artificial intelligence tools. Artificial intelligence for surgery, should support the process of diagnosis and medical procedures. Machine learning, which is a key technology in the field of intelligence, begins to play a special role in this extension. Technologies such as machine learning, neural networks can be used to analyze data characterizing patients, through their categorization, to the creation of advanced predictive models determining the probability of a patient suffering from a specific disease. This group of tools also includes neural networks that use professional activity and are to map the mechanisms of expert inference.

The development of artificial intelligence technologies contributes to the search for solutions that would be useful in the field of healthcare. The use of artificial intelligence solutions in other areas of decision support indicates the possibility of their effective application also in the area of medical diagnostics. Technologies such as machine learning and neural networks can be used to analyze patient data, by categorizing them, and to create advanced predictive models that determine the likelihood of a patient suffering from a specific disease. This is of particular importance in the implementation of diagnostic processes, which, through the support of artificial intelligence solutions, can become significantly more effective.

It is indicated that the use of artificial intelligence tools in medicine may be of particular importance in two areas: diagnostics and predicting the results of medical procedures. Following the application of artificial intelligence by financial institutions and insurance companies, it seems that one of the directions of using artificial intelligence in medicine may be the creation of automated algorithms for predicting the risk of developing certain types of diseases.

Therefore, the article aims to indicate the possibilities of using artificial intelligence tools in this area. The authors aim to propose a method of predicting the probability of developing a specific disease. On the one hand, the creation of appropriate inference algorithms will enable the analysis of larger information resources constituting patient characteristics by automating them. On the other hand, it will contribute to personalizing the approach to prevention.

The research hypothesis defined by the authors is as follows below.

Using machine learning methods, it is possible to detect groups of people at risk of developing the analyzed disease. Thanks to the identification of risk groups, it is possible to prioritize and personalize preventive actions, which will result in their greater effectiveness.

The research objective of the article is to use the shared database with regard to the characteristics of a selected group of patients and to indicate the relationship between these pieces of data. On the basis of the identified dependencies, the creation of inference algorithms will allow the discovery of relationships between individual characteristics in order to identify groups with a high or increased risk of developing a specific disease.

The solution developed as a result of the research conducted and described in the article will use Data Science techniques which, by validating the predictive mechanisms, should have the characteristics of an expert judgment.

The next section of the article presents an overview of thematically related research. The research method is defined in the next section of the article. The next point contains a presentation and discussion of the results of the conducted research procedure.

The theoretical research carried out, as is the case with the indicated research procedure, contributes to the discussion by the authors on the possibilities and scope of application of the examined tools. The article ends with a conclusion.

2. Relates works

2.1. Smart medicine

Intelligent medicine (technology + medicine) relieves the health service, improves the possibilities of prevention, allows for the provision of services in isolation and enables cost reduction. Thanks to the data pre-selection carried out by intelligent accessories, the doctor treating a patient can extract the information that is a clue as to the existence of a specific disease state from a multitude of information. Artificial Intelligence (AI) improves diagnostics [1] and increases safety thanks to algorithms that are able to interpret data from various sources and devices in real-time. Telemedicine allows one to perform preventive and control examinations outside a standard medical facility, consult their results, monitor one's health without leaving home, as well as issue prescriptions and referrals [2]. Virtual Reality (VR) plays an increasingly important role in patient care, providing the opportunity to see patients from a different perspective [3]. The 5G network provides a greater speed and contributes to the development of telemedicine, enabling the dissemination of medical and diagnostic services as well as real-time treatment [4].

2.2. *Machine learning in medicine, AI in medicine*

The complexity and rise of data in healthcare cause that artificial intelligence (AI) is increasingly often applied within the field. Artificial Intelligence is a collection of technologies. The most common of these is machine learning, which uses statistical techniques to fit models to data, e.g. a Random Forest algorithm to identify five different levels of severity of anxiety, depression and stress [5].

A more complex form of machine learning is a neural network. It is used for categorization applications such as determining whether a patient will develop a specific disease [6]. It presents problems in terms of inputs, outputs and weights of variables or 'features' that associate inputs with outputs [7]. The most promising route for AI in medicine is the development of automated risk prediction algorithms which can be used to guide health care [8].

ML techniques used in medicine cover diagnosis and prediction of results. As a result, it is possible to identify high-risk medical emergencies such as relapse or transition to another disease state, e.g. predict progression from pre-diabetes to type 2 diabetes using routinely collected electronic health data [9]. An example is the use of machine learning classification methods and comparing them with different statistical measures [10].

2.3. *Risk identification*

Risk prediction models are typically based on a limited number of predictors in all patient groups. Data-driven (machine learning - ML) techniques can improve risk prediction performance by discovering new risk predictors and learning the complex interactions between them. Risk prediction algorithms are typically developed using multivariate regression models that combine information about a limited number of well-known risk factors and assume that all such factors are related to (e.g.in cardiovascular disease - CVD) outcomes in a linear fashion, with limited or no interactions between different factors [11].

Data-driven techniques based on machine learning (ML) can improve risk prediction performance through the use of large data repositories to identify new predictors of risk and more complex interactions between them.
Only a few studies have explored the potential benefits of using ML methods to predict CVD risk, focusing only on a limited number of ML methods [12, 13] or a limited number of risk predictors [14].

2.4. Tailored interventions

There are two ways to improve adherence: single interventions and multicomponent interventions. Single interventions use a single method such as treatment regimen management, patient education, convenient medication administration, reminders, psychotherapy, etc. Multicomponent interventions use multiple methods simultaneously.

Tailored interventions include identifying factors that influence the patients' adherence behaviour, discussing the patient's illness and treatment representations and suggesting a way to overcome barriers and improve medication adherence [15].

Various interventions have been developed to improve patient compliance [16]. Most of the interventions described have not attempted a one-to-one approach to identify specific causes or barriers interfering with the ability to adhere to treatment regimens for individual patients [17].

2.5. Association rules

Many articles in the literature discuss various aspects of association rules. The issues are discussed at various levels of detail. The basic level discusses definitions and reviews the literature sources. At the advanced level, the authors deal with the mathematical approach to the problem of association rules. The effectiveness of the application of association rules largely depends on the source data. The efficiency of the algorithm for the analysis of large data sets is discussed in the publication [18]. However, most algorithms are imperfect due to the measures used in the rule evaluation process and the prioritization techniques applied at the attribute level, which may play a critical role in the rule generation process [19]. The most effective and the most frequently used technique in designing recommender

systems is filtering data according to attributes [20]. Association rules are created on the basis of research resulting from observation of the environment. The obtained results motivate us to further probe data sets and explore association rules [21]. Association rules are used for the investigation of road accidents involving vehicles with hazardous materials on express roads [22], among other things. Investigating the factors contributing to ship accidents is a high priority issue in the maritime safety analysis. In this study, a method based on association rules was proposed for the analysis of these component factors [27]. The increasing complexity of software projects requires more and more sophisticated methods of analysis and testing. Identification of defective software components is essential to ensure software quality, while also improving the effectiveness software testing activities [26]. Another application of association rules is the construction of a semantic network using the RDF format [23]. Association rules may have their measures: support, confidence, comprehensibility and curiosity in two objective functions [24]. In building rules, one can use the method of identifying functional dependencies between attributes [25],[27].

Association rules are a way to discover interesting relationships between variables in large data sets. Association rules are similar to decision rules. The decision (the right side of the implication) is not predetermined, i.e. we do not know which attribute is to be based on. This is an example of learning without a teacher (similar to grouping algorithms): the algorithm does not have a predetermined correct answer; instead, it is intended to describe the internal relationships between the attributes, e.g.:

$$a_1=v_1 \wedge a_9=v_9 \Rightarrow a_4=v_4$$

In order to define an association rule, a few basic concepts should be introduced, such as a set of transactions and a set of so-called goods.

Definition 1. Items are a set of binary attributes $I = I_1, I_2, \ldots, I_n$. Let $T = t_1, t_2, \ldots, t_m$ be the collection of all transactions. Each transaction t_i is represented by a binary vector, where $t_i[k] = 1$ if t_i includes I_k and $t_i[k] = 0$ otherwise. We will call an itemset X with $X \subseteq I$. If the set contains k items, it is a k-itemset. The transaction t_j contains a set of items X if X is a subset t_j.

Definition 2. We'll call an association rule such an implication $X \rightarrow Y$, where X and Y are sets of items with $X \cap Y = \emptyset$. The predecessor of this implication (X) will be called LHS (left-hand side), and the successor (Y) - RHS (right-hand side). The strength of an association rule is measured by two indicators: support and confidence.

Definition 3. The rule support is the ratio:

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N},$$

where N is the total number of transactions.

Definition 4. The certainty of a rule is determined by the following formula:

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

Support and confidence are very important parameters in finding rules. The support for the rule should be high, as it determines the statistical presence of the goods in total. The lower the support, the more likely the rule is to be a coincidence and nothing more. In addition, support is often linked to a business point of view, as it is not profitable, for example, to promote products that are rarely bought together. Rule certainty determines the probability with which the RHS will appear in transactions that contain LHS.

The problem of finding association rules can be defined as follows:

Definition 5. For a given set of transactions T, find rules having support greater than minimum support and certainty greater than minimum certainty, where minimum support and certainty are defined thresholds. The problem can be solved by checking the supports and certainty for all possible rules, but this approach is time-consuming and redundant, so many algorithms break the task into two parts:

- finding the so-called frequent/large itemsets,
- generating rules from frequent patterns found in the previous step.

Definition 6. A frequent pattern is a set of items X for which $\sigma(X) \geq$ minsupp, where minsupp is fixed minimum support.

Examples of algorithms for finding frequent patterns are as follows: AIS, APRIORI, SETM, DHP, Partition, ECLAT, FP-GROWTH.
The article will use the APRIORI algorithm. APRIORI is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. Frequent item sets determined by APRIORI can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.
Definition 7. The maximum frequent pattern is a frequent pattern X for which none of its immediate supersets is a frequent pattern. Such patterns provide a compact representation of the set of patterns found and are useful in cases where the found patterns can be very long because then an enormous number of subsets do not need to be listed. However, finding such patterns for the generation of association rules has a certain disadvantage - only the maximum pattern support is given. Nothing is known about the support of the subsets, so one then has to count them for each of them, which is not effective.
Examples of algorithms are as follows: MAX-MINER, DepthProject, MAFIA, GenMax.
Definition 8. A closed frequent pattern is a frequent pattern of X for which none of its immediate supersets has the same support as it does.
These patterns address the problem of subset support that exists with maximum frequent patterns. Examples of algorithms are as follows: CLOSE,CLOSET,CHARM.

Generating rules
The rules should be extracted from the generated set of frequent patterns. Each k-set represents k^k - 2 potential rules (all combinations except $X \rightarrow \emptyset$ and $\emptyset \rightarrow X$). Generation can be done by dividing the file into two subsets: X and Y − X, and then checking the certainty threshold for the obtained rule. The support threshold no longer needs to be checked because each rule is generated from a frequent pattern that already meets this condition. It takes time to test all the possibilities, so the following theorem is useful:
Theorem 1. If the rule $X \rightarrow Y - X$ does not satisfy the minimum certainty condition, no rule of the form $X' \rightarrow Y - X'$, where $X' \subset X$, also fails. This allows for limiting the number of rules whose certainty should be verified by starting the search with rules that have one element in the successor. Then new ones are built on the basis of those already found.

3. Research methodology

The main aim of the research is to propose a method for the identification of risk groups in order to provide patients with the relevant intervention. Such intervention can be prevention or delivery of relevant education prepared in a personalized way. The authors do not aim to propose a precise intervention directed to identified risk groups. The proposition of a universal method for identifying risk applicable in various medical cases is the main contribution of the authors.
This study was conducted based on the dataset published at Kaggle. It includes 11 clinical features for predicting stroke events. The 12th attribute specifies whether the patient has suffered a stroke or not. Published on 2021-01-26, the dataset includes data of 5,110 patients.
The dataset consists of the following attributes:
1) id: unique identifier
2) gender: "Male", "Female" or "Other" - quantitative
3) age: age of the patient - qualitative
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension - quantitative
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease - quantitative
6) ever_married: "No" or "Yes" - quantitative
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed" - quantitative
8) Residence_type: "Rural" or "Urban" - quantitative
9) avg_glucose_level: average glucose level in blood - qualitative
10) bmi: body mass index - qualitative
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"* - quantitative

12) stroke: 1 if the patient had a stroke or 0 if not – binary

Example data is presented in Figure 1. One row represents 1 patient.

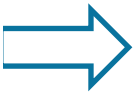
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Figure 1. First 5 rows of the original dataset

The method consists of the following steps:

1. Data preparation. Association rules require binary attributes that contain information if a specific event or a specific characteristic regards the described patient or not.
   As far as quantitative attributes are concerned, they have to be transformed into corresponding sets of binary attributes. The sklearn.preprocessing.LabelBinarizer class included in the scikit-learn library is used in this task. An example of a quantitative feature of the smoking status into a set of 3 binary attributes is presented in Figure 2.

id	smoking_status
9046	formerly smoked
51676	never smoked
31112	never smoked
60182	smokes
1665	never smoked



id	smoking_formerly smoked	smoking_never smoked	smoking_smokes
9046	1	0	0
51676	0	1	0
31112	0	1	0
60182	0	0	1
1665	0	1	0

Figure 2. Transformation of quantitative attributes into binary attributes

   Qualitative attributes have to be transformed into quantitative attributes first and then into binary ones. For transformation into a quantitative attribute, the standard interpretation or a cluster analysis can be employed here.

2. Definition of a min. support parameter and generation of Frequent Itemsets meeting the defined threshold of support. In the research, the authors set minimum support to 0.002, which means that all Frequent sets, including a minimum of 10 patients, were generated. Frequent itemsets will be generated using the APRIORI algorithm implemented in the mlxtend† library.
3. Definition of a metric and its minimum value as parameter and generation of Association Rules meeting the defined threshold of a selected metric. The authors selected lift as a metric and set its minimum value at 0.01, assuming that rules with such a low level of lift are extremely interesting because they were responsible fora negative correlation between rule antecedents and analysed consequents (in this case, stroke). Association rules will be generated using the mlxtend library as well.
4. Selection of association rules where the consequence is a stroke.
5. Analysis of most informative rules – those with the highest lift for identification of the risk group and those with the lowest lift for the definition of the safety group.
6. Split of an initial dataset based on a selected feature in order to identify differences in rules generated on

† Mlxtend (machine learning extensions) is a Python library of useful tools for the day-to-day data science tasks. The project is released under a permissive new BSD open source license. http://rasbt.github.io/mlxtend/

subsets and better understand risk influencers.
7. Generation of rules for separate subsets and their evaluation with an expert.
8. Application of rules for specific patients in order to assess whether they belong to the risk group.

4. **Research Findings**

After the data preparation, the authors generated a set including only binary attributes. Based on the prepared data, frequent itemsets were generated for the whole dataset as a foundation of rules. In the following step, the authors generated rules. An example of 5 rules with the highest list metric is presented in Figure 3. Those example rules don't deliver any valuable knowledge about stroke probability.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
650549	(work_type_Self-employed, heart_disease, bmi_overweight)	(age_high, Male, Residence_type_Rural)	0.005479	0.026810	0.002153	0.392857	14.653285	0.002006	1.602901
650564	(age_high, Male, Residence_type_Rural)	(work_type_Self-employed, heart_disease, bmi_overweight)	0.026810	0.005479	0.002153	0.080292	14.653285	0.002006	1.081344
650193	(no_hypertension, age_high, smoking_Unknown)	(work_type_Self-employed, heart_disease, glucose_normal)	0.023092	0.008415	0.002544	0.110169	13.092235	0.002350	1.114353
650176	(work_type_Self-employed, heart_disease, glucose_normal)	(no_hypertension, age_high, smoking_Unknown)	0.008415	0.023092	0.002544	0.302326	13.092235	0.002350	1.400235
146429	(glucose_metabolic_consequences, stroke)	(Residence_type_Urban, heart_disease, work_type_Private)	0.010959	0.015068	0.002153	0.196429	13.035714	0.001988	1.225693

Figure 3. Example 5 rules with calculated metrics

The number of generated rules was enormous, amounting to 4,917,232. The authors concentrated on those rules where "consequents" is (stroke) to identify a feature or a set of features that impact the analysed phenomenon (see Table 1).

Authors identified a group of highly impacting rules where the lift is greater than 3, which means that a person to whom such rule is applicable is at least 3 times more exposed to stroke than a typical person. The number of rules with lift > 3 is amounts to 800. The top 3 rules with the highest lift are presented in Table 1.

Table 1. Rules with the highest lift value

rule id	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
146408	(Residence_type_Urban, heart_disease, glucose_metabolic_consequences, work_type_Private)	(stroke)	0.005088	0.048728	0.002153	0.423077	8.682422
624860	(ever_married, heart_disease, work_type_Private, no_hypertension, glucose_metabolic_consequences)	(stroke)	0.005479	0.048728	0.002153	0.392857	8.062249
146378	(no_hypertension, heart_disease, glucose_metabolic_consequences, work_type_Private)	(stroke)	0.006067	0.048728	0.002348	0.387097	7.944034

Those rules regard relatively small groups of patients. Support measure at a level of 0.002 means that they were generated on groups of 10 people, so the probability that they could be applicable to a random patient is relatively low. To focus on more general rules, the authors narrowed their scope down to those with support higher than 0.01. The number of such rules is 26. The authors believe that those rules constitute valuable knowledge that can streamline further medical research. The top 4 rules are presented in Table 2.

Table 2. Top rules with the highest lift value with support > 0.01

rule id	antecedents	consequents	antecedent support	consequent support	support	confidence	lift

3746	(age_high, work_type_Private)	(stroke)	0.065753	0.048728	0.014090	0.214286	4.397590
46600	(no_heart_disease, age_high, work_type_Private)	(stroke)	0.052250	0.048728	0.010959	0.209738	4.304258
46432	(no_hypertension, age_high, work_type_Private)	(stroke)	0.050294	0.048728	0.010372	0.206226	4.232182

Using the exact mechanism, we can identify safe groups (far from the risk group). In such a case, analysis of rules with the lowest lift value can help. The top 4 rules are presented in Table 3.

Table 3. Top rules with the lowest lift value

rule id	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
51682	(no_hypertension, no_heart_disease, glucose_normal)	(stroke)	0.585910	0.048728	0.016830	0.028724	0.589479
50842	(no_hypertension, Female, glucose_normal)	(stroke)	0.364579	0.048728	0.010568	0.028986	0.594843
4418	(no_hypertension, glucose_normal)	(stroke)	0.607241	0.048728	0.018787	0.030938	0.634908
51542	(no_hypertension, no_heart_disease, Residence_type_Rural)	(stroke)	0.423092	0.048728	0.013112	0.030990	0.635976

When the researcher defines a differentiator splitting the whole analysed set into subsets, generation of separate sets of rules and a comparison can deliver valuable knowledge. The authors split the set into 2 groups defined by genders in order to compare metrics for the same antecedents and consequents. Example rules are presented in Table 4.

Table 4. Rules generated separately for both genders

rule id	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	Gender
111	(age_high)	(stroke)	0.135697	0.051064	0.025532	0.188153	3.684669	Male
95	(age_high)	(stroke)	0.141235	0.047078	0.024708	0.174941	3.715943	Female
2559	(work_type_Private, age_high)	(stroke)	0.067612	0.051064	0.015603	0.230769	4.519231	Male
2183	(work_type_Private, age_high)	(stroke)	0.064441	0.047078	0.013022	0.202073	4.292250	Female
103	(smoking_smokes)	(stroke)	0.150918	0.047078	0.006344	0.042035	0.89288	Female
97	(smoking_smokes)	(stroke)	0.159338	0.051064	0.010875	0.068249	1.336548	Male

Analysis of those rules shows that age in both genders is a significant feature determining stroke possibility, although females' lift is slightly higher and confidence lower than in males.
A comparison of rules where "smoking_smokes" feature is antecedents is interesting. It transpires that smoking females are less stroke liable than an "average female" (lift = 0.89), contrary to males where those smoking ones are more stroke liable (lift = 1.33).

5. Discussion

The main advantage of association rules generation is that the researcher can quickly and automatically identify a potentially valuable hypothesis related to an examined disease. Such knowledge can streamline further medical research.
One of the highly recognised drawbacks of the application of AI models in medicine is that they are illegible for

humans. Most machine learning models and deep learning models are very efficient and accurate, but they work as a so-called black-box. They can predict the probability of falling sick by a given patient, but they cannot explain the conclusions, which is crucial in current medicine. The authors believe that an approach based on association rules is helpful because it delivers understandable knowledge that a physician or medical researcher can base their actions on. Moreover, the presented approach and obtained results may shed light on new possibilities such as building new models to support medical decisions using data that does not come from standard sets of medical parameters (although often collected during an interview e.g. data on the nature of professional work, shown as significant in the current study or data on the patient's family situation). Despite the fact that relationships between job involvement and the risk of ischemic disorders were demonstrated several decades ago (see the Framingham Heart Study [28]), still, the vast majority of the existing models for medical decisions, including risk stratification, is based on typically medical parameters (e.g. the classic ESRS[29] scale, which allows for assessing the risk of stroke based on age, comorbidities and smoking status). Information such as the type of professional activity or family situation served rather as a background or context to which the doctor could refer while advising the patient how they may change their behaviour in order to be healthier or was presented in some scientific research [30] suggesting that it requires further studies. We believe that the presented model and the detected associations may be a significant signal encouraging the modification of the existing tools for risk assessment.

6. Conclusions

Intelligent methods of analyzing data from the environment are the subject of research by analysts and data science. It is not easy to imagine the world today without examining the relationships between the facts. Association rules should be considered as one of the approaches to data analysis. The most common application area of association rules is sales and marketing, but they can be very efficient also in other areas. Based on medical data, association rules have been proven to be effective in examining relationships between features. The analysis of this data allowed for the definition of relationships between particular attributes with a high probability. The general research hypothesis that association rules can detect relationships in medical data has been positively verified.

References

[1] Seong Ho Park , Kyunghwa Han (2018) Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 286 (3).

[2] A.S.Albahri, Jwan K.Alwan, Zahraa K.Taha, Sura F.Ismail, Rula A.Hamid, A.A.Zaidan, O.S.Albahri, B.B.Zaidan, A.H.Alamoodi, M.A.Alsalem (2021) IoT-based telemedicine for disease prevention and health promotion: State-of-the-Art. Journal of Network and Computer Applications, Volume 173, 102873.

[3] Wai Kan Yeung, Andy, Anela Tosevska, Elisabeth Klager, Fabian Eibensteiner, Daniel Lexar, Jivko Stoyjanov, Marija Glisic, Sebastian Zeiner, Stefan Tino Kulnik, Rick Crutzen, Olivier Kimberger, Maria Kletecka-Pulker, Atanas G. Atanasov, Harald Willschke (2021) Virtual and Augmented Reality Applications in Medicine: Analysis of the Scientific Literature. Journal of Medical Internet 23 (2).

[4] Hantao Ge, Lu Yeye (2020) 5G applications and trends in the field of smart medicine. Information and Communications Technology and Policy 46 (12): 15-20.

[5] Anu Priya, Shruti Garg, Neha Prerna Tigga (2020) Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. Procedia Computer Science, Volume 167: 1258-1267.

[6] Puja Guptaa, Shruti Garg (2020) Breast Cancer Prediction using varying Parameters of Machine Learning Models. Procedia Computer Science, Volume 171: 593-601.

[7] Davenport Thomas, Ravi Kalakota (2019) The potential for artificial intelligence in healthcare. Future healthcare Journal 6(2): 94–98.

[8] Sidney-Gibbons Jenni A.M, Chris J. Sidney-Gibbons (2019) Machine learning in medicine: a practical introduction. BMC Med Res Methodology 19, (64).

[9], João Lopes, Tiago Guimarães, Manuel Filipe Santos (2020) Predictive and Prescriptive Analytics in Healthcare: A Survey. Procedia Computer Science, Volume 170: 1029-1034.

[10] Neha Prerna, Tigga, Shruti Garg (2020) Prediction of Type 2 Diabetes using Machine Learning Classification Methods, Procedia Computer Science, Volume 167: 706-716.

[11] Alaa Ahmed M., Thomas Bolton, Emanuele Di Angentonio, James H.F. Rudd, Mihaela Van der Schaar (2019) Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PlosOne. https://doi.org/10.1371/journal.pone.0213653.

[12] Bharath Ambale-Venkatesh, Xiaoying Yang, Colin O. Wu, Kiang Liu, W. Gregory Hundley, Robyn McClelland, Antoinette S. Gomes, Aaron R. Folsom, Steven Shea, Eliseo Gullar, David A. Bluemke, Joao A.C. Lima (2017) Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circulation research 2017 (121):1092–1101.

[13] Ahmad T., Lund L.H., Rao P., Ghosh R., Warier P., Vaccaro B., Dashlorm U., O'Connor C.M., Felker G.M., Desai N.R. (2018) Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. Journal of the American Heart Association. 7 (8).

[14] Weng S.F., Reps J., Kai J., Garibaldi J.M., Qureshi N. (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data?. Jornal PlosOne. https://doi.org/10.1371/journal.pone.0174944.

[15] Xu, Hai-Yan, Yong-Ju Yu, Qian-Hui Zhang, Hou-Yuan Hu, Min Li (2020) Tailored Interventions to Improve Medication Adherence for Cardiovascular Diseases. Front Pharmacol. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7751638/.

[16] Van der Laan, D.M., Elders P.J.M., Boons C.C.L.M., Bosmans J.E., Nijpels G., HugtenburgJ.G. (2017) The (cost-)effectiveness of a patient-tailored intervention programme to enhance adherence to antihypertensive medication in community pharmacies: study protocol of a randomised controlled trial. Trials 18 (29).

[17] Linn A J., Vervloet M., Van Dijk L., Smit E.G., Van Weert J.C.M., (2011) Effects of eHealth Interventions on Medication Adherence: A Systematic Review of the Literature. Journal of medical Internet research. 13 (4).

[18] Osadchiy T., Poliakov I., Olivier P., Rowland M., Foster E. (2019) Recommender system based on pairwise association rules, in: Expert Systems with Applications, Volume 115, January 2019, Pages 535-542, https://doi.org/10.1016/j.eswa.2018.07.077, Elsevier.

[19] Jaber Alwidian, Bassam H.Hammo, Nadim Obeid (2018) WCBA: Weighted classification based on association rules algorithm for breast cancer disease, Applied Soft Computing, https://doi.org/10.1016/j.asoc.2017.11.013, Elsevier.

[20] Maryam Khanian Najafabadi, Mohd Naz'ri Mahrin, Suriayati Chuprat, Haslina Md Sarkan (2017) Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data, Computers in Human, https://doi.org/10.1016/j.chb.2016.11.010, Elsevier.

[21] Rim Rekik, Ilhem Kallel, Jorge Casillas, Adel M. Alimi (2018) Assessing web sites quality: A systematic literature review by text and association rules mining, International journal of information, https://doi.org/10.1016/j.ijinfomgt.2017.06.007, Elsevier.

[22] Jungyeol Hong, Reuben Tamakloe, Dongjoo Park (2020) Application of association rules mining algorithm for hazardous materials transportation crashes on expressway, Accident Analysis & Prevention, https://doi.org/10.1016/j.aap.2020.105497, Elsevier.

[23] Molood Barati, Quan Bai, Qing Liu (2017) Mining semantic association rules from RDF data, Knowledge-Based Systems, https://doi.org/10.1016/j.knosys.2017.07.009, Elsevier.

[24] Kamel Eddine Heraguemi, Nadjet Kamel, Habiba Drias (2018) Multi-objective bat algorithm for mining numerical association rules, International Journal of Bio-Inspired Computation, https://doi.org/10.1504/IJBIC.2018.092797,inderscienceonline.com.

[25] Federico Antonello, Piero Baraldi, Ahmed Shokry, Enrico Zio, Ugo Gentile, Luigi Serio (2021) Association rules extraction for the identification of functional dependencies in complex technical infrastructures, Reliability Engineering & System Safety, https://doi.org/10.1016/j.ress.2020.107305, Elsevier.

[26] Diana-Lucia Miholca, Gabriela Czibula, Istvan Gergely Czibula (2018) A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks, Information Sciences, https://doi.org/10.1016/j.ins.2018.02.027, Elsevier.

[27] Jinxian Weng, Guorong Li (2019) Exploring shipping accident contributory factors using association rules, Journal of Transportation Safety & Security, https://doi.org/10.1080/19439962.2017.1341440, Taylor & Francis.

[28] Miśkowiec Dawid, Kwarta Paulina, Witusik Andrzej, Pietras Tadeusz (2013) Wzór zachowania typu A jako predyktor choroby niedokrwiennej serca–czy wciąż aktualny problem. Postępy Psychiatrii i Neurologii, 22(2), 129-136.

[29] Christian Weimar, Hans-Christoph Diener, Mark J. Alberts, P. Gabriel Steg, Deepak L. Bhatt, Peter W.F. Wilson, Joachim Röther (2009) The Essen stroke risk score predicts recurrent cardiovascular events: a validation within the Reduction of Atherothrombosis for Continued Health (REACH) registry. Stroke, 40(2), 350-354.

[30] Eleonor I. Fransson, Solja T. Nyberg, Katriina Heikkilä, Lars Alfredsson, Jakob B. Bjorner, Marianne Borritz, Hermann Burr, Nico Dragano, Goedele A. Geuskens, Marcel Goldberg, Mark Hamer, Wendela E. Hooftman, Irene L. Houtman, Matti Joensuu, Markus Jokela, Anders Knutsson, Markku Koskenvuo, Aki Koskinen, Meena Kumari, Constanze Leineweber, Thorsten Lunau, Ida E.H. Madsen, Linda L. MagnussonHanson, Martin L. Nielsen, Maria Nordin, Tuula Oksanen, Jaana Pentti, Jan H. Pejtersen, Reiner Rugulies, Paula Salo, Martin J. Shipley, Andrew Steptoe, Sakari B. Suominen, Töres Theorell, Salla Toppinen-Tanner, Jussi Vahtera, Marianna Virtanen, Ari Väänänen, Peter J.M. Westerholm, Hugo Westerlund, Marie Zins, Annie Britton, Eric J. Brunner, Archana Singh-Manoux, G. David Batty, and Mika Kivimäki (2015) Job strain and the risk of stroke: an individual-participant data meta-analysis. Stroke, 46(2), 557-559.