Recommendation of Medical Exams to Support Clinical Diagnosis Based on Patient's Symptoms

Cristiana Neto[1], Diana Ferreira[1], Hugo Cunha[2], Maria Pires[2], Susana Marques[2], Regina Sousa[1], and José Machado[3](✉)

[1] Algoritmi Research Center, University of Minho, 4710 Braga, Portugal
{cristiana.neto,diana.ferreira,regina.sousa}@algoritmi.uminho.pt
[2] University of Minho, Campus Gualtar, Braga 4710, Portugal
{a84656,a86268,a84167}@alunos.uminho.pt
[3] Department of Informatics, University of Minho, 4710 Braga, Portugal
jmac@di.uminho.pt

Abstract. Nowadays, it is essential that the error in the decisions made by health professionals is as small as possible. This applies to any medical area, including the recommendation of medical exams based on certain symptoms for the diagnosis of diseases. This study aims to explore the use of different Machine Learning techniques to increase the confidence of the medical exams prescribed by healthcare professionals. A successful implementation of this proposal could reduce the probability of medical errors in what concerns the prescription of medical exams and, consequently, the diagnosis of medical conditions. Thus, in this paper, six Machine Learning models were applied and optimized, namely, RF, DT, k-NN, NB, SVM and RNN, in order to find the most suitable model for the problem at hand. The results obtained with this study were promising, achieving high accuracy values with RF, DT and k-NN.

Keywords: Recommender System · Medical Exams · CRISP-DM · Classification

1 Introduction

The field of medicine is possibly the one that presents the greatest challenges in the integration of machine learning techniques. Starting from the basis of learning, these challenges lie in the nonexistence of datasets and in the difficulty of creating them due to inherent privacy issues. Clinical data are of a highly complex nature and are regularly incomplete and coming from various sources so they do not follow the same standards and records for the same problem may be incompatible [6].

The decisions made by health care professionals in a clinical diagnosis have direct impact on patients' treatment outcome. Due to the accelerated medical and technological growth, new options appear regularly, resulting in difficulties in

ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023

J. M. Machado and H. Peixoto (Eds.): AISCOVID 2022, LNICST 485, pp. 91–100, 2023.
https://doi.org/10.1007/978-3-031-38204-8_8

choosing the most appropriate exams for patients [16]. Thus, the need to create recommender systems in order to assist professionals in the decision-making process becomes evident. In a generic way, a recommender system can be defined as a system that guides users in a personalized way to interesting or useful objects in a large space of possible objects or produces such objects as outputs [5]. In a medical perspective these objects can be the medical examinations that the patients will have to undergo and the users the health care professionals who will have to prescribe them.

One of the most important workflows in a hospital environment, that can be enhanced by the referred technologies, is the CMD (Complementary Means of Diagnosis) workflow. This workflow ranges from the request of the CMDs, to their scheduling and results reporting. Because of its importance, optimizing and enhancing this workflow is a key point in ensuring not only the proper functioning of the hospital institution but also a better healthcare deliver.

In this context, the current research has emerged, consisting in the development and exploration of machine learning algorithms for decision support in recommender exams for patients according to their symptoms. It should be noted that the recommendation is based on their symptoms only and not on their diseases, making the problem at hand more difficult since the intended goal is creating algorithms that can make a sort of intermediate diagnosis, managing to map the symptoms to the necessary exams without the need to take the intermediate step which is to think about which diseases the patient may have.

In this way, this study contributes to the optimization of the CMD workflow, since its successful implementation will reduce the prescription of unnecessary exams, as well as the overload of medical equipment, and consequently will improve the financial burden of the health institution.

The remainder of this paper is organized as follows: the next section presents an analysis of research papers related to the topic addressed in this study; Sect. 3 presents the methodology process carried out; Sect. 4 presents and discusses the results obtained; finally, Sect. 5 outlines the main contributions of this study and some ideas for future work.

2 Related Work

There are several articles and sources available on recommender systems in health care and on the respective best methods for mapping symptoms to diseases. This article has a greater goal and intends to go further by mapping the symptoms directly to the intended medical exams, not requiring a diagnosis beforehand. There is a greater scarcity of research in this particular area, and it is intended that the research carried out and exposed in this article can be an asset for future research work and a good basis for further advances.

Focusing first on general recommender systems in the health field, it is common knowledge that there are treatments for diseases that can be time consuming and a great monetary burden. To avoid this, there is a need for systems that can detect disease symptoms as quickly as possible and even help professionals

make better choices when treating patients. Thus, recommender systems have already been proposed to predict risk factors (such as possible complications or future illnesses) that a certain patient with a chronic disease may face in the future [10,13].

In these particular systems, it is applied Collaborative Filtering, which recommends items to a user based on the following idea: "If users shared the same interests in the past, then they would have similar tastes". This approach can be interpreted in the context of these systems as follows: "Patients who share similar diseases and health status might face the same risk factors" [16]. Similarly, IBM's artificial intelligence machine, Watson Health, is already able to recommend suitable treatments for patients based on the outcomes of other patients and evidence-based medicine. According to IBM, 81% of healthcare executives who are familiar with Watson Health believe that it has a positive impact on their business [15]. This demonstrates that using technology and analytics has become increasingly important in healthcare.

The research and development of predictions in the domain of medical examinations is still quite early and has not been extensively explored. We argue that this is due to the specifics of bench-marking criteria in medical scenarios and the enormous context complexity of the medical domain. Risk perceptions towards data security and privacy, as well as trust in safe technical systems play a central role particularly in the clinical context. These aspects predominate in the acceptance of such systems.

3 Methodology

The benchmarking process followed the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, one of the most popular methodologies used in Machine Learning and Data Mining projects worldwide.

Figure 1 illustrates the methodology's life cycle, dividing it into six different stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment [8].

In the following subsections, each phase of the CRISP-DM will be discussed in more detail.

3.1 Business Understanding

Considering the complexity of medical data, as it is often unstructured, incomplete, non-standardized and stems from various sources or because large parts of data are not generated in a computer (as typical recommendations are) but stem from paper-based health-records that are often digitized afterwards, it is a challenge to provide accurate medical recommendations.

At this stage of the project, it is important to have a clear understanding of the main goals in order to ensure that the process is carried out rigorously and that an efficient recommendation system is therefore achieved. Hence, a set of goals were established for this study: to achieve a competitive advantage in the

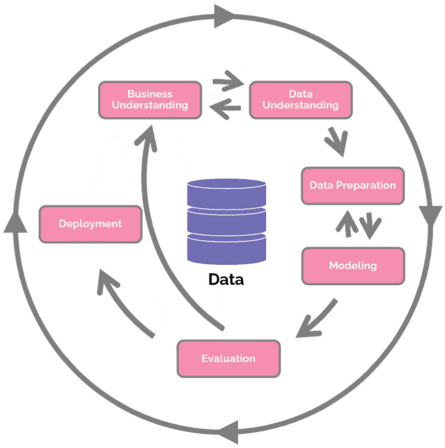

Fig. 1. Stages of the CRISP-DM Methodology.

evaluation of medical exams by health professionals; to be able to make personalized recommendations taking into account the type and number of symptoms; and to find the recommendation algorithm and parameterization that leads to the highest overall performance in the recommendation system.

3.2 Data Understanding

The medical sources used in this study are found in literature as open access for research purposes. Therefore, the data used in this study comes from the Disease Symptom Prediction [14] dataset, which is publicly available. The first dataset contains 4920 entries regarding 41 diseases and the symptoms experienced by different subjects suffering from that disease. There are 131 different symptoms in the dataset and for every disease, there is exactly 120 entries with combinations of symptoms experienced. This dataset also included an association of every symptom to a severity weight on a scale from one to seven. Due to the sensitivity of medical data and lack of datasets the mapping of the exams required to detect diseases had to be done manually through intensive research of reliable medical sources. Thus, a second dataset mapping every disease, present in the first dataset, was created. The number of medical exams per diseases ranges from two to seven, as it can be seen in Fig. 2. Some diseases are linked to very specific sets of medical exams while others require nothing more than a simple physical examination or a blood test to be detected by the professional. In the end, 102 different medical exams were compiled. Figure 3 displays the most common exams prescribed, where it is visible that prescribing blood tests and being physically examined by an health care professional is an important standard practice to understand and guide further diagnosis.

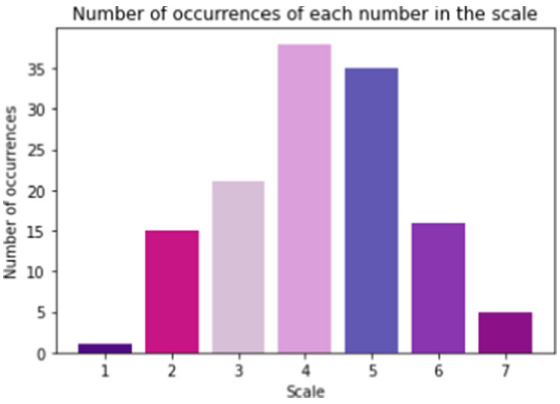

Fig. 2. Distribution of the severity of the diseases.

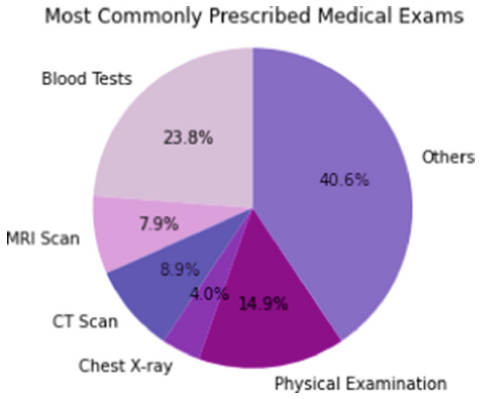

Fig. 3. Most Prescribed Medical Exams.

3.3 Data Preparation

Preparing the data is a crucial step to maximize the performance of the chosen models, therefore the chosen datasets had to be merged into one large dataset and the raw data properly treated to be able to be fed to the models afterwards.

The first step in this phase, consisted in the transformation of each symptom to a column using one-hot enconding. Consequently, it was associated the value 0 in the cases the symptom did not applied to the disease and the value 1 otherwise. After this step, since the public dataset used included a severity scale associated with the symptoms, each symptom in the dataset was mapped to its respective severity. Then, the constructed dataset that included the exams associated with each disease was also transformed using one-hot enconding and later merged with the previous dataset, using the disease as a common attribute in this operation.

After merging all the data, since it was intended to abstract the disease, which is the common factor of both symptoms and medical exams, this attribute was eliminated as it was not relevant to the study. Finally, the dataset was randomly divided using 80% of the data to train the models and 20% to test them.

3.4 Modeling

In this stage, six Machine Learning models were implemented and the hyper-parameters were optimized in order to determine which model performs best in the evaluation/results stage. The models included in this study were: Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (k-NN), Naïve Bayes (NB), Support Vector Machine (SVM) and Recurrent Neural Network (RNN). An exhaustive search over specified parameter values was performed for the first five algorithms.

Decision Tree. The goal of a DT is to create a training model that can predict the class or value of a target variable by learning simple decision rules inferred from the features of the training data. In DTs, in order to predict a class label for a record, we start with the rules at the root of the tree [8,12]. We compare the values of the root attribute with the values of the record's attribute, then we follow the branch corresponding to that value and afterwords we move on to the next node. For optimization purposes, we used Gini's index and entropy as criteria to measure which features should be in the nodes. The actual feature for each node can be chosen randomly from a distribution based on the metric of each feature or just by choosing the one with the best value, the thresholds picked in each node are always the most optimal. There is no pruning applied to the tree, however we tested limiting the tree growth to a max of 1, 10 or 20 nodes of depth as well as without a depth limit.

Random Forest. RF is a meta-estimator that uses the average of multiple DT classifiers, where each DT can be trained with a fraction of the dataset. Similarly to the DT classifier, the growth can be limited to 1, 10 or 20 nodes. The premise of this algorithm is that by using multiple tree classifiers with high randomness we can reduce the overall variance perceived in the resulting classification [11]. Since this is a meta-estimator, it is possible to define how many simple estimators to use. In this study, we evaluated the performance with 10, 20, 30, 50, 100, 200 and 1000 classifiers.

K-Nearest Neighbours. The k-NN classifier is a type of instance-based learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the K nearest neighbors of the point.

In this study, the number of neighbors tested was 3, 5, 7, 11, 13, 15, 17, and 25. The value assigned to a query point when using uniform weights is computed from a simple majority vote of the nearest neighbors; however, when using weights based on distance, it assigns weights proportional to the inverse of the distance from the query point. The distance can be calculated using either the Euclidean distance formula or the Manhattan distance, and the search algorithm can use brute force or an acceleration structure such as a ball tree or KD-tree [2].

Naïve Bayes. The NB methods are a set of supervised learning algorithms based on Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable [1]. The main difference between each implementation of the NB classifier is how the likelihood of the features is computed.

Using the Gaussian formula, we can define the smoothing added to the variance as 1e−11, 1e−10 or 1e−9 of the maximum variance reported. Using the Bernoulli formula, we can specify whether the algorithm should learn the prior probabilities before training. In addition, we can also specify the additive smoothing parameter as 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9 or 1, which also works for the polynomial implementation of NB algorithm.

Support Vector Machines. SVMs are a set of supervised learning methods [4] whose goal is to find a hyperplane in an N-dimensional space, where N is the number of features, that clearly classifies the data points. Because SVMs are binary classifiers, N-binary classifier models had to be generated before they could be used in this context. One-vs-Rest is a heuristic method that fits each classifier against all the other classes [3] to achieve a multi-class classification.

Three different kernels were tested `poly, rbf, linear`, along with the optimization of the gamma value, which defines how far the influence of a single training example reaches the c value, which is the regularization parameter that controls miss-classification, and the class weight, which affects directly the c value.

Recurrent Neural Network. Neural Networks (NNs), commonly known as Artificial Neural Networks (ANNs) are a part of machine learning and are at the centre of deep learning algorithms. Their structure and nomenclature are based on the human brain, mirroring the communication between organic neurons. ANNs are comprised of node layers, containing an input layer, one or more hidden layers, and an output layer. With input data, weights, a bias (or threshold), and an output, each node represents a separate linear regression model. Before delivering information to the network's next layer, each node's output goes through a non-linear activation function. Otherwise, no data is transmitted to the network's next layer. Large volumes of training data are essential for neural networks to develop and enhance their accuracy over time [9].

Generally, neural networks perform tasks involving supervised learning, learning from data sets where the right answer has already been chosen. The

networks then improve the accuracy of their forecasts by fine-tuning themselves to identify the proper response on their own. To achieve this, the network compares initial outputs with a given correct target. Depending on how much the initial outputs deviated from the goal values, a cost function is employed to adjust them. A crucial step in how a neural network learns a specific task is the back propagation across all neurons and connections to modify the biases and weights [7].

3.5 Evaluation

The goal of this study is to recommend a list of possible medical exams for a set of symptoms. The output of the neural network is not a list of medical exams. Because the network's output does not always converge to a binary list, the results must be processed before they can be evaluated. Starting with an array of booleans encoded with the one hot encoding standard, we apply a form of thresholding by averaging the maximum and minimum values in the resulting array.

In terms of evaluation, several metrics could be used, however when it comes to a medical recommender system, it is important to choose metrics that are truly useful and return relevant information. In this stage, we discussed the difference between the impact a false positive and a false negative, so we thought it would be interesting to create a new metric that shows simple statistics about the correctness of each answer.

Although the ultimate goal is to achieve 100% precision on every prediction, a 95% correct answer is considered extremely precise. With precision less than 95%, the error becomes overwhelming, so the average of the answers should not round this value. Hence, results less than 95% correctness should be accepted only in the context of watching the model evolve, as the distribution of the results is expected to shift from an average of 70% correct matches to greater than 95%.

It is important to keep in mind that the list of possible exams is not much more than 100 which means that a 90% correctness implies that the model recommended incorrectly over 10 exams, which might be catastrophic in a clinical setting depending on the patient's condition and on the healthcare professional's interpretation.

4 Results and Discussion

The results obtained in this study are compiled on Table 1, referring to the algorithm with the best parameters found by performing grid search. It can be observed that k-NN is the quickest model, taking only 3 s per fold and reaching 100% for both accuracy metrics. SVM is the second quickest with only 4 s per fold. RF and DT also achieved 100% for both accuracy metrics with an execution time of 6 and 10 s per fold, respectively. As for the NB algorithm, the highest accuracy was obtained applying the Bernoulli distribution. The real accuracy

only achieved 82%, which means that in a universe of 1170 sets of tests, 86 were classified incorrectly. The incorrect classification of a set means that the recommendation system is not able to recommend the minimum required exams to detect the disease. On the other hand, the neural network's results are even lower and its training time is eighty times higher than k-NN's execution time.

Table 1. Best Results obtained using Grid Search and Cross Validation.

Model	Accuracy	Real Accuracy	Execution Time (s)
RF	100%	100%	30
DT	100%	100%	50
k-NN	100%	100%	15
NB (Bernoulli)	93%	82%	26
SVM	100%	100%	20
RNN (Optimal)	73%	62%	1200

5 Conclusion

Although recommender systems in medicine have made significant progress in recent years, continuous innovation and improvement are still required. Currently, there is a high demand for technological assistance for healthcare professionals in order for them to make the most accurate and error-free decisions possible. This study focused mostly on the implementation of different algorithms into the proposed recommender system, to serve as a launching point for future research in the field of health care, particularly, in medical exams. We discovered that using models such as RF, DT, and k-NN has great potential in this case, with an accuracy close to 100%. We argue that because of the shape and distribution of the data, these results must be carefully understood, and the way in which accuracy metrics are implemented in this situation should also be considered.

This study demonstrates that it is possible to combine human expertise (medical knowledge hidden in past decisions on the datasets under use) with computational power (artificial intelligence algorithms) to improve health care and provide a comprehensive recommendation with all of the necessary medical exams to detect a set of diseases that might be manifesting in the patient. In this case, it is advantageous to recommend as many medical exams as necessary to determine what is causing the symptoms; thus, the results are very satisfactory because the bare minimum of necessary tests is always recommended. Furthermore, the accuracy metrics developed in this study ensure that it addresses the previously mentioned concern and explains the high results obtained. Ideally, the models chosen to perform the benchmark on this study should be applied to a more diverse and realistic database, in this scenario it is expected that the RNN would perform better given the increase in the amount of data.

In the future, it would be interesting to apply the models to larger datasets to verify the generalization of the models and to improve professional criticism and patient health.

Acknowledgements. This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020. Diana Ferreira and Cristiana Neto thank the Fundação para a Ciência e Tecnologia (FCT) Portugal for the grants 2021.06308.BD and 2021.06507.BD, respectively. The grant of Regina Sousa is supported by the project "Integrated and Innovative Solutions for the well-being of people in complex urban centers" within the Project Scope NORTE-01-0145-FEDER-000086.

References

1. Naive Bayes. https://scikit-learn.org/stable/modules/naive_bayes.html
2. Nearest neighbors. https://scikit-learn.org/stable/modules/neighbors.html#neighbors
3. Onevsrestclassifier. https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html
4. SVM. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
5. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
6. Calero Valdez, A., Ziefle, M., Verbert, K., Felfernig, A., Holzinger, A.: Recommender systems for health informatics: state-of-the-art and future perspectives. In: Holzinger, A. (ed.) Machine Learning for Health Informatics. LNCS (LNAI), vol. 9605, pp. 391–414. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50478-0_20
7. Chow, T.W.S., Cho, D.S.Y.: Neural Networks and Computing: Learning Algorithms and Applications, vol. 7. World Scientific (2007)
8. Ferreira, D., Neto, C., Lopes, J., Duarte, J., Abelha, A., Machado, J.: Predicting the survival of primary biliary cholangitis patients. Appl. Sci. **12**(16), 8043 (2022)
9. Maind, S.B., Wankar, P., et al.: Research paper on basic of artificial neural network. Int. J. Recent Innov. Trends Comput. Commun. **2**(1), 96–100 (2014)
10. Nasiri, M., Minaei, B., Kiani, A.: Dynamic recommendation: disease prediction and prevention using recommender system. Int. J. Basic Sci. Med. **1**(1), 13–17 (2016)
11. Neto, C., Brito, M., Lopes, V., Peixoto, H., Abelha, A., Machado, J.: Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients. Entropy **21**(12), 1163 (2019)
12. Neto, C., Peixoto, H., Abelha, V., Abelha, A., Machado, J.: Knowledge discovery from surgical waiting lists. Procedia Comput. Sci. **121**, 1104–1111 (2017)
13. Patil, P.: Disease symptom prediction (2020). https://www.kaggle.com/itachi9604/disease-symptom-description-dataset
14. Patil, P.: Disease symptom prediction (2020). https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset
15. Stark, B., Knahl, C., Aydin, M., Elish, K.: A literature review on medicine recommender systems. Int. J. Adv. Comput. Sci. Appl. **10**(8) (2019)
16. Tran, T.N.T., Felfernig, A., Trattner, C., Holzinger, A.: Recommender systems in the healthcare domain: state-of-the-art and research issues. J. Intell. Inf. Syst. **57**(1), 171–201 (2021)