# MSc Dissertation

# Final Report

# MSc in Integrated Machine Learning Systems

**Student:** Diogo Sousa

**Student number:** 20187253

**Project Title:**

Natural Language Processing for Medical Diagnosis

University College London

Dept. of Electrical and Electronic Engineering

2020-21

# Abstract

The study presented in this paper is focused on designing and developing a machine learning system that could provide accurate and efficient medical diagnosis based on a question answering task.

Medical domain is always changing and evolving side by side with the technological field. Since early stages of computation this field was one of the main attractions of study because of its hard tasks to solve and because of its importance in everyone’s life, therefore it is highly significant to develop a system that could provide any help to physicians on more complex cases.

The number of complex cases in medicine will increase alongside the advances in technology because of higher accuracy in imaging or blood tests and therefore more new information from different sources and different evidence will be presented to the physicians when performing the diagnosis.

During this research it was found that when leading with medical data, that is very non-linear and independent, the simpler models such as Naïve Bayesian Classifier or Semi-Naïve Bayesian Classifier perform better than more complex models using artificial neural networks. For that reason, this study will start by designing and developing a machine learning algorithm based on Bayesian Network to achieve good performance, transparency and provide good explanation on decision making and then try to take the performance to higher level adding more complexity into the model.

Table of Contents

[MSc Dissertation 1](#_Toc79855657)

[Final Report 1](#_Toc79855658)

[MSc in Integrated Machine Learning Systems 1](#_Toc79855659)

[Abstract 2](#_Toc79855660)

[Introduction and Problem Statement 4](#_Toc79855661)

[Background theory 5](#_Toc79855662)

[System Requirements 5](#_Toc79855663)

[**Good Performance** 5](#_Toc79855664)

[**Data (missing and noisy)** 6](#_Toc79855665)

[**Transparency & Explanation** 6](#_Toc79855666)

[**Data privacy** 6](#_Toc79855667)

[Related work 6](#_Toc79855668)

[**Symbolic Learning** 7](#_Toc79855669)

[**Statistical Learning** 8](#_Toc79855670)

[**Neural Networks** 10](#_Toc79855671)

[**Performance** 11](#_Toc79855672)

[Methodology 13](#_Toc79855673)

[Data Selection 15](#_Toc79855674)

[Data Pre-processing 15](#_Toc79855675)

[Classifier 16](#_Toc79855676)

[Planned work to complete the project 17](#_Toc79855677)

[Results to date 18](#_Toc79855678)

[Implementation 19](#_Toc79855679)

[Data sets 19](#_Toc79855680)

[**Collection** 19](#_Toc79855681)

# Introduction and Problem Statement

The focus under study is the design and development and algorithm that would be use on chatbot for medical diagnosis.

The healthcare domain is a sector that is always under change and interested, so it is comprehensive that it has so much interested from the computation sector and so much work around has been done. Since the early steps of computation this domain was always of the most interested ones to make research on. However, recently big steps have been made by the machine learning in the medical field.

Several studies have presented algorithms to provide more personalised, accurate, efficient and wider medical solutions in all societies across the globe with the help of deep neural networks, natural language processing, computer vision or robotics.

There has been put much effort and development in these studies because medical diagnosis is one of the most important stages of the medical system. Bad decisions or misinterpretations from physicians can lead to treatment suspension or even ultimately to a loss of life. Every patient is unique and therefore provides different information even on the same disease, therefore it is important to develop am algorithm that will use inputs from the client to perform an accurate prediction on a medical diagnosis, or no diagnosis in case of no disease.

An algorithm of question-answering platform will be design, implemented and tested to improve the quality of medical diagnosis, providing the specific requirements for a model in this domain, such as transparency, data privacy, explanation and good performance, in order to provide good support to physicians.

To obtain this algorithm natural language processing techniques and statistical learning methods (Naïve Bayesian Classifier) will be used as they have shown good results in the past when dealing with medical data.

# Background theory

Since early days of computation algorithms were developed to enable modelling and analysis of data, those algorithms can be separated into three fronts of machine learning raised: Symbolic learning, statistical methods and neural networks.

Developing computer systems for medical diagnosis based on machine learning is not something that has recently started, in fact one of the most promising times for medical data analysis was in the very beginning with the usage of Concept Learning System (CLS) from Hunt et al. [1] to build decision trees for medical diagnosis and prognosis in 1966.

However, through the years a big number of disease prediction models have been studied and developed to be used in medical diagnosis with the use of data mining and machine learning techniques in all three branches, such as Symbolic Learning, Neural Networks, Naïve Bayesian classification or Regression models.

Since this is a field with high responsibility in society for a machine learning system to be considered effective and useful to be applied in helping with medical diagnosis several specific requirements need to be taken into account when trying to design and develop a system for this type of task. Those requirements are good performance, ability to deal with missing data, ability to deal with noisy data, transparency of knowledge, ability to explain decisions, ability to reduce the number of needed tests to obtain good results and protecting the privacy of data [2].

## **System Requirements**

### Good Performance

The developed algorithm must provide high quality responses from the available data and provide an accurate as possible diagnosis. This obtained performance should be better than the performance obtained by physicians when using the same data, provided by the patients. A recent study from Esteva et al [3] provides an illustrative example of a good performance algorithm in medical diagnosis with a deep convolutional neural networks system being able to classify skin cancer with a level of competence comparable to dermatologists.

### Data (missing and noisy)

A crucial challenge when building a machine learning system is acquiring a big, representative and diverse data set. The data set used to train the model should represent as good as possible the population where the model will be applied to infer on to prevent unwanted and unconscious bias.

When dealing with medical diagnosis it is necessary to guarantee complete description records from diversified patients containing all the needed information without uncertainty or errors. However, when using large data sets it is possible to apply measures for the model to be prepared to be trained with noisy data, mapping noisy inputs to noisy outputs.

### Transparency & Explanation

The outcome provided by the machine learning should be complete and provide total information on what knowledge was used to achieve that result. It is important that the physician understands the process that knowledge was generated because these tools should also be used as learning process for the physicians and not only replacement.

### 

### Data privacy

When using very confidential medical data it is important to guarantee the security of sensitive personal information and reduce the minimum possible the transfer of this data on totally unsecure channels.

## **Related work**

In this section it will be given a description and analysis of three branches mentioned above (Symbolic Learning, Statistical Learning and Neural Networks) and the most used algorithms used for each of the branches when developing machine learning systems for medical diagnosis.

### Symbolic Learning

Symbolic learning is the field of machine learning where symbolic models are explicitly induced from data in contrast with methods like neural networks that normally produce black-box models. These produced models can be inspected, modified and verified by humans with expertise in the data under analysis. For that reason, they have the potential to start being used as part of knowledge base in the domain of study. The most common methods of this branch are the usage of decision trees and decision rules to perform symbolic learning [4].

This was the method firstly used as an attempt to solve issues in medical field with help of computation. In 1966 Hunt et. Al [1] used a Concept Learning System developed by themselves in for order to build decision trees for medical diagnosis and prognosis.

The information used in this type of problems can be obtained from multiple sources and different approaches, therefore the number of records to be used can be high making it difficult to organise and understand all the information for a human investigator. The complexity of this process could then be reduced using computation to investigate by its own the patterns existent on the data selecting the possible routes much faster than a human.

Although these trees were very helpful on accelerating this process, they would become very large because there is always new information to be added from new symptoms or description from a patient’s record.

After the first implementation of Hunt et al [1] this subject became highly studied by different investigators and multiple algorithms have been developed to solve several problems in the medical domain such as Iterative Dichotomizer 3 (IDE) [5] that would be later applied to solve oncology diagnosis problems or CART applied on cardiology and oncology diagnosis and prognosis [6].

A brief description of the most used Symbolic Learning algorithms in medical domain - Assistant-R, Assistant-I and Lookahead Feature Construction (LFC) - will be presented below.

#### **Assistant-R**

This algorithm was developed as a reimplementation of the Assistant Learning System and the main difference to the latter is that a non-myopic heuristic to quantify the quality of attributes is used for attribute selection, even with highly dependent attributes [2].

#### **Assistant-I**

Assistant-I is an extension of Assistant-R that uses information gain as criteria picker.

#### **Lookahead Feature Construction**

This algorithm is built on methods for top-down decision tree induction, generating at the end much simpler binary decision trees. To obtain this the algorithm runs projections over several attributes before deciding on the current node. At each node, new binary attributes are initiated, resulting from conjunction, disjunction or negation of the original attributes.

### 

### Statistical Learning

Statistical Learning is the application of knowledge from statistics and functional analysis into the machine learning algorithm development field. This theory generally deals with the problem of predicting based on data, using a perspective of some unknown probability distribution and using some training set to obtain this probability distribution that will in the future lead to a better prediction of future testing cases. Examples of this type of learning are Regression Models or Naïve Bayesian Classifiers.

At the beginning these methods were considered less effective when dealing with high number of non-linear and dependent data samples, which is the case of medical historical data because of the different number of sources.

The most widely used Statistical Learning algorithm in the field here under study is Naïve Bayesian Classifier because it is an extremely simple algorithm, however very efficient and effective.

#### **Naïve Bayesian Classifier**

As mentioned above this algorithm has been widely used when dealing with medical data problems because of its simplicity, efficiency and effectiveness. It can be defined as probabilistic classifier, which is built on Bayes’ Theorem, applying very independent assumptions on the data set.

These classifiers assume the presence of a specific feature in some class is not dependent on the presence of any other feature, because of this assumption it aims to the probabilities of assigned classes using joint probabilities.

Bayesian formula calculates the probability of a sample to be classified on class C, given the values Vi of all the features for that sample:

Despite independence being a poor assumption, this algorithm is able to compete very well with much more complex algorithms, sometimes even when dealing with data sets with some feature dependence [7]. Naïve Bayesian algorithms have shown good performances in text classification, medical diagnosis and systems performance management.

The main advantage that this algorithm provides, generally and especially for medical diagnosis, is that all the information available in the model is used to make the final prediction and therefore the explanation is simpler and, because of that, well received from physicians. All attributes are given a strength positive or negative for a given class and the decision can be explained with sum of all these strengths for that class, therefore the usage of this algorithm to solve medical diagnosis tasks is very well accepted by physicians because it is a similar process to what they normally do when solving the same task.

In 1991 Kononenko [8], developed and extension of Naïve Bayesian Classifier, called Semi-Naïve Bayesian Classifier. This algorithm is different from previous existent classifier because it explicitly tries to find dependencies between values of different features. If a relation is found between those two values, then they cannot be considered as conditionally independent. The performance for this algorithm is traded-off between knowledge of dependency and reliability of approximation of probabilities [2]. The term for the probability of a sample to be assigned class C is now:

For decision making, in this case, the calculation of information gains is very similar with the exception when a relation is found and, in that case, instead of attribute values, the process uses the joined values.

### 

### Neural Networks

Neural networks were first developed in 1962 by Rosenblatt [9], using single-layered perceptrons. This topic was after proved as not being able to solve non-linear problems and almost every scientist dropped their studies in this field. Neural networks gained more attention when an algorithm with backpropagation learning rule for multi-layered feed forward neural networks was developed because this solution would be used to solve hard tasks in the medical diagnosis.

However, physicians were not very receptive to learn from this model because this type of model are normally used as black box classifiers, not providing the transparency needed of how the knowledges was generated. The decision on these models is obtained adjusting the weights between neurons and determine the appropriate values for those connections.

Artificial Neural Networks are computational models that tries to use a similar nature to the human brain. It is a network highly connected with elements operating in parallel. Each sub-group of this elements is called layer.

The first layer in the model is called input layer and the last layer is called output layer. In the middle of these two layers, it is possible to add additional layers, those are called hidden layers.

The most used widely used neural networks algorithm is Feed Forward because of its successful performance in classification, forecasting and problem solving. Feed Forward Normally this type of algorithm includes a backpropagation of the error technique. However, this technique arises problems of overfitting. Overfitting is when a model is trained deeply to that specific data set and learns the data set and not the patterns existent on it, therefore it will not predict accurately when facing a completely new data set on testing phase.

To overcome this issue, it is possible to use a weight elimination technique, which will penalise high weights on connections between neurons, removing them and reducing model’s complexity, enabling the obtention of accurate information on the training data set to improve the model’s performance.

### Performance

Kononenko [2] compared the performance of seven algorithms using eight medical data sets, evaluating some system requirements mentioned above. The general results were that Naïve Bayesian and Semi-Naïve Bayesian classifiers obtained the best performance overall, being very good at the explanation requirement and handling missing data. While in terms of performance they performed at the same level as Neural Networks and k-NN (k-Nearest Neighbours).

Regarding transparency, symbolic learning algorithms (Assistant-R and Assistant-I) obtained a very good result being the best from the 7. On the bottom of overall performance was Neural Networks using backpropagation, having a very good performance however a poor transparency and explanation. These requirements are as important as the performance to be accepted by the physicians and patients.

Another study was done by Ster and Dobnikar [10] and the results were also favourable to Naïve Bayesian and Semi-Naïve Bayesian classifiers obtaining the highest score in three of the five used data sets.

Both these studies show that high complexity models do not contribute much to classification accuracy, despite the high consume of resources, when dealing with medical data sets and are not as good as assisting physicians since they act as black boxes. Therefore, the best models to use in this study are statistical learning models, specifically Naïve Bayesian Classifier and Semi-Naïve Bayesian Classifier. Those are the models that perform best and the ones that provide higher assistance to physicians with complete explanation and transparency of decision making.

Recently the medical domain has seen an increasing in the number of studies using hybrid data mining approaches, in order to improve the performance, which means using multiple models from different branches to perform the same tasks, such as artificial neural networks with multivariate adaptive regression splines, evolutionary decision trees, fuzzy AIS with k-NN or polynomial fuzzy decision tree.

However, for now, these techniques do not show much higher performance when comparing with simple Naive Bayesian Classifier or Semi-Naïve Bayesian Classifier for the complexity that these models represent. For that reason, when dealing with problems within the medical domain it is better to start with one of these models as base and then increase the complexity if the performance is achieved, therefore in this study we will initially focus on presenting a good performance with Naïve Bayesian Classifier on medical diagnosis.

# Methodology

In this section are presented the methods used and steps performed to address the questions raised earlier, that is concluded with a machine learning system that will provide help to physicians on the medical diagnosis. This system will provide information on the most probable medical diagnosis when symptoms are given as input in the form of text.

To answer the questions under study the methodology implemented is based on the knowledge acquired during the literature review section, with simple adaptions in order to be more adequate to the time and resources available in this research.

Before starting the modelling of our solution for this problem there was debate on several available options to obtain the pretended results. The first option considered was the manually construction of a Bayesian Network that will then be traverse to obtain the most accurate diagnosis. Another available modelling solution is to give the symptoms as input and train a Naïve Bayes Classifier than will then, given a certain list of symptoms, provide the most probable diagnosis.

In order to construct a Bayesian Network (BN) manually it is necessary to have previous knowledge of experts, in this case physicians, and it will have high complexity and be very time consuming to construct. On the other hand, the increasing clinical data available would allow for the network to be more accurate and reliable, since the human knowledge is, in several cases of this field, more concise and robust than the information collected from a data set of limited size.

The increasing of data is also the most common point in favour to choose the network’s learning with algorithmic methods. Recent studies show that when trying to solve classification problems, like the one under study in this research, the usage of Naïve Classifiers perform better than more complex networks, constructed manually [11]. For that reason, the approach selected to model the solution was to learn the Bayesian Network with Naïve Classifiers algorithms.

Natural Language Processing (NLP) is a subfield of linguistics with artificial intelligence concerning the interactions between humans and computers, using human language. Particularly how computers process human language to obtain the insights laying on the speech. The solution developed in this study is based in NLP because it is pretended that the system will process the text introduced by humans and return the response in human readable information too.

Both approaches described above need to take NLP into the pipeline. However, the selected approach is less complex to incorporate with NLP since Naïve Classifiers, like most of machine learning models, must be given processed data as they expect numerical feature vectors with a fixed sized instead of raw text data with variable length.

The figure below (Figure 1) shows the steps taken to obtain the final machine learning solution.

**Diagram

Description automatically generated**

Figure 1 - Solution pipeline

## **Data Selection**

The first step of our modelling is to define the data of interest, that will be used to train the machine learning system. In the case of our study the focus relies on medical data. Recent studies in the field have produced solutions with similar goals, however for specific diseases or medicine subfields like cancer or heart conditions.

The aim of the solution proposed in this study is to offer the possibility to obtain information on as many diseases and medical fields as possible and, therefore it is necessary to select data from multiple sources and with multiple schemas.

After a process of analysis of medical data is was decided that the most useful information was provided by previously constructed data sets available on UCI Machine Learning Repository [12], on Kaggle [13], medical data retrieved through PyMed python library [14], that provides access to PubMed [15] data via PubMed API and data acquired directly from WebMD [16] and Mayo Clinic [17] websites.

This data was collected in the form of .csv files containing symptoms as features to the corresponding medical diagnosis.

## **Data Pre-processing**

As mentioned above the data collected is from several sources and, therefore it is presented in multiple schemas, data types and data values.

During the processing phase, the first step is to verify missing values on the data acquired and take an action regarding corresponding entries. Missing values can be related to human error or machine error if the entries are recorded by any device. In medical data field, recently, several tools based on artificial intelligence have been developed to help physicians on medical records through speech recognition. This type of data set construction can lead to failing values in the presence of system error. On the other hand, some data sets acquired have binary information regarding presence of symptoms. In this case missing values can be related to human error and the miss of a value entry.

Regarding the last mentioned type of data sets need to be process in such way that the features are changed to a list of symptoms instead of having each possible symptom as a feature and then use binary values to mark the presence or absence of that symptom in the corresponding diagnosis.

In the processing phase it is also where the data representation is simplified to be used as input for the training model, as mentioned above this process is called NLP and it can be based on Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF).

Bag-of-words refers to the creation of set of all the existent words in a data set, where each word is tokenized and has its own numeric representation. This process is less complex and requires less resources to implement, for that reason the initial approach of our system will implement this technique.

In this technique, after each symptom has its own numeric representation a counting of each token occurrences in each document is done and the resulting list of occurrences is the new feature for that document and will be used as input to train and test the model.

This process is done once for all system and then the same tokenization is used for each test done too.

## **Classifier**

The classifier to be used in this research will follow the knowledge acquired from the previous work on the field described on literature review chapter and, therefore a Naïve Bayes Classifier will be developed.

However, the system will also be developed applying Artificial Neural Networks to compare the results obtained and assess the advantages and disadvantages of each model in this task. When selecting the final classifier to use in the system it is necessary analyse the trade-offs on the system requirements identified in the literature review such as good performance, handling data missing and noisy data and transparency and explanation.

In the Artificial Neural Networks field of machine learning algorithms this proposal will develop a Feedforward Neural Network with Backpropagation, since recent studies show evolution on this field and good results when using medical data.

# Planned work to complete the project

In order to complete the project achieving the desired goals it is needed to select the most adequate classifier for this data. The process will start with Multinomial Naïve Bayes Classifier, moving to another Naïve Classifiers like Gaussian Naïve Bayes Classifier, Complement Naïve Bayes Classifier and Bernoulli Naïve Bayes Classifier and verify which classifier obtains the best score. The timeline to achieve this goal is 10th of July.

After this first it will also be tested the solution using Artificial Neural Networks to have a comparison of results and verify the hypothesis stated in introduction of this research.

All these tests will be done using a testing data set after testing the model with a training data set and validating with a validation data set, then the database will be incremented to have more information available and increase the score of the system. It is expected for the database to increase by 20th of July.

Finally, the last step of this system will be the development of the chat feature that will rely on the continuous reading text input from the stdin and reply in real time with the most probable medical diagnosis or ask for more symptoms if the information provided is not sufficient to provide a diagnosis with at least 50% of probability. This feature is aimed to be concluded on 31st of July.

# Results to date

Medical machine learning solutions require high quantity of data, which needs to be as much verified as possible, in order to produce results that would obtain recognition and acceptance by the medical community. For that reason, most of the time spent until now on this research was on searching for useful data that could be used to train the model.

As mentioned above the aim of this proposal is to provide assistance in multiple medical subfields instead of focusing in one specific area of medicine, therefore it is necessary to collect data on all those subfields. It was a decision done at the beginning of the research to focus on obtaining this data prior to the development, this decision might have cost more concrete results to the date and could be better to focus on some specific area at the beginning to start developing the model.

Data focusing on several diseases was collected and to start the system modelling it was used one data set that allows the classification of multiple subfields and other data sets focusing on specific fields were also collected but not used in the first stage. These data sets will, after, be used to tune the model.

The data set used was processed in order to have one feature that includes a list of symptoms in human text separated by space and a corresponding classification, in this case the name of the disease.

After the pre-processing stage a simple bag of words was created for the existing data because, as mentioned earlier, the classifiers need to be given numerical vectors instead of raw text. This will change the symptoms feature values to be changed from the raw words to the corresponding tokens for the same words.

# Implementation

In this section it is presented the work developed to achieve the final solution proposed for the problem under study. It is divided in sub-sections such as Data sets, Training pipeline and Hyper-parameter tuning, Learning process and Classification.

## 

## **Data sets**

### Collection

As mentioned earlier for this project multiple sources of data were selected to provide a broader range of diseases and different samples of symptoms for the same disease. Being this project related to providing health assistance and having the goal of obtaining accurate and reliable performance it is mandatory to have reliable data sets and knowledge on this specific field.

However, for the same reason it is also extremely difficult to obtain this type of data from existent solutions or previous studies, because it is of high importance that all the health records keep the personal information in private, and for most of the existent solutions the data used is based in real data acquired from real clinical records.

Due to lack of official information from similar solutions, the search for data moved to the available free data from related work, but it was paid special attention to acquire data from certified sources such as UCI Machine Learning Repository, Kaggle competitions, WebMD [16] or Mayo Clinic [17]. The data was collected through two different techniques: .csv files download, previously created and used for other purposes on healthcare study, and .csv files created by hand to expand the final data set size.

After initial search it was possible to observe that the available data sets from these sources were very specific to certain medical fields like Cancer or Image processing, or more recent data sets were only focusing on COVID-19 data, which is comprehensible since those are the areas, for different reasons, where more attention and resources were put into recently. Nevertheless, these sources provided some useful data sets for this study namely *Dermatology Data set* [17]*, Acute Inflammations Data Set* [18]and *Disease Symptom Prediction* [19]*.*

These data sets were all from different previous studies on this subject and, for that reason they have all different characteristics in terms of features, sample sizes and number of samples.

Despite collecting good data from these sources, the amount of data was not enough for the goals proposed at the beginning of this study, in such way that the initial test of model implementation provided an accuracy score of 100% which would not be possible in a normal environment. The main reason for this result was the data set size and simplicity of the data used to train and test the model.

For that reason, there was evidence that more data was needed to improve the quality of the product, increasing the number of diseases covered by diagnosis.

After deep search for more existing data sets that would provide more useful data, it was decided to create a new data set from scratch, giving the opportunity to design the feature set accordingly and add the number a specific number of instances for each of the selected diseases.

In order to create the new data set, WebMD website was accessed manually, using its *Health A-Z* section to obtain the symptoms for each specific disease. To select the subset of the symptoms most commonly and less commonly present with the correspondent diagnosis was extremely important the background knowledge acquired by me from my BSc in Nursing, helping on creating more accurate data set. The final data set acquired from WebMD added 56 new diseases to the count and adding more than 6000 new entries, providing information for a more accurate solution in the end.

After combining all information acquired from the different sources, the final data set used to train and test the model have more than 11000 entries regarding 90 diseases. Having 120 instances for each disease in average in order to guarantee a data set as balanced as possible.

Each disease has multiple combinations of possible symptoms following medical sources to give more occurrences to symptoms that are more commonly associated with the corresponding diagnosis.

The following table presents the diseases used in the final data set and the number of samples for each disease.

|  |  |  |  |
| --- | --- | --- | --- |
| Disease | Samples | Disease | Samples |
| Achalasia | 120 | Heart attack | 230 |
| Acne | 120 | Heart failure | 140 |
| AIDS | 120 | Hemorrhoids | 200 |
| Alcoholic hepatitis | 120 | Hepatitis A | 275 |
| Allergy | 120 | Hepatitis B | 280 |
| Alzheimer’s Disease | 60 | Hepatitis C | 340 |
| Anemia | 120 | Hepatitis D | 120 |
| Angina | 120 | Hepatitis E | 120 |
| Ankylosing spondylitits | 90 | Hypertension | 235 |
| Anorexia nervosa | 120 | Hyperthyroidism | 120 |
| Appendicitis | 120 | Hypoglycemia | 120 |
| Arrhythmia | 120 | Hypothyroidism | 120 |
| Arthritis | 230 | Impetigo | 120 |
| Asthma | 230 | Infectious arthritis | 50 |
| Atopic dermatitis | 190 | Irritable bowel syndrome | 105 |
| Atrial fibrillation | 40 | Jaundice | 120 |
| Bipolar disorder | 120 | Kidney stone | 130 |
| Carpal tunnel syndrome | 120 | Laryngitis | 85 |
| Cataracts | 70 | Lichen planus | 70 |
| Cervical spondylosis | 120 | Lyme disease | 120 |
| Chicken pox | 120 | Malaria | 240 |
| Chronic cholestasis | 120 | Migraine | 240 |
| Chronic dermatitis | 50 | Multiple sclerosis | 90 |
| Cold | 230 | Myocarditis | 130 |
| Concussion | 120 | Narcolepsy | 100 |
| COPD | 105 | Nephritis | 60 |
| Coronary artery disease | 90 | Osteoarthritis | 120 |
| Crohn’s disease | 120 | Paralysis (brain hemorrhage) | 120 |
| Deep vein thrombosis | 100 | Parkinson’s disease | 60 |
| Degenerative disk disease | 30 | Paroxysmal positional vertigo | 120 |
| Dengue | 120 | Peptic ulcer disease | 120 |
| Depression | 120 | Pityriasis rosea | 50 |
| Diabetes I | 120 | Pityriasis rubra pilaris | 20 |
| Diabetes II | 240 | Pneumonia | 120 |
| Diarrhea | 120 | Psoriasis | 230 |
| Diverticulitis | 60 | Schizophrenia | 120 |
| Drug reaction | 120 | Sciatica | 100 |
| Ear infection | 150 | Seborrheic dermatitis | 60 |
| Endometriosis | 140 | Sinus infection | 120 |
| Fibromyalgia | 140 | Sleep apnea | 90 |
| Flu | 120 | Tuberculosis | 120 |
| Food poisoning | 100 | Typhoid | 120 |
| Fungal infection | 120 | Ulcerative colitis | 110 |
| Gastroenteritis | 120 | Urinary tract infection | 80 |
| GERD | 242 | Varicose veins | 120 |

Table 1 - Disease data set

In addition to performing a prediction of diagnosis based on a list of symptoms provided by the user, the system developed also provides some information regarding the diagnosis and what are the next actions to take in order to get treatment or a more concise diagnosis.

To provide this information more data had to be collected for each of the diseases presented in the final data set. This data was collected manually from both Mayo Clinic [17] and WebMD [16] websites and stored in a file called disease.csv, which would have only 3 columns of data (Disease, Description, Action) and 90 rows corresponding to each one of the diseases presented in the table above.

### Pre-processing

jjkk

Bibliography

|  |  |
| --- | --- |
| [1] | E. Hunt, J. Martin and P. Stone, “Experiments in Induction,” *Academic Press,* 1966. |
| [2] | I. Knonenko, “Machine Learning for Medical Diagnosis: History, State of the Art and Prespective,” *Artificial Intelligence in Medicine,* vol. 23, pp. 89-109, 2001. |
| [3] | A. Esteva, B. Kuprel, A. Novoa, J. Ko, S. Swetter, M. Blau and S. Thrun, “Dermatologist‐level classification of skin cancer with deep neural networks,” *Nature,* vol. 542, pp. 115-118, 2017. |
| [4] | S. Dzeroski, “Application of symbolic machine learning to ecological modelling,” *Ecological modelling,* vol. 146, no. 1-3, pp. 263-273, 2001. |
| [5] | J. Quinlan, “Discovering rules from large collections of examples,” *Expert systems in the microelectronic age,* 1979. |
| [6] | L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and regression trees, Belmont, 1984. |
| [7] | P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss,” pp. 103-130, 1997. |
| [8] | I. Knonenko, “Semi-naive Bayesian classifier,” in *European Working Session on Learning-91*, Berlin, 1991. |
| [9] | F. Rosenblatt, Principles of neurodynamics, Washington: Spartan Books, 1962. |
| [10] | B. Ster and A. Dobnikar, “Neural networks in medical diagnosis: Comparison with other methods,” 1996. |
| [11] | P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian network classiﬁer,” *Mach Learn,* no. 29, pp. 103-130, 1997. |
| [12] | UCI, “Machine Learning Repository,” [Online]. Available: https://archive.ics.uci.edu/ml/index.php. [Accessed 2021]. |
| [13] | Kaggle, “Symptom Checker,” [Online]. Available: https://www.kaggle.com/rabisingh/symptom-checker?select=Training.csv. [Accessed 2021]. |
| [14] | G. Wobben, “PyMed,” [Online]. Available: https://pypi.org/project/pymed/. [Accessed 2021]. |
| [15] | NIH, “National Library of Medicine,” [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/. [Accessed 2021]. |
| [16] | W. M. Team, “Health A-Z,” [Online]. Available: https://www.webmd.com/a-to-z-guides/common-topics. [Accessed 08 2021]. |
| [17] | N. Ilter and H. Guvenir, “Dermatology Data set,” 1998 01 01. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Dermatology. [Accessed 07 2021]. |
| [18] | A. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh and J. Weston, “ParlAI: A Dialog Research Software Platform,” *Facebook AI Reasearch,* 8 March 2018. |

**Appendices**

(No limit to the number of pages)

Add appendices with additional results, proofs, etc., as appropriate to your project. Listing of source codes is discouraged unless the particular portions of the source code are critical to the development of your project and they are discussed in the main part of your text.