

**Course Title : Master’s in Data Analytics**

**Lecturer : David Williams & Amir Sajad Esmaeily**

**Module : B9DA105**

**Module Name : Applied Research Process for Data Analytics**

**Proposed Title : Disease symptoms identification using machine leaning based chatbot**

**Student Name : Jayesh Aswar**

**Student No : 10531366**

**Supervisors : Abhishek Kaushik, Dr. Shahram Azizi**

**Date : March 2020**

**Index**

**1. Introduction** …………………………….**3**

**2. Research Hypothesis**  …………………………….**4**

**3. Literature Review** ………………….……….....**4**

**4. Methodologies**  …………………………….**5**

* 1. **Process to be implemented in classification**
  2. **Data Pre-processing**
  3. **Advanced methods for Analytics**
  4. **Web Development Framework**

**5. Risk analysis** .………………………………….**7**

**6. References** .………………………………….**8**

**7. Gantt Chart: Milestone & Deliverable of the project …………………….8**

1. **Introduction:**

As we know about the value of our health, most specially during this covid-19 outbreak where whole world is facing health problems, due to different kind of diseases we face health problems. One of the deadliest health problems the world is facing is cardiovascular diseases. Every year millions of people loss their life due to these kind of health problems. As problems arise world always try to solve those problems either by medically or technological and use hybrid solutions with alternatives. There are so many medical and health related problems where machine learning and artificial intelligence try to help with proper diagnosis and reduce the loss of such diseases. There is so many applications of machine learning which is helpful to many diseases.

* 1. **Objectives:**

Due to covid-19 pandemic we all experienced that it is risky to go to hospital frequently for check-up of our body, so we try to build a machine learning chatbot platform for the users. So that user can get basic diagnosis of their health based on symptoms. Our main objective is to reduce the risk of cardiac arrest or any similar kind of emergency heart failure situation by giving virtual symptoms diagnosis on basic level using the chatbot platform. Based on user inputs our chatbots classify the type of disease and give suggestions based on criticality of disease. We try to facilitate a virtual assistant to suggest user on prior based information regarding physical symptoms, like for example any user has hypertension and also feeling chest pain along with blood pressure problems at that time we can classify as a cardiac problem and send warning to the user about it and also give suggestions and contact information for the best rated hospital for such problems.

My thesis mainly focuses on **checking for** **symptoms of disease and classify it using naïve bayes algorithm and make it interactive on web based chatbot platform**.

* 1. **Keywords:**
* Symptoms classification
* naive Bayes classifier method for symptoms identification
* Interactive chatbot web-based platform

1. **Research Hypothesis:**

* How to classify the text data input by user?
* How to identify and check the similarity of symptoms with user input?
* What should be the chatbot platform characteristics?
* Which web framework will work suitable for our requirement?
* What is the most optimum way of converting text to tokens and then running an algorithm with that data?

1. **Literature Review:**

[1] <https://www.semanticscholar.org/paper/Disease-Diagnosis-System-By-Exploring-Machine-Sunny-Kulshreshtha/035d1d9317e7d4032eb1c6f15c5f0c7925f8e6fd>

This research paper gives broad idea about the how to deal with pure text data as, hospital related symptoms datasets are in pure text based. So that we can get best accuracy according to our hypothesis without convert it to numerical format.

It gives clear idea about the algorithmic techniques and tools to use along with most suitable algorithm with experiments. It describes various types of requirements as well like functional, performance and user characteristics. It also gives information of system design and architecture flow.

[2] <https://ieeexplore.ieee.org/document/8251305>

This paper gives us idea about application of various algorithm with different accuracies to classify the diseases and forecasting through data analytics.

[3] <https://dialogflow.com/docs>

This documentation gives idea about set up of google dialogflow platform for chatbot purpose. It provides wide range of API information to implement a chatbot into a web platform.

1. **Methodologies:**

To achieve our hypothesis task, we need to follow several steps with dataset.

* 1. **Process to be implemented in Classification:**

**A picture containing clock

Description automatically generated**

Above diagram shows how a text data get process in various stages of machine learning, like first we extract the feature vector then application of algorithms and labeling also get into process. Thus, all this information gives overall idea of process for to achieve the classification of text data.

* 1. **Data Pre-processing:**

For any kind of machine learning application data pre-processing task is very important. Every kind of datasets has null values or missing values which needs to process. Apart from that there is label encoding process which also convert data to suitable format of machine learning. Data normalization is also part of the pre-processing.

* 1. **Advanced Analytics:**

After preprocessing step, we move on towards algorithms selection process, on clean data which algorithm is more suitable and which is not suitable that thing we can decide based on experiment.

We first try the K-nearest neighbors algorithm which gives very less accuracy score, so that we can say that it is not most suitable technique to classify our data. Same kind of results we get it into decision tree algorithm technique.

A close up of a device

Description automatically generated

The given data set uses vectors to map values and deliver a coherent picture. For the test data a simple text data vector has been taken and results to mapped from that. Due to size of the original dataset, there we use two vectors, mainly the symptoms, and disease. Thereby matching each symptom to the required disease. The probability that the disease is the one specified is then generated, based upon the algorithm used.

Vectors are then fed into a machine learning algorithm. For this application, we have used algorithm such as naïve Bayes. The machine learning algorithm then output a predictive model, based upon the algorithm’s parameter.

Predictive model is used to generate an accurate prediction of the given dataset. The predicted model then compared with the actual model, and the accuracy of the system is measured. Measured accuracy is then determined to be an accurate value of the system.

Our final output is predicted disease, then we must check that it is the same as generated input symptoms. In whole process of predictive model we use the Scikit-learn library which is very powerful library for machine learning.

As we are building a chatbot platform, we try to transmit the data using RabbitMQ protocol service of python. To build a chatbot platform we use Google’s dialogflow API which is very effective and provide easy configuration options.

* 1. **Web Development Framework:**

To build an enterprise level application for multi-specialty hospital or any enterprise application we need a powerful framework to turn everything online and end user can understand. To satisfy the design and performance both we use Django framework which is opensource enterprise application building framework. It provides ease of user access with better look and feel design. Django also provides to integrate the external design which can give better view to end user.

1. **Risk Analysis:**

If we talk about risk factor in the application, it is in health sector, where so many companies facing health sues due to some technical errors or mistakes. This application is deal with health-related data, which is very sensitive, any kind of errors in the results can causes bad impact on human life.

Apart from that data variability is also a problem, sometimes some mismatch in results can make your prediction false.

In future, as deep learning algorithms are getting more efficient, we can use this kind of dataset for deep learning models using tensorflow or pytorch which are most powerful framework. In future chatbots also can get integration with whole elements and algorithms so that it can reduce the different training time of model.

1. **References:**

# Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)

<https://ieeexplore.ieee.org/document/8251305>

# Diagnosis of Alzheimer's Disease Using Machine Learning

<https://ieeexplore.ieee.org/document/8697386>

# Detection of Rheumatoid Arthritis Using Machine Learning

<https://ieeexplore.ieee.org/document/9004234>

* Disease diagnosis system by exploring machine learning algorithms

<https://www.semanticscholar.org/paper/Disease-Diagnosis-System-By-Exploring-Machine-Sunny-Kulshreshtha/035d1d9317e7d4032eb1c6f15c5f0c7925f8e6fd>

1. **Gantt Chart: Milestones and deliverables of the project:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Task Name** | **Q1 2020** |  | **Q2 2020** |  | **Q3 2020** |
| **May 20** | **June 20** | **July 20** | **Aug 20** | **Sept 20** |
| **Planning** |  |  |  |  |  |
| **Research** |  |  |  |  |  |
| **Design** |  |  |  |  |  |
| **Implementation** |  |  |  |  |  |
| **Follow Up** |  |  |  |  |  |