**IFET COLLEGE OF ENGINEERING**

**(AN AUTONOMOUS INSTITUTION)**

**DISEASE PREDICATION FROM VARIOUS SYMPTOMS USING MACHINE LEARNING**

**Submitted by**

1. **VAMSIKRISHNA I (421118104093)**
2. **RAMKUMAR R (421118104070)**
3. **RAJESHWARAN S(421118104069)**
4. **YOGESHWARAN J (421118104102)**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



**STUDENT PROJECTS SCHEME (2021-2022)**

**TAMILNADU STATE COUNCIL FOR SCIENCE AND TECHNOLOGYFORMAT FOR STUDENT PROJECT PROPOSAL**

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | Name of the Student (s) | : | I. VAMSIKRISHNA  R. RAMKUMAR  J. YOGESHWARAN  S. RAJESHWARAN |
|  | one valid e-mail id | : | Ivamsi12@gmail.com |
| 2. | Name of the Guide | : | Mrs. PADMAPRIYA N |
|  | Department / Designation | : | CSE, ASP. |
|  | Institutional Address | : | IFET College of Engineering  (An Autonomous Institution), Villupuram |
|  | Phone No. & Mobile No | : | 9442646743 |
| 3. | Project Title | : | DISEASE PREDICTION FROM VARIOUS SYMPTOMS USING MACHINE LEARNING ALGORITHMS. |
| 4. | Sector in which your Project proposal is to be Considered | : | Engineering & Tech. (Computer Science) |
| 5. | Project Details | : | Annexure-1 |
| 6. | Has a similar project been carried out in your college / elsewhere? If so furnish details of the previous project and highlight the improvements suggested in the present one | : | NO |

**CERTIFICATE**

This is to certify that Mr./Ms. **I.VAMSIKRISHNA, R.RAMKUMAR, S.RAJESHWARAN, J.YOGESHWARAN** is a bona-fide final year student of P.G. Science / U.G. Engineering / P.G. professional courses of our college and it is also certified that two copies of utilization certificate and final report along with seminar paper will be sent to the Council after completion of the project by the end of April 2020.

**Signature of the Guide Signature of the HOD Signature of the Principal/**

**Head of the Institution**

**ANNEXURE-1**

**Disease prediction from various symptoms using machine learning**

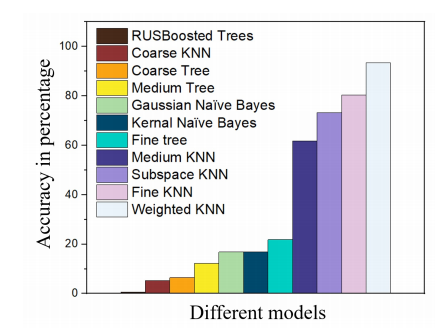
**Introduction:**

Medicine and healthcare are some of the most crucial parts of the economy and human life. There is a tremendous amount of change in the world we are living in now and the world that existed a few weeks back. Everything has turned gruesome and divergent. In this situation, where everything has turned virtual, the doctor and nurses are putting up maximum efforts to save people’s lives even if they have to danger their own. Virtual doctors are board-certified doctors who choose to practice online via video and phone appointments, rather than in-person appointments but this is not possible in the case of emergency. Machines are always considered better than humans as, without any human error, they can perform tasks more efficiently and with a consistent level of accuracy. A disease predictor can be called a virtual doctor, which can predict the disease of any patient without any human error. While ML projects vary in scale and complexity, their general structure is the same. Several rule-based techniques were drawn from machine learning to recall the development and deployment of the predictive model. Several models were initiated by using various machine learning (ML) algorithms that collected raw data and then bifurcated it according to gender, age group, and symptoms. The data-set was then

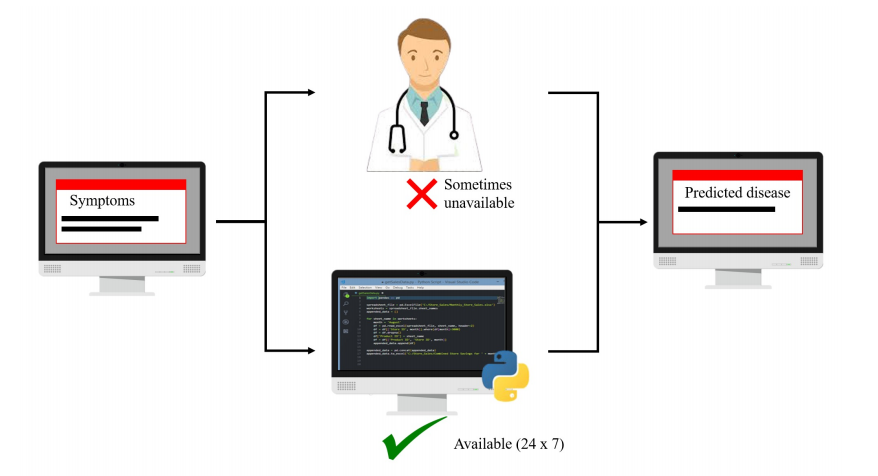
processed in several ML models like Fine, Medium and Coarse Decision trees, Gaussian Naive Bayes, Kernel Naive Bayes, Fine, Medium and Coarse KNN, Weighted KNN, Subspace KNN, and RUS Boosted trees. According to ML models, the accuracy varied. While processing the data, the input parameters data-set was supplied to every model and the output is obtained.

**LITERATURE SURVEY:**

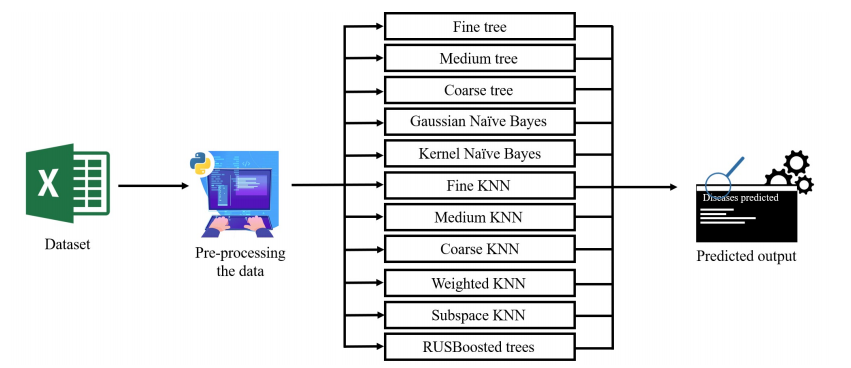
Numerous research works have been carried out for the prediction of the diseases based on the symptoms shown by an individual using machine learning algorithms. They included 3744 unvaccinated adults and adolescent patients of influenza who had fever and at least 2 other symptoms of influenza. Out of 3744, 2470 were confirmed to have influenza by the laboratory.

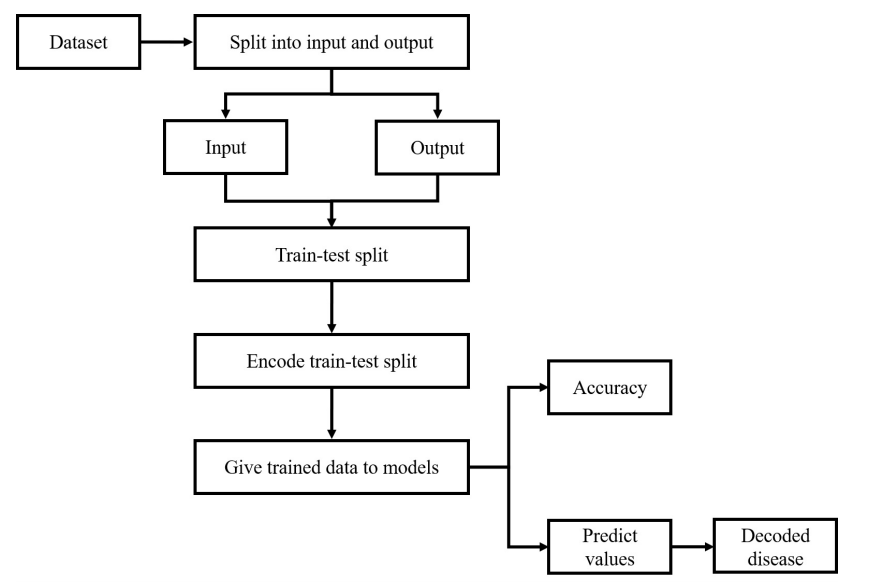
****

**PROPOSED SYSTEM FOR DISEASE PREDICTION:**

****

**PROPOSED SYSTEM DURING TRAINING:**

****

**FUNCTIONING OF MACHINE LEARNING MODELS:**

**PROPOSED METHODOLOGY:**

From an open-source dataset, an excel sheet was created where we listed down all the symptoms for the respective diseases. After which depending on the diseases, age and gender were specified as a part of the dataset. We listed down around 230 diseases with more than 1000 unique symptoms in all. The symptoms, age, and gender of an individual were used as input to various machine learning algorithms.

**K-nearest neighbors (KNN):**

The K-nearest neighbors (KNN) algorithm used is a type of supervised machine learning algorithm. It simply calculated the distance of a new data point to all other training data points. The distance can be of Euclidean or Manhattan type. After this, it selects the K-nearest data points, where K can be any integer. Lastly, it assigns the data point to the class to which the majority of K data points belong.

**Fine, Medium, and Coarse KNN:**

We need to assign the integer values of K to find the distance. So, in our fine KNN model, we assigned a low value of K which means it approximately uses only one neighbor for the prediction. Similarly, the medium KNN model uses approximately 10 neighbors and the coarse KNN uses 100 neighbors. Since the neighbors for each, the model differs the accuracy percentages also varied with a wide range. Among all the three models our fine KNN gave us a very high accuracy whereas the coarse KNN resulted in a low prediction value.

**Weighted KNN:**

It is a modified version of KNN. In KNN we chose an integer parameter K and by using that parameter we found where the major predicted values lied. But if the value of K is too small the algorithm is much more sensitive to the points that are outliers. Also, if the value of K is too large then all the points that are almost very close to the K value are selected. To overcome this issue the weighted KNN gave more weight to the points that were nearest to the K value and the less weight to the points that were farther away. We were able to get the highest accuracy using this model. Also among all the KNN models, this model gave us the best results.

**Naive Bayes:**

It is a machine learning algorithm for classification problems and is based on Bayes’ probability theorem. The primary use of this is to do text classification which involves high dimensional training data sets. We used the Bayes theorem that can be defined as:

P(h | d) = P(d | h).P(h) / P(d)

Where P(h | d) is the probability of hypothesis h given the data d. This is called the posterior probability. P(d | h) is the probability of data d given that the hypothesis h was true. P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h. P(d) is the probability of the data (regardless of the hypothesis).

**Gaussian Naive Bayes:**

It follows the same procedure as the Naive Bayes. But for Naive Bayes we need a categorical dataset and for Gaussian Naive Bayes we need a dataset that has all the continuous features. Our dataset consisted of continuous features of symptoms, age, and gender so it was mandatory to use this model. The accuracy using this model was not a very high value.

**Kernel Naive Bayes:**

Our dataset had some numerical attributes such as age so we implied Kernel Naive Bayes to predict the medicines. The steps followed for this algorithm are similar to the Naive Bayes. The major benefit of using this algorithm is that it provides estimators that have a nonparametric nature. If there is no prior knowledge that the dataset used is parametric or not this model can give more accurate results. The results given by this model were almost the same as shown by the Gaussian Naive Bayes.

**Decision trees:**

Decision trees algorithm belongs to the family of supervised learning algorithms. It is used for regression and classification. In the decision tree, for prediction, it uses the method of tree diagram at the top. It contains a root node after which it gets split in the dominant input feature and then it again gets split. These processes continue till all input is placed and at the node, the extreme last node contains the weights on the bases of these weights it classifies the input. In a coarse tree, the maximum number of splits from each node is 4. Whereas in a Medium tree, the maximum number of splits from each node is 20. In a fine tree, the maximum number of splits from each node is 100.

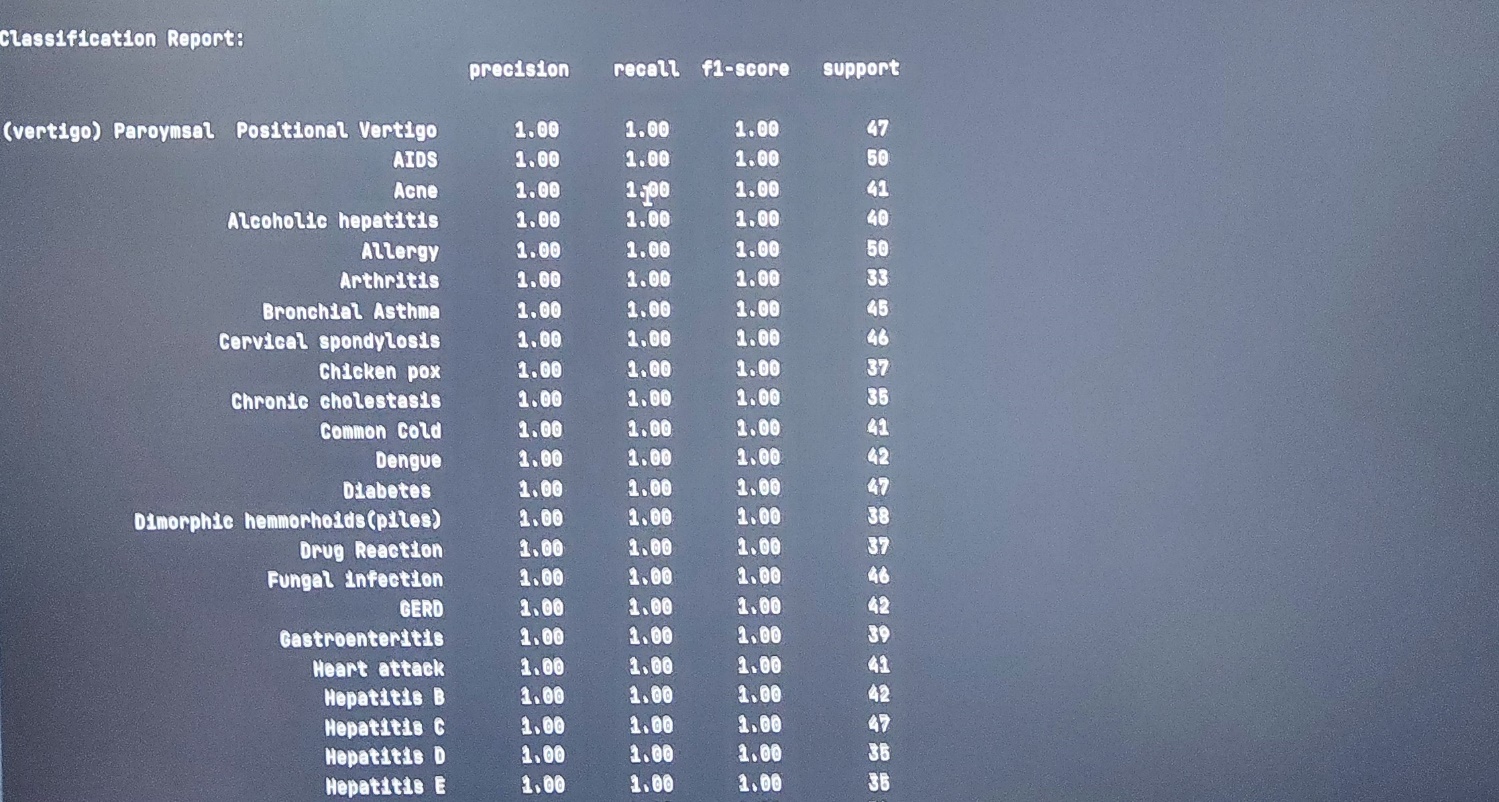
**Sub Space KNN:**

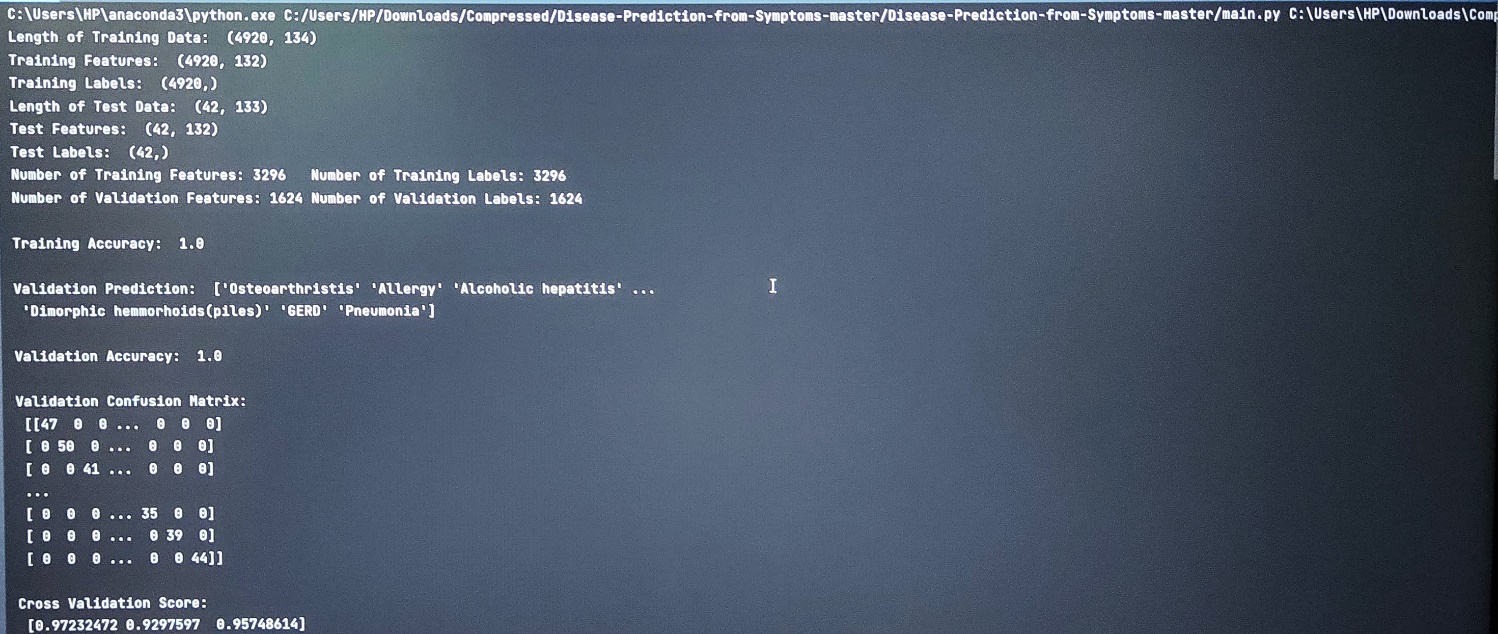
The Sub-Space KNN method is similar to bagging except that the features are randomly sampled, with replacement. Informally, this causes individual learners to not over-focus on features that appear highly predictive/descriptive in the training set, but fail to be as predictive for points outside that set. For this reason, random subspaces are an attractive choice for problems where the number of features is much larger than the number of training points.

**RUS Boost algorithm:**

Our data was required to be trained properly to get proper and good accuracy. RUS Boost algorithm is used for improving the performance of the trained data set acquired from the skewed data set. RUS Boost is a hybrid data sampling/boosting algorithm. This algorithm is one of the methods producing the fastest results amongst the hybrid boosting algorithms.

**Results and Discussion:**





The result is fully based on the dataset of disease prediction under the prediction statistics of the precision, re-cell, fi-score, support value. The confusion matrix can represent the score at the diagonal position to show the validation of the DISEASE PREDICTION in the multi-dimensional array to plot the accuracy as 47, 50, 41,35,39,44. The cross-validation score will be in the single dimension array as [0.97232472 0.9297597 0.95748614].

**Conclusions:**

The manuscript presented the technique of predicting the disease based on the symptoms, age, and gender of an individual patient. The Weighted KNN model gave the highest accuracy of 93.5 % for the prediction of diseases using the above-mentioned factors. Almost all the ML models gave good accuracy values. As some models were dependent on the parameters, they couldn’t predict the disease and the accuracy percentage was quite low. Once the disease is predicted, we could easily manage the medicine resources required for the treatment. This model would help in lowering the cost required in dealing with the disease and would also improve the recovery process.