A machine learning aid to predict diseases based on lifestyle and symptoms

[1]Dr Padmashree T, [2]Dr Harsha S, [3]Haripriya V Joshi
[1] RV College of Engineering,[2] RNS Institute of Technology, [3] RV College of Engineering
[1] padmashreet@rvce.edu.in,[2] harshahassan@gmail.com, [3]joshiharipriya8@gmail.com

Abstract
Currently with this pandemic situation, health is given topmost priority. Lifestyle and diet are the two foremost features that are measured to effect numerous illnesses. Sicknesses are mostly triggered by grouping of alteration, lifestyle choices and environments [4]. Taking precaution, early detection of diseases and awareness about possible health breakdown can create a drastic change which can eventually lead to a good health. This study aims to predict lifestyle diseases an individual is susceptive to and create awareness about healthy life style.
A huge dataset consisting of patient attributes like unhealthy eating habits, obesity, alcoholic, stress and many other parameters are taken as input and would be modelled thereby classifying a lifestyle disorder. Based on the output, symptoms would be given as input and using Naïve Bayes having class label as Disease name and symptoms as parameters, the model would predict current or future occurrence of disorders. Based on symptoms and habits, the proposed system is capable of suggesting home remedies to control the disease from becoming worse until patient meets doctor.
After applying the algorithm and building the model, lifestyle disorder prediction yielded 98% accuracy whereas symptom-based disorder prediction yielded 92.42 % accuracy along with respective home remedy suggestion.

Keyboard: Artificial intelligence, , feature extraction, lifestyle disorder , Naïve Bayes, Support Vector Machine.

I. Introduction

A description prepared by the World Health Organization and World Economic Forum declares that India will experience a cumulative loss of $235.6 billion by 2018 due to gloomy routines and flawed régime [11]. Routine and food are the two chief influences that are measured to effect receptivity to many illnesses.

Illnesses are mostly triggered by the blend of alteration, existence collections, and environments. Adding to it, classifying lifestyle problems based on person's living is one of the utmost vital things a person can do to help the doctor comprehend and identify genetically linked conditions. Illnesses which are related to the individual's lifestyles are recognized as lifestyle diseases.

This learning aims to comprehend Multinomial Naïve Bayes and use it to find probability of suffering from lifestyle diseases that a person might be vulnerable to. The model also predicts diseases based on symptoms if the user is suffering from any, and suggests home remedy till the availability of doctor.

The objective of this work is to build a system which would be effective to interact with the user through chat bot, predict the disease based on symptoms and the lifestyle of the user. The system must be able to suggest possible home remedy for the predicted disease

Structure of Paper

The rest of the paper is as organized as follows,
Section 2 proceeds with literature survey and findings.
The methodology is as presented in Section 3, In Section

4 the implementation details using KNN and Naïve Bayes algorithm is discussed, result outcomes are as shown in section 5. Finally, the conclusion and future work are followed in sections 6 and 7.
The rest of the paper is as organized as follows, Section 2 proceeds with literature survey and findings.
The methodology is as presented in Section 3, In Section 4 the implementation details using KNN and Naïve Bayes algorithm is discussed, result outcomes are as shown in section 5. Finally, the conclusion and future work are followed in sections 6 and 7.
The rest of the paper is structured as follows: section 2 proceeds with literature survey and findings. The methodology is as presented in section 3, In section 4 the implementation details using Multinomial Naïve Bayes algorithm is discussed, result outcomes are as shown in section 5. Finally, the conclusion and future works are followed in section 6.
The rest of the paper is as organized as follows, Section 2 proceeds with literature survey and findings. The methodology is as presented in Section 3, In Section 4 the implementation details using KNN and Naïve Bayes algorithm is discussed, result outcomes are as shown in section 5. Finally, the conclusion and future work are followed in sections 6 and 7.

II. **Literature review**

Illnesses that are related with the method an individual or collection of individuals live are recognized as lifestyle ailments. Health care industry gathers huge illness connected information that is inappropriately collected to learn concealed data that can be used for operative choice creation [13]. In [3] by Mrunmayi Patil, et al, authors have used SVM to forecast lifestyle diseases that a specific person might be vulnerable. A Deoxyribonucleic acid testing that analyses individual lifestyle costs around ten to twenty thousand rupees. Lifestyle disorder prediction using SVM was modeled from IISC, Bangalore resulting in 92.3% accuracy. System based disorder prediction using several algorithms was modelled yielding maximum of 92% accuracy. There is no model which bridges the gap between predicting lifestyle disorder and probable disease an individual is susceptive to.
They have suggested and simulated a financially feasible model as a substitute to deoxyribonucleic acid testing that examines a person's routine to recognize likely coercions that form the basis of diagnostic tests and illness controlling, which might be raised due to unnatural regimes and dangerous liveliness consumption, bodily latency. The replicated system will demonstrate to be an intelligent substitute to sense likely syndromes caused by unnatural lifestyles. SVM Algorithm is used obtaining 92.30% accuracy.
Health care practices comprises of assembling all types of persistent information which would assist the doctor properly to identify the well-being state of the patient. The risk for the stated lifestyle diseases increases with "body mass index" (BMI) in direct proportion. BMI is a number calculated by "Weight/Height2" and is used to assess body composition [7]. This information could be modest indications detected by the patient, early analysis by a doctor or a thorough examination outcome from a research laboratory. Therefore, this information is only applied for examination by a surgeon who then determines the illness using his/her individual therapeutic practice. This intellect has been utilized with Naive Bayes and Random Forest Classification algorithm to a set of rules to categories numerous illnesses using appropriate datasets to test whether the affected person is stricken by that disorder or not. A presentation examination of the illness information for both algorithms is considered and associated.
In [2] by V. Jackins ae. Al, the consequences of the model predicts the efficiency of the classification methods on a dataset, as well as the feature and complications of the dataset utilized. Data mining can be successfully applied in therapeutic field [4]. The goal of this

research is to learn a prototypical model for the analysis of diabetes, coronary heart disease and cancer among the available dataset [5-10]. The data is selected from different sources and pre-processed. Later, classifiers like Bayesian and random forest models are built. Lastly, the correctness of the prototype is intended and are constructed on the competence intentions. Bayesian Classification community validates the correctness of 74.46, 82.35 and 63.74% for diabetes, coronary heart disorder and most cancers data [12]. Likewise, class with Random Forest classifier suggests the correctness of 74.03, 83.85 and 92.40. The accuracy of final results of Random Forest version for the 3 illnesses is more than the accuracy values of Naïve Bayes classifier.

In [1] authors have mentioned that health and medicine are gaining a lot of importance in today's advancing world, where evolving technology is being used to combat almost all the known diseases. This project proposes a system which takes as input the symptoms of the patient to predict the disease, which is followed by recommending the appropriate medicine. This system consists of a database system module, data preparation module, disease prediction module, medicine recommendation module, model evaluation and data visualization module.

A Decision tree map, Naive Bayes model and Random Forest algorithm are used to achieve the objectives [1]. This paper deals with the implementation of a system which performs the dual function of prediction of diseases and the recommendation of medicines, which adds to the capabilities of the existing systems. After model building using Naive Bayes, Random Forest, the accuracies obtained is 84%, 86% respectively which are quite low. So, in an attempt to improve the accuracy, the project uses Decision Tree as the primary algorithm with 92% accuracy.

III. Methodology

In this section the algorithms that are part of implementation process are being explained., the proposed architecture is as shown in figure 1, defines the high-level design of the system.

The various steps involved in the proposed methodology are as follows:

Step 1: Lifestyle parameters are taken as input and likelihood of lifestyle disorder is predicted.

Step 2: Symptoms are taken as input and disease is predicted by the model.

Step 3: The previously obtained output i.e., disease name is given as input. Home remedy is suggested by model for the predicted disease.

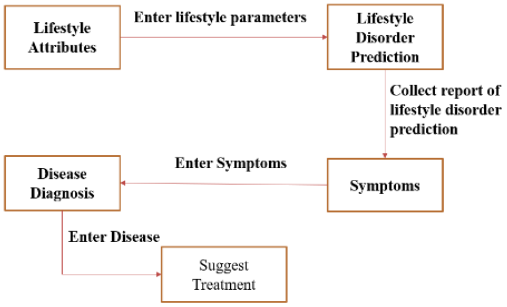

Fig 1: System Architecture

A. Algorithm

Naïve Bayes Algorithm

The Naive Bayes classifier uses the Bayes theorem of probability for prediction of unknown class. It assumes the effect of a particular feature in a class is independent of other features,

finds its application especially on large data sets and supports sophisticated classification methods as illustrated in equation 1.

$$P(q\,|t\,) = \frac{P(t\,|p)P(q)}{P(t)} \quad \text{...eq(1)}$$

where in,

P(q | T) = $P(t_1\,|c) * P(t_2\,|c) * \ldots * P(t_n\,|c) * P(q)$

P(q|t) is the posterior probability of class (target) given predictor (attribute).

P(q) is the prior probability of class.

P (t|q) is the likelihood which is the probability of predictor given class.

P(t) is the prior probability of predictor.

Algorithm Naïve Bayes:
Begin
[S1]: compute the marginal probability for given class labels
[S2]: define the possibility of expected outcomes with each attribute for each class
[S3]: compute the conditional probability for each class using equation 1.
[S4]: the output of prediction is the class with the highest posterior probability.
End

Gaussian naïve Bayes: A bell-shaped curve is generated as shown in figure 2, that refers to the symmetric value of the mean features based on Gaussian distribution.

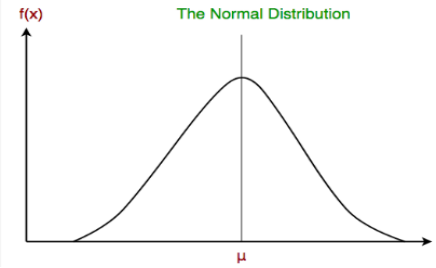

Fig 2: bell curve generated

Normal Distribution Since the likelihood of the features is gaussian, the conditional probability is calculated by using the equation 4 as below.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad \text{.....eq(2)}$$

he Naive Bayes classifier uses the Bayes theorem of probability for prediction of unknown class. It assumes the effect of a particular feature in a class is independent of other features, finds its application especially on large data sets and supports sophisticated classification methods as illustrated in equation 3.

Where P(k): Prior probability of k.
P(D): Probability of the data, (irrespective of the hypothesis).
P(k|D): Posterior probability i.e. the probability of hypothesis m given the data D.
P(D|k): Posterior probability of data D given that the hypothesis m was true.

IV. Experimental results

The experimental results are shown below.

Table 1: Unit Testing of symptomatic based disease prediction model

Test case:	1
Name of test:	Test model on validation set

Sample Input:	Congestion Chest pain Runny nose
Expected Output:	Hypertension
Actual Output:	Hypertension
Remarks:	PASS

Table 1 shows the Unit Testing of symptomatic based disease prediction model where the user will enter feature values and the model will output the target variable value.

Table 2 Unit Testing of lifestyle disorder prediction model

Test case:	2
Name of test:	Test model on validation set
Sample Input:	Eating: often Physical: active Obesity: yes Sleep: Sometimes Stress: Fair Smoking: No Alcohol: no Hereditary: no Gender: male Age: Senior citizen
Expected Output:	Lifestyle disorder chances: No
Actual Output:	Lifestyle disorder chances: No
Remarks:	PASS

Table 7.2 shows the Unit Testing of lifestyle disease prediction model where the user will enter feature values and the model will output the target variable values.

Table 3: Unit Testing of treatment suggestion model

Test case:	3
Name of test:	Test model on validation set
Sample Input:	Allergy
Expected Output:	Print drugs, creams and ayurvedic home remedies
Actual Output:	Print drugs, creams and ayurvedic home remedies
Remarks:	PASS

Table 3 shows the Unit Testing of treatment suggestion model where the user will enter feature values and the model will output the target variable value.

Table 4: Integration Testing for Disease prediction through UI

Test case:	4
Name of test:	through UI
Sample Input:	Symptoms, lifestyle parameters, predicted disease.
Expected Output:	Successfully display the likelihood of lifestyle disorder, symptomatic disorder and suggest home remedy.
Actual Output:	Successfully display the likelihood of lifestyle disorder, symptomatic disorder and suggest home remedy.
Remarks:	PASS

Table 4 shows the Integration Testing for the application where user will enter the input values and the relevant output is displayed.

The figure 3 shows snapshot of the UI where symptoms are given the form of input and the disease is predicted by the model.

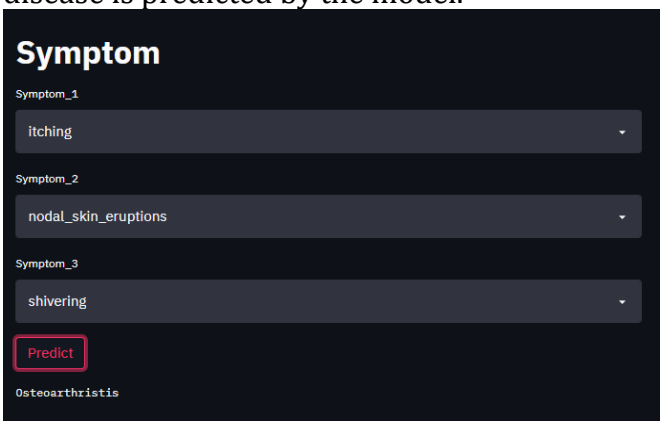

Fig 3: Symptom based disease prediction

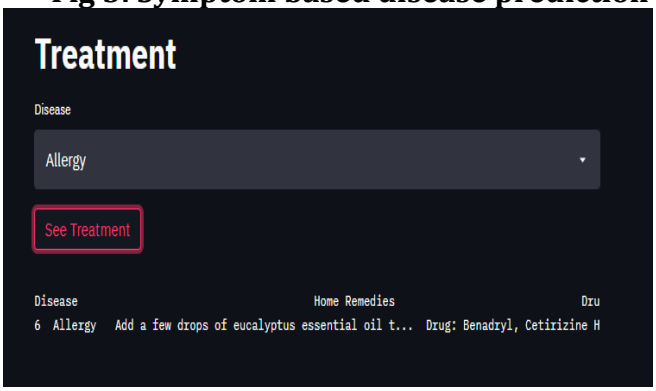

Fig 4: Home remedy suggestion.

The figure 4 shows snapshot of the UI where disease predicted by the model is given the form of input and the home remedy suggested by the model.

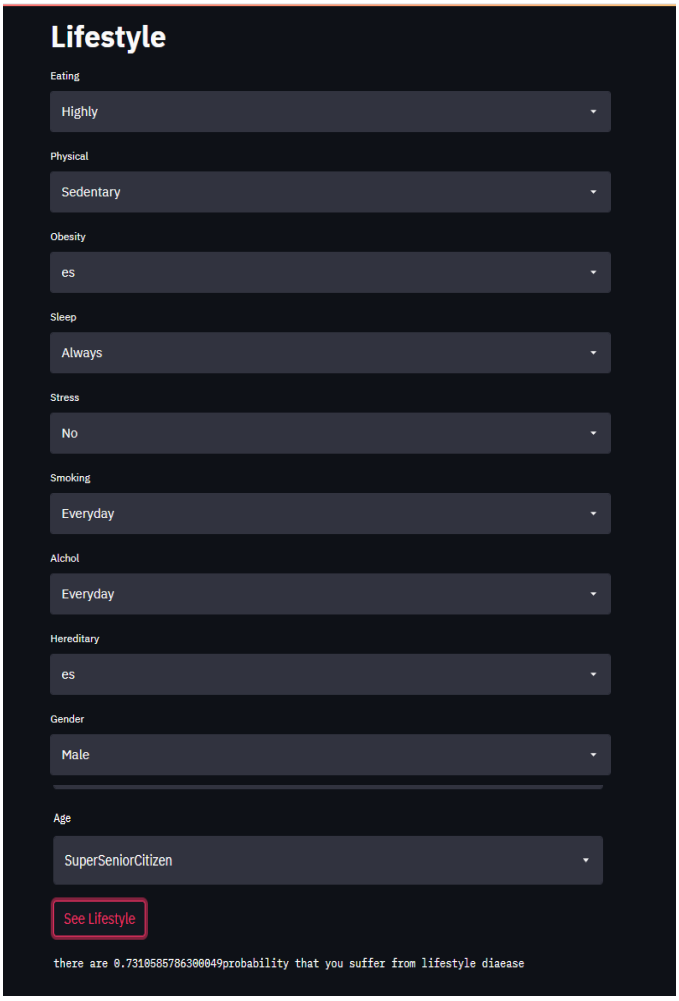

Fig 5: Lifestyle disease prediction

The figure 5 shows snapshot of the UI where lifestyle attributes are given the form of input and the lifestyle disorder is predicted by the model.

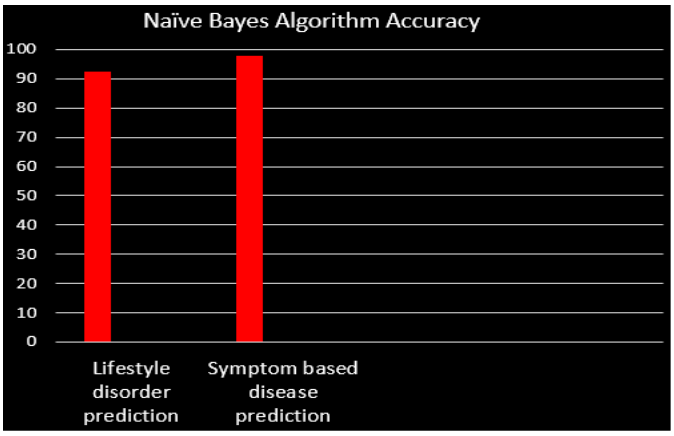

Fig 6: Evaluation of Naïve Bayes Algorithm

The model is tested for all the datasets and the accuracies are **92.42 of symptomatic disease prediction and 98% for Lifestyle disorder prediction applying Naïve Bayes algorithm described in figure 6.**

V. CONCLUSION AND FUTURE WORK.

Based on inputs, the proposed system is capable of predicting the likelihood of lifestyle disorder and symptomatic disorder. The proposed system can suggest treatment based on symptoms and disease identified.

The data repository makes itself a unique as there is no such repository which gives all the information about symptoms, different frequencies of lifestyle disorder parameters for the diseases and home remedy possible for disease the model works on.

The built model is useful by doctors to predict occurrence of lifestyle disorder and the symptomatic disorder the patient is suffering from. It also suggests home remedy to cure the symptomatic disease with least side effects.

In future, this model can be applied to all the diseases that are based on age, lifestyle, body type and genetic based. Deep learning techniques can be applied to make it more reliable in real time.

References

[1] Dr.T.Venkat Narayana Rao, Anjum Unnisa, Kotha Sreni Ruchika Rachakonda. "Symptom based disease prediction and medicine recommendation system", High Technology Letters, Volume 26, Issue 7, 2020.

[2]V. Jackins, S. Vimal, M. Kaliappan , Mi Young Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naïve Bayes", The Journal of Supercomputing ,2020.

[3] Mrunmayi Patil , Vivian Brian Lobo , Pranav Puranik ,Aditi Pawaskar , Adarsh Pai , Rupesh Mishra "A Proposed Model for Lifestyle Disease Prediction Using Support Vector", IEEE – 43488,2018.

[4] Hossain R, Mahmud S.H, Hossin M.A, Noori S.R.H. and Jahan, H. PRMT: "Predicting Risk Factor of Obesity among MiddleAged People Using Data Mining Techniques". Procedia Computer Science, 132, pp. 1068–1076, 2018.

[5] Sayali Ambekar and Dr.Rashmi Phalnikar, "Disease prediction by using machine learning", International Journal of Computer Engineering and Applications, vol. 12, 2018.

[6] Hossain R, Mahmud S.H , Hossin M.A, Noori S.R.H. and Jahan, H, PRMT: "Predicting Risk Factor of Obesity among Middle Aged People Using Data Mining Techniques". Procedia Computer Science, 132, 2018.

[7] Csige, I. Ujvárosy, D. Szabó, Z. L˝orincz, I. Paragh, G. Harangi, M. Somodi, S. "The impact of obesity on the cardiovascular system". J. Diabetes Res. 2018,

[8] Mishra A.K, Keserwani P.K., Samaddar S.G, Lamichaney, H.B. and Mishra, A.K, "A decision support system in healthcare prediction. In Advanced Computational and Communication Paradigms "(pp. 156–167) Springer, Singapore. 2018.

[9] Kazeminejad A, Golbabaei S. and Soltanian-Zadeh H, "Graph theoretical metrics and machine learning for diagnosis of Parkinson's disease using rs-fMRI. "In Artificial Intelligence and Signal Processing Conference (AISP), (pp. 134–139). IEEE 2017.

[10] Kanchan B.D. and Kishor M.M, “Study of machine learning algorithms for special disease prediction using principal of component analysis”. In Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016 International Conference on (pp. 5–10). IEEE. 2016.

[11] Sharma, M. and Majumdar, P.K., “Occupational lifestyle diseases: An emerging issue”. Indian Journal of Occupational and Environmental medicine, 13(3), pp. 109–112. 2009.

[12] Milgram, J., Cheriet, M. and Sabourin, R., 2006. “One against one” or “one against all”: Which one is better for handwriting recognition with SVMs? Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule (France), Suvisoft, 2006.

[13] Suzuki A, Lindor K, St Saver J, Lymp J, Mendes, F Muto, A., Okada, T. and Angulo,” Effect of changes on body weight and lifestyle in nonalcoholic fatty liver disease”. Journal of Hepatology, 43(6), pp. 1060–1066, 2005.