ऋते ज्ञानान्न मुक्तिः

GLA UNIVERSITY
MATHURA
Recognised by UGC Under Section 2(f)

DEPARTMENT OF COMPUTER ENGINEERING & APPLICATIONS
INSTITUTE OF ENGINEERING & TECHNOLOGY

B.Tech. IV Year CSED

Project Report

On

“Disease-Predictor”

Under the supervision of

Dr. Mayank Srivastava

Submitted by

1. **Abhinav Bhardwaj (02 / 181500009)**
2. **Aditya Singh Chauhan (08 / 181500046)**
3. **Ankit Parmar (12 / 181500099)**
4. **An**

vit Gupta (16 / 181500127)

Group No: G-07

Odd Semester, 2021-22

Table of Contents

Chapter 1

Introduction

1.1 Motivation and Overview

Presently, the medical practitioners prognosticate the diseases of a patient based on their knowledge and the experience and understanding they have developed over time. Even after a great experience with the medical field, the prognosis made by doctors could be inaccurate as they are also human. Our approach will give a single platform for predicting various illnesses. Our system will examine the reports and draw conclusions based on them. The model's new predictions will be utilized as a prior dataset to train the model for future inputs to improve the system's performance. Our technology will act as a resource for doctors, allowing them to be more confident in their judgments and reducing the likelihood of incorrect diagnoses.

People today suffer from a range of diseases as a result of their lifestyles and the environment in which they live. As a result, being able to detect sickness at an early stage is essential. A doctor's ability to establish an accurate diagnosis solely on symptoms, on the other hand, is limited.
Making accurate disease forecasts is the most challenging problem. Data mining is essential for sickness prediction in order to overcome this problem. The amount of data in medical science expands exponentially each year. As the volume of data in the medical and healthcare professions has expanded, so has the need for precise medical data analysis, which has been aided by early patient care.

The main motive of the project is to provide a web application to the medical practitioners that will help in diagnosing the patient. Our web application will ease the work of doctors in diagnostics of the disease of a patient. Our platform will be using training datasets for creating such a model and the predicted result will also be added as a training data to enhance system's performance over time.

1.2 Objective

- Our system will provide a single platform which can be used to predict different diseases.
- Our system will analyze the reports and will provide the conclusion based on that.
- The new results predicted by the model will also be used as previous dataset to train the model for future input for enhancing the system's performance.
- The new results predicted by the model will also be used as previous dataset to train the model for future input for enhancing the system's performance.
- Our system will serve to doctors as a source which will help doctors to trust their decisions and also reduce the chances of faulty diagnostics

1.3 Scope

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. The main objective of this research is to develop a prototype Disease Prediction System (DPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bayes and KNN. So, it provides effective treatments, it also helps to reduce treatment costs and also enhances visualization and ease of interpretation. With immense knowledge and accurate data in that field. Large corporations invest heavily in this kind of activity to help focus attention on possible events and risks that are involved. Such work brings together all available past and current data, as a basis on which to develop reasonable expectations about the future.

Chapter 2

Literature Survey

A. Common Diseases

Dahiwade[9] et al. suggested a machine learning-based approach for diagnosing and predicting common illnesses. The symptoms dataset came from the UCI ML repository, and it included symptoms for a wide range of disorders. To predict numerous illnesses, the system employed CNN and KNN as classification approaches. Furthermore, the proposed solution was complemented with additional information on the tested patient's living habits, which proved to be useful in determining the amount of risk associated with the anticipated disease. In terms of processing time and accuracy, Dahiwade[9] et al. compared the outcomes of the KNN and CNN algorithms. CNN had an accuracy of 84.5 percent and a processing time of 11.1 seconds, respectively. Statistics show that the KNN algorithm performs poorly when compared to the CNN method. In light of this research, Chen[22] et al. [10] found that CNN outperformed traditional supervised algorithms including KNN, NB, and DT. The authors determined that the suggested model was more accurate, which they attribute to the model's capacity to recognise complicated nonlinear interactions in the feature space. Furthermore, CNN finds high-importance characteristics that result in a more precise description of the condition, allowing it to effectively anticipate diseases of great complexity. This conclusion is well-supported, with actual evidence and statistical reasons to back it up. However, the models offered lacked information, such as Neural Networks characteristics such as network size, architectural type, learning rate, and back propagation technique, among others. Furthermore, the performance analysis is only evaluated in terms of accuracy, which calls into question the veracity of the reported conclusions. Furthermore, the authors failed to account for the bias problem that the tested algorithms encounter. Incorporating additional feature variables, for example, might dramatically improve the performance metrics of underperforming algorithms.

B. Kidney Diseases

Using The Kidney Function Test (KFT) dataset, Serek et al. planned a comparison investigation of classifier performance for chronic kidney disease (CKD) identification. The classifiers employed in this study are KNN, NB, and RF, and their performance is measured in terms of F-measure, precision, and accuracy. According to the findings, RF outperformed NB in terms of F-measure and accuracy, but NB outperformed RF in terms of precision. Vijaya ani's goal for this project was to use SVM and NB to detect renal disorders. Acute Nephritic Syndrome, Acute Renal Failure, Chronic Glomerulonephritis, and Chronic Kidney Disease were identified using the classifiers. In addition, the study looked at which classification algorithm performed best in terms of accuracy and execution time.. According to the results, SVM outperformed NB in terms of accuracy, making it the best performing algorithm. NB, on the other hand, categorised data in a short amount of time. Several more empirical research focused on identifying CKD; Charlenae et al. and Kotuku et al. determined that the SVM classifier is the best for renal illnesses since it handles semi-structured and unstructured data effectively. As a consequence of this versatility, SVM was able to handle bigger feature spaces, resulting in great accuracy when diagnosing complicated kidney illnesses. Although the data corroborate the conclusion, the previous claim that alternative hyper-parameters were not tested when evaluating the performance of ML algorithms weakens the conclusion. According to Uddin[3], exploring the hyper-parameter space might result in varied accuracy outcomes and higher ML algorithm performance.

C. Heart Diseases

Marimuthu et al. used supervised machine learning techniques to forecast cardiac disease. Gender, age, chest pain, gender, goal, and slope were used to organise the data characteristics [16]. The utilised machine learning algorithms were DT, KNN, LR, and NB. According to the research, the LR algorithm had the highest accuracy of 86.89 percent, making it the most successful among the other algorithms. Dwivedi sought to improve the precision of heart disease prediction in 2018 by taking into consideration new characteristics such as resting blood pressure, serum cholesterol in mg/dl, and maximum heart rate reached. The employed dataset came from the UCI ML lab, and it included 120 heart disease positive samples and 150 heart disease negative samples. Artificial Neural Networks (ANN), SVM, KNN, NB, LR, and Classification Tree Dwivedi assessed them all. The findings of Polaraju and Vahid et al., who found that Logistic Regression outperformed other approaches including ANN, SVM, and Adaboost, support this conclusion. The studies excelled in conducting in-depth analyses of machine learning models. For example, for each ML algorithm, several hyper-parameters were evaluated to converge to the greatest feasible accuracy and precision values. Despite this benefit, the learning models are limited in their ability to target illnesses with greater accuracy and precision due to the modest size of the imported datasets.

D. Breast Diseases

Shubair attempted to diagnose breast cancer using ML algorithms, particularly RF, Bayesian Networks, and SVM, and Yao[21] came to the conclusion that the RF method performed better than SVM. The researchers used the UCI Repository to access the Wisconsin original breast cancer dataset and used it to compare the learning models in terms of important criteria including accuracy, recall, precision, and ROC graph area. The classifiers were put to the test using the K-fold validation technique, with K set to 10. SVM outperformed other methods in terms of recall, accuracy, and precision, according to the simulation findings. However, the ROC graph indicated that RF had a better likelihood of correctly classifying the tumour. Yao[21], on the other hand, tested a variety of data mining approaches, including RF and SVM, to find the best algorithm for breast cancer prediction. As a result of the findings, the classification rate, sensitivity, and Random Forest algorithm's accuracy, sensitivity, and specificity were 96.27 percent, 96.78 percent, and 94.57 percent, respectively, whereas SVM's accuracy, sensitivity, and specificity were 95.85 percent, 95.95 percent, and 95.53 percent, respectively. The former gives more accurate estimations of the amount of data obtained in each feature characteristic. Furthermore, RF is the best method for classifying breast illnesses since it scales effectively for big datasets and reduces the risk of variation and data overfitting. Multiple performance indicators were offered in the research, which helped to solidify the main point. However, including a preprocessing stage to prepare raw data for training has been shown to be detrimental to ML models. According to Yao[21], removing sections of data affects image quality, and hence the ML algorithm's effectiveness is hampered. Parkinson's disease (PARKINSON'S DISEASE) Chen[22] et al. proposed a successful Parkinson's disease (PD) diagnostic method based on Fuzzy k-Nearest Neighbor (FKNN).. The goal of the study was to compare the suggested SVM-based and FKNN-based techniques. For the creation of an optimum FKNN model, the Principal Component Analysis (PCA) was used to combine the most discriminating features. The dataset was collected from the UCI repository and included a variety of biological voice measurements from 31 patients, 24 of whom had Parkinson's disease. The results of the experiments show that the FKNN strategy outperforms the SVM methodology in terms of sensitivity, accuracy, and specificity. Behroozi[23]'s goal for this study was to develop a novel classification framework for diagnosing Parkinson's disease, which was strengthened by a filter-based feature selection technique that improved classification accuracy by 15%. The framework's categorization was determined by using separate classifiers for each subset of the data. dataset to account for the loss of important data KNN, SVM, Discriminant Analysis, and NB were chosen as classifiers. SVM came out on top in every performance indicator, according to the findings. Eskidere[24] also compared the performance of SVM with other classifiers such as Least Square Support Vector (LS-SVM), General Regression Neural Network (GRNN), and Multi-layer Perceptron Neural Network to track the course of PD (MLPNN). The results showed that LS-SVM is the most effective model. The proper comparison of decoders with their ideal performance metric supports this result. ML algorithms are meant to maximize a variety of performance measures, according to Leveson (e.g., Neural Networks optimizes squared error whereas KNN and SVM optimize accuracy). Furthermore, the writers are very skilled in proposing detailed frameworks. SVM parameters such as the kernel and regularization value, for example, were thoroughly discussed. ML models, on the other hand,

E. Parkinson’s Disease

Chen[22] et al. [22] reported a successful Parkinson's disease (PD) diagnostic method based on Fuzzy k-Nearest Neighbor (FKNN). The goal of the study was to compare the suggested SVM-based and FKNN-based techniques. For the creation of an optimum FKNN model, the Principal Component Analysis (PCA) was used to combine the most discriminating features. The dataset was collected from the UCI repository and included a variety of biological voice measurements from 31 patients, 24 of whom had Parkinson's disease. The experimental findings have indicated that the FKNN approach advantageously achieves over the SVM methodology in terms of sensitivity, accuracy, and specificity. In line of this study, Behroozi[23] [23] aimed to propose a new classification frame work to diagnose PD, which was enhanced by a filter-based feature selection algorithm that increased the classification accuracy up to 15 percent . The classification of the framework was characterized by applying independent classifiers for each subset To account for the loss of useful information, use the dataset. KNN, SVM, Discriminant Analysis, and NB were chosen as classifiers. SVM came out on top in every performance indicator, according to the findings. In addition, Eskidere[24] [24] compared the effectiveness of SVM with other classifiers such as Least Square Support Vector (LS-SVM), General Regression Neural Network (GRNN), and Multi-layer Perceptron Neural Network to track the course of PD (MLPNN). The results showed that LS-SVM is the most effective model. The proper comparison of decoders with their ideal performance metric

[25] supports this result. ML algorithms are meant to maximize a variety of performance parameters, according to Lavesson[25] [25]. (e.g., Neural Networks optimizes squared error whereas KNN and SVM optimize accuracy). Furthermore, the writers are very skilled in proposing detailed frameworks. Parameters of SVMs, such as the kernel and the training set, are examples The importance of regularization value was discussed in detail. However, before examining the performances, the ML models were not calibrated. [26] calibration, according to Caruana[26], significantly improves the categorization of a few learning models, particularly NB, SVM, and RF.

CONCLUSION

The application of various machine learning algorithms allowed for the early diagnosis of a variety of ailments, including heart, kidney, breast, and brain disorders. SVM, RF, and LR algorithms were the most extensively used prediction algorithms in the literature, with accuracy being the most widely used performance indicator. When it came to forecasting common illnesses, the CNN model proved to be the most accurate. Furthermore, because of its consistency in handling high-dimensional, semi-structured, and unstructured data, the SVM model exhibited improved accuracy in most cases for renal illnesses and PD. Because of its capacity to scale effectively for big datasets and its propensity to prevent overfitting, RF exhibited advantage in the likelihood of correct illness classification for breast cancer prediction. Finally, when it came to forecasting cardiac illnesses, the LR algorithm proved to be the most accurate. To improve illness efficiency in the future, more advanced machine learning algorithms will be required. prediction. Furthermore, learning models should be adjusted more often after the training period to improve performance. Furthermore, to minimize overfitting and improve the accuracy of deployed models, datasets should be enlarged on diverse demo images. Finally, to improve the performance of learning models, more relevant feature selection approaches should be applied.

REFERENCES

[1] A. Gavhane[1], G. Kokkula[1], I. Pandya, and K. Devadkar[1], “Prediction of heart disease using machine learning,” in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.

[2] Y. Hasija[2], N. Garg, and S. Sourav, “Automated detection of dermatological disorders through image-processing and machine learning,” in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 1047–1051.

[3] S. Uddin[3], A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019.

[4] R. Katarya[4] and P. Srinivas, “Predicting heart disease at early stages using machine learning: A survey,” in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302–305.

[5] P. S. Kohli[5] and S. Arora, “Application of machine learning in disease prediction,” in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.

[6] M. Patil[6], V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra, “A proposed model for lifestyle disease prediction using support vector machine,” in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1–6.

[7] F. Q. Yuan[7], “Critical issues of applying machine learning to condition monitoring for failure diagnosis,” in 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2016, pp. 1903–1907.

[8] S. Ismaeel[8], A. Miri, and D. Chourishi, “Using the extreme learning machine (elm) technique for heart disease diagnosis,” in 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), 2015, pp. 1–3.

[9] D. Dahiwade[9], G. Patle, and E. Meshram, “Designing disease prediction model using machine learning approach,” Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, no. Iccmc, pp. 1211–1215, 2019.

[10] S. Jadhav[10], R. Kasar, N. Lade, M. Patil[6], and S. Kolte, “Disease Prediction by Machine Learning from Healthcare Communities,” International Journal of Scientific Research in Science and Technology, pp. 29–35, 2019.

[11] R. Saravanan[11]and P. Sujatha, “A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification,” in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 945– 949.

[12] Y. Amirgaliyev[12], S. Shamiluulu, and A. Serek, “Analysis of chronic kidney disease dataset by applying machine learning methods,” in 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1–4. [13] V. S and D. S, “Data Mining Classification Algorithms for Kidney Disease Prediction,” International Journal on Cybernetics & Informatics, vol. 4, no. 4, pp. 13–25, 2015.

[14] A. Charleonnan[13], T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, “Predictive analytics for chronic kidney disease using machine learning techniques,” 2016 Management and Innovation Technology International Conference, MITiCON 2016, pp. MIT80–MIT83, 2017.

[15] P. Kotturu[15], V. V. Sasank, G. Supriya, C. S. Manoj, and M. V. Maheshwarredy, “Prediction of chronic kidney disease using machine learning techniques,” International Journal of Advanced Science and Technology, vol. 28, no. 16, pp. 1436–1443, 2019.

[16] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, “A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach,” International Journal of Computer Applications, vol. 181, no. 18, pp. 20–25, 2018.

[17] A. K. Dwivedi, “Performance evaluation of different machine learning techniques for prediction of heart disease,” Neural Computing and Applications, vol. 29, no. 10, pp. 685–693, 2018.

[18] K. Polaraju, D. Durga Prasad, and M. Tech Scholar, “Prediction of Heart Disease using Multiple Linear Regression Model,” International Journal of Engineering Development and Research, vol. 5, no. 4, pp. 2321–9939, 2017. [Online]. Available: www.ijedr.org

[19] S. Pouriyeh[19], S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, “A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease,” in 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 204– 207.

[20] P. P. Sengar[20], M. J. Gaikwad, and A. S. Nagdive, “Comparative study of machine learning algorithms for breast cancer prediction,” Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, pp. 796–801, 2020.

[21] D. Yao[21], J. Yang, and X. Zhan, “A novel method for disease prediction: Hybrid of random forest and multivariate adaptive regression splines,” Journal of Computers (Finland), vol. 8, no. 1, pp. 170–177, 2013.

[22] H. L. Chen[22], C. C. Huang, X. G. Yu, X. Xu, X. Sun, G. Wang, and S. J. Wang, “An efficient diagnosis system for detection of Parkinson’s disease using fuzzy k-nearest neighbor

approach," Expert Systems with Applications, vol. 40, no. 1, pp. 263–271, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2012.07.014

[23] M. Behroozi[23] and A. Sami, "A multiple-classifier framework for Parkinson's disease detection based on various vocal tests," International Journal of Telemedicine and Applications, vol. 2016, 2016.

[24] O. Eskidere[24], F. Ertas¸, and C. Hanilc¸i, "A comparison of regression ¨ methods for remote tracking of Parkinson's disease progression," Expert Systems with Applications, vol. 39, no. 5, pp. 5523–5528, 2012.

[25] N. Lavesson[25], Evaluation and Analysis of Supervised Learning Algorithms and Classifiers, 2006.

[26] R. Caruana[26] and A. Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics," Proceedings of the 23rd international conference on Machine Learning, pp. 161–168, 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.3232

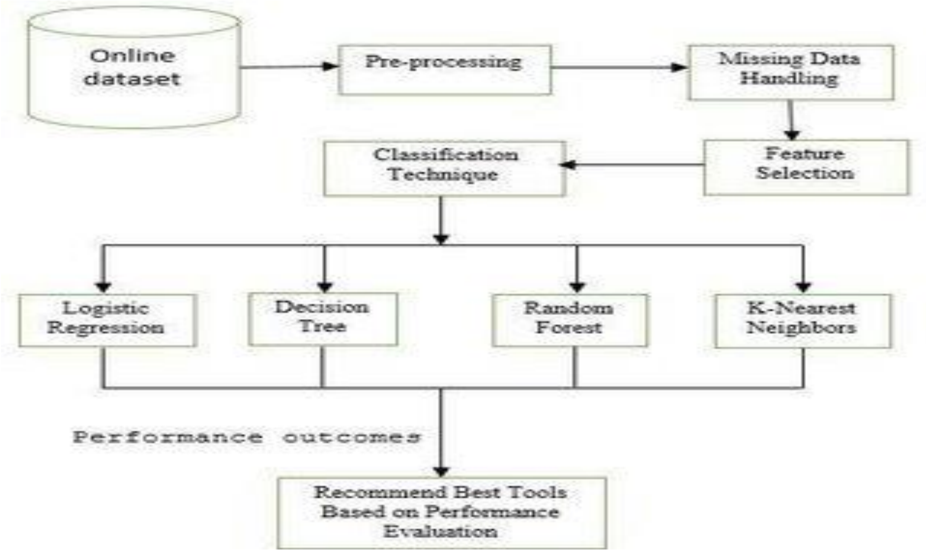
Chapter 3
Proposed Model

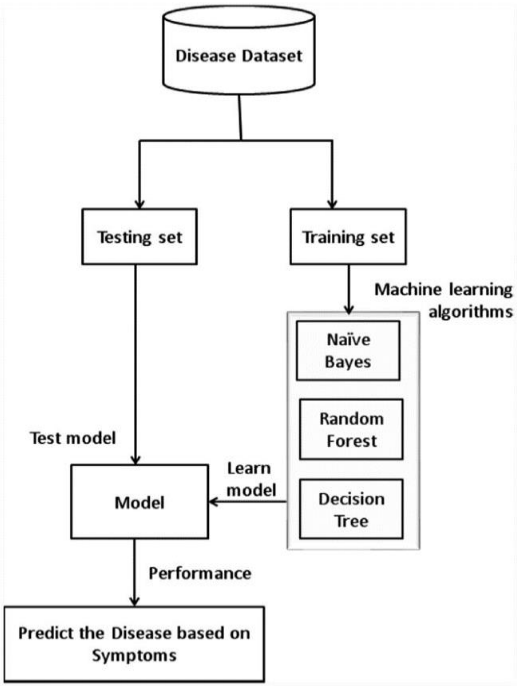
A. Data Source

The dataset used here for predicting disease is taken from Kaggle. Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Gold bloom and Jeremy Howard. Nicholas Gruen was founding chair succeeded by Max Lev chin. Equity was raised in 2011 valuing the company at $25 million. On 8 March 2017, Google announced that they were acquiring Kaggle.

B. Architecture Diagram





C. Description of Algorithms

In this section describe two main algorithms are used in this system namely i) Decision tree Classification algorithm and ii) Naïve Bayes Classifier Algorithm.

A) Decision Tree Classification Algorithm

The decision tree is a supervised machine learning algorithm. It handles both the categorical data and numerical data. Based on certain conditions it gives a categorical solution such Yes/No, True or false, 1 or 0. For handling medical dataset the Decision tree Classification algorithm is widely used. The result of this model differing from the other models like the k-nn model, SVM model. The output consists of horizontal and vertical line splits based on the condition depends on the dependent variables. The accuracy level of this algorithm is quite higher than the other algorithms. The reason for the higher accuracy of this algorithm is these model analyses the dataset in the tree shape format. Thus, each and every attribute of the dataset is been analyzed. Thus, the accuracy rate of this model is higher. This model analyzes the data in the tree-shaped structure. Tree shaped diagram determines the course of actions. The decision tree model analyzes the data on the basis of three nodes namely

- **Root node** - this main node, on basis of this node all other perform it function

- **Interior node** - the condition of dependent variables is handled by this node

- **Leaf node -** the final result is carried on a leaf node.

Formula for finding root node (Information Gain)

Information Gain = Class Entropy - Entropy Attributes

To find Class Entropy**:**

(Pi + Ni) =−PP+Nlog2PP+Nlog2NP+N.

Here => P, Possibilities of Yes.
=> N, Possibilities of No.

To find Entropy Attributes:

Entropy attribute=∑Pi+NiP+N.

B) Naïve Bayes Classification Algorithm

Naïve Bayes classifier is a supervised algorithm which classifies the dataset on the basis of Bayes theorem. The Bayes theorem is a rule or the mathematical concept that is used to get the probability is called Bayes theorem. Bayes theorem requires some independent assumption and it requires independent variables which is the fundamental assumption of Bayes theorem.

Bayes theorem on Mathematical Representation:

P(A\B)=P(B\A)∗P(A)P(B)

Here,
P (A) => independent probability of A (prior probability)
P (B) => independent probability of B

P (B\A) => conditional probability of B given A (likelihood)
P (A\B) => conditional probability of A given B (posterior probability).

Naïve Bayes is a simple and powerful algorithm for predictive modeling. This model is the most effective and efficient classification algorithm which can handle massive, complicated, non-linear, dependent data. Naïve comprises two part namely naïve & Bayes where naïve classifier assumes that the presence of the particular feature in a class is unrelated to the presence of any other feature.

C) Machine Learning Algorithm K-Nearest Neighbor (KNN)

The K-Nearest Neighbors method is based on the Supervised Learning approach and is one of the most basic Machine Learning algorithms. The K-NN method assumes that the new case/data and existing cases are comparable and places the new case in the category that is most similar to the existing categories. The K-NN method saves all available data and classifies a new data point based on its similarity to the existing data. This implies that fresh data may be quickly sorted into a well-defined category using the K-NN method. The K-NN approach may be used for both regression and classification, however it is more commonly utilized for classification tasks. The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data.

It's also known as a lazy learner algorithm since information doesn't learn from the training set right away, instead storing it and retrieving it later.

The k-nearest neighbors algorithm (k-NN) is a non-parametric classification technique created by Evelyn Fix and Joseph Hodges in 1951[1] and later extended by Thomas Cover in statistics. [2] It is employed in the categorization and regression of data. In both circumstances, the input is a data set with the k closest training samples. Depending on whether k-NN is used for classification or regression, the following is the result:

The outcome of k-NN classification is a class membership. An item is categorized by a majority vote of its neighbors, with the object allocated to the most common class among its k closest neighbors (k is a positive integer, typically small). If k = 1, the item is simply assigned to that single nearest neighbor's class. The result of k-NN regression is the object's property value. This number is the average of the k closest neighbors' values.

k-NN is a special case of a variable-bandwidth, kernel density "balloon" estimator with a uniform kernel.

The naive version of the algorithm is easy to implement by computing the distances from the test example to all stored examples, but it is computationally intensive for large training sets. Using an approximate nearest neighbor search algorithm makes k-NN computationally tractable even for large data sets. Many nearest neighbor search algorithms have been proposed over the years; these generally seek to reduce the number of distance evaluations actually performed.

k-NN has some strong consistency results. As the amount of data approaches infinity, the two-class k-NN algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data). Various improvements to the k-NN speed are possible by using proximity graphs.

For multi-class k-NN classification, Cover and Hart (1967) prove an upper bound error rate of

$$R^* \ \leq \ R_{k\mathrm{NN}} \ \leq \ R^* \left(2 - \frac{MR^*}{M-1}\right)$$

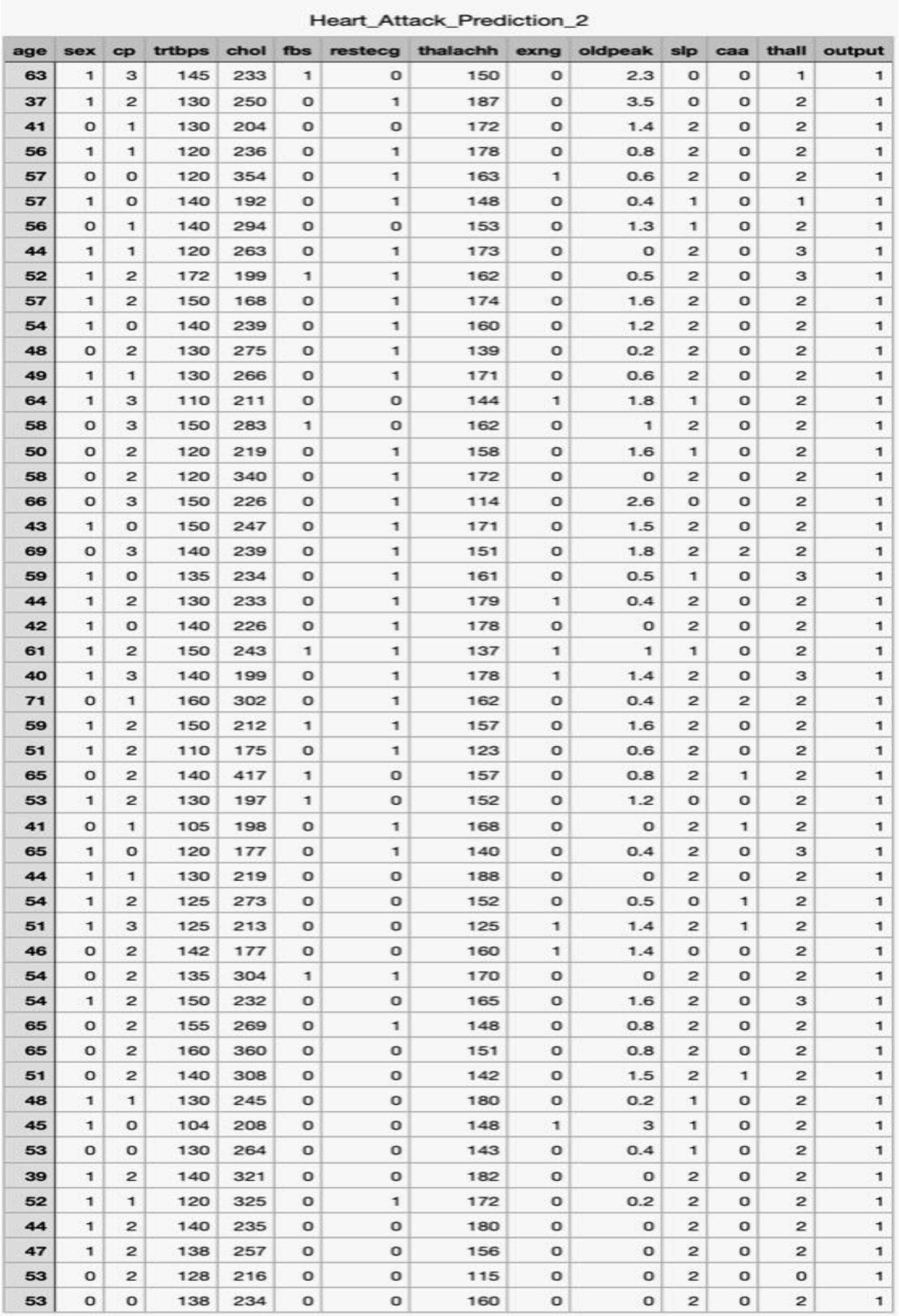
Where,

R* is the Bayes error rate (which is the minimal error rate possible),

R_{kNN} is the *k*-NN error rate, and

M is the number of classes in the problem

D. Datasets



Heart_Attack_Prediction_2

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1
40	1	3	140	199	0	1	178	1	1.4	2	0	3	1
71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
51	1	2	110	175	0	1	123	0	0.6	2	0	2	1
65	0	2	140	417	1	0	157	0	0.8	2	1	2	1
53	1	2	130	197	1	0	152	0	1.2	0	0	2	1
41	0	1	105	198	0	1	168	0	0	2	1	2	1
65	1	0	120	177	0	1	140	0	0.4	2	0	3	1
44	1	1	130	219	0	0	188	0	0	2	0	2	1
54	1	2	125	273	0	0	152	0	0.5	0	1	2	1
51	1	3	125	213	0	0	125	1	1.4	2	1	2	1
46	0	2	142	177	0	0	160	1	1.4	0	0	2	1
54	0	2	135	304	1	1	170	0	0	2	0	2	1
54	1	2	150	232	0	0	165	0	1.6	2	0	3	1
65	0	2	155	269	0	1	148	0	0.8	2	0	2	1
65	0	2	160	360	0	0	151	0	0.8	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
48	1	1	130	245	0	0	180	0	0.2	1	0	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
53	0	0	130	264	0	0	143	0	0.4	1	0	2	1
39	1	2	140	321	0	0	182	0	0	2	0	2	1
52	1	1	120	325	0	1	172	0	0.2	2	0	2	1
44	1	2	140	235	0	0	180	0	0	2	0	2	1
47	1	2	138	257	0	0	156	0	0	2	0	2	1
53	0	2	128	216	0	0	115	0	0	2	0	0	1
53	0	0	138	234	0	0	160	0	0	2	0	2	1

Chapter 4

Requirement Analysis

4.1 Hardware Requirements

- System : Intel Core i3, i5, i7 and 2 GHz Minimum
- RAM : 512Mb or above
- Hard Disk : 10 GB or above
- Input Device : Keyboard and Mouse
- Output Device : Monitor or PC

4.2Software Requirements

- Operating System : Window 7, 10 or higher Versions
- Platform : Jupiter Notebook
- Front End : HTML, CSS, JavaScript, React JS
- Back End : none
- Programming Lang : Python

Chapter 5

Conclusion

Finally, I'd like to state that this project Disease prediction using machine learning is extremely useful in everyone's day-to-day lives, but it's especially important for the healthcare sector, because they're the ones who use these systems on a daily basis to predict the diseases of patients based on their general information and symptoms. Now that the health industry plays such an important role in curing patients' diseases, this is often a useful tool for the health industry to inform the user, and it's also useful for the user if he or she doesn't want to travel to the hospital or other clinics, because simply by entering the symptoms and other relevant information into the form, the user can learn about the disease to which he or she is subjected, and the health industry can benefit as well. Simply ask the user for their symptoms and enter them into the system, and they will give you the precise and, to some degree, accurate diseases in a matter of seconds. If the health sector embraces this idea, doctors' workload will be decreased and they will be able to forecast the patient's sickness more readily. The purpose of disease prediction is to provide predictions for a variety of common diseases that, if left untreated and often ignored, can progress to deadly disease and cause a slew of problems for the patient. This project may be updated in the future by adding additional attributes to the dataset and making it more interactive for the users. It can also be done as a mobile application. We'll make changes to the system by linking it to the hospital's database.