ABSTRACT  
With the start of the Industrial Revolution in the 1870s, the world has witnessed an exponential rise in data volumes, ushering in the digital age. The International Data Corporation (IDC) projects that by 2025, there will be 175 zettabytes in the global datasphere, up from 33 zettabytes in 2018. Since COVID-19, it has become more crucial than ever to use sophisticated machine learning (ML) algorithms to detect possible outbreaks; nevertheless, proper datasets are required for efficient training.  
  
In the healthcare sector, machine learning and data analytics are closely related because ML technologies depend on strong health datasets to predict accurate results. Applications of machine learning (ML) in healthcare span a wide range, from individualized therapies for genetic disorders, chronic illnesses, and surgical procedures to disease analysis via medical imaging. In the medical field, predictive analytics can be used to anticipate disease outbreaks, customize treatments, and enhance operational effectiveness in areas like staff scheduling and medication creation. For example, Pfizer used decision-making techniques to boost its drug development success rate from 2% to 21% between 2010 and 2020.  
  
Using a healthcare dataset downloaded from Kaggle in 2024, this study analyzes five machine learning algorithms: random forest, K-nearest neighbor (KNN), support vector machine (SVM), decision tree, and logistic regression. The dataset is processed through Python using the Anaconda interface. It consists of 349 patients and ten features (illness, fever, cough, fatigue, difficulty breathing, age, gender, blood pressure, cholesterol level, and outcome variable).  
  
Based on statistical study, the patients' average age is 46, and their gender distribution is balanced. The dataset shows a strong link between blood pressure and cholesterol as well as a high prevalence of chronic diseases. When ML algorithms are tested, Random Forest and KNN exhibit the best accuracy, whereas SVM has the worst performance. The study highlights how crucial it is to choose a model depending on the properties of the dataset.  
  
Gradient boosting, feature importance analysis, bootstrapping, and error estimates are some suggestions for enhancing ML performance. These tactics seek to improve the accuracy and dependability of the model.  
  
This paper underscores the transformative power of data analytics in healthcare, promoting evidence-based practices and better decision-making to enhance patient outcomes.

# **1. INTRODUCTION**

Since the 1870s , industrial revolution has been led to endless loop since which continues into digital age as we experience at today also data volume around the world doubling for every 3 years (Reinsel D. , 2018).According to International Data Corporation (IDC) it has predicted the global datasphere will grow from 33 Zettabytes up to 175 Zettabytes between 2018 - 2025.In recent years especially after COVID-19 outbreak,monitoring of either potential outbreaks or another cases across various regions gained crucial role and it can be handled by using advance level machine learning algorithms.However to use this sophisticated technology which has to be equipped with appropraite datasets due to training stage.

ML techonologies and data analytics area are related to each other strictly in healthcare industry becuase without health database ML technologies can’t be exist.In other words which means that ML technologies are trained based on suitable database then which tries to forecast the most reliable outcome towards to the goal.

As mentioned before there are several areas for healthcare industry which can be integrated into combination of data analytics and machine learning.

Firstly,machine learning algorithms can analyze vairous disease such as cancer,multiple sclerosis (MS),rheumatism etc. by medical imaging.Also as a latest trend predictive analytics is very useful to forecast potential outtbreaks around the world.

Secondly,another usage area is providing personalized treatment to patient across genetic issues,chronic disease and surgical operations. Furthermore, the medications and therapies we develop are provided based on statistical averages and evaluated on large populations. For instance, only half of people who take any particular prescription medication that is now available find it to be effective. Only over 60% of people who take antidepressants report feeling better. This same research is increasing our ability to predict which medical treatments will be safe and effective for each patient, and which ones will not be. (Salehi B, et al. 2009).

Thirdly, data analytics is able to bring well approach to staff scheduling which contributes time saving as well as economical income for healthcare instutions.Moreover this level technology may lead to decrease staff demands for jobs such as appointment scheduling,billing etc. as they can be automated easily.

Next field is very critical and it is about drug development.Data science (analyst) help to researcher to analyze clinical trials accurately for either developing new drugs or improving existing ones. For example by using decision making method Pfizer has been increased end to end success rate regarding to drug development from %2 to %21 between 2010-2020 period (Kathy Fernando,et al. 2021). Furthermore another major point that this science is able to track long term side effect on of medicines on various patient groups.

Lastly in Germany the health-care reforms are too slow and too cautious, because policy change could worsen the situation for important stakehold-ers like hospitals as well as sickness funds (Christoph F Dietrich, [Petra Riemer-Hommel](https://www.researchgate.net/profile/Petra-Riemer-Hommel?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19) 2012) also German citizens think there is little time to discuss with doctor by patients.Therefore ML powered chat portals for patients which can assist to people towards to potential problem of their health by using suitable datasets.At the same time for this type chatbots there are virtual versions available.

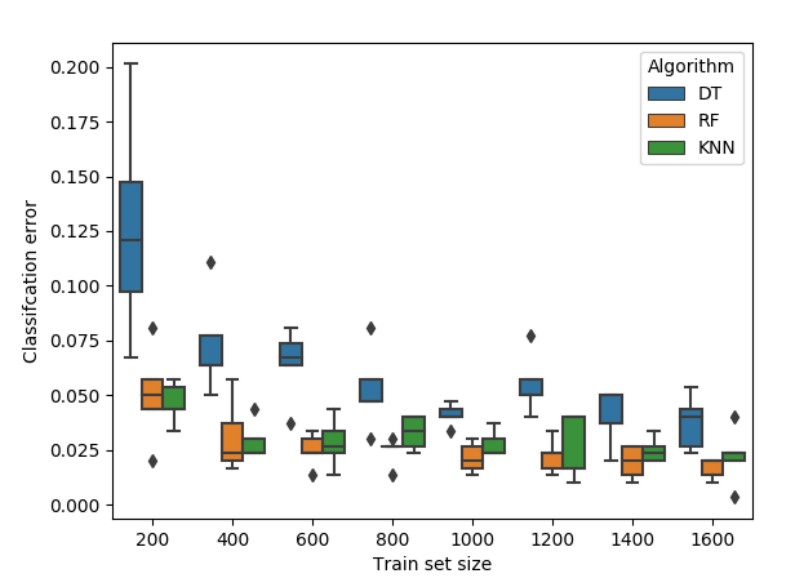
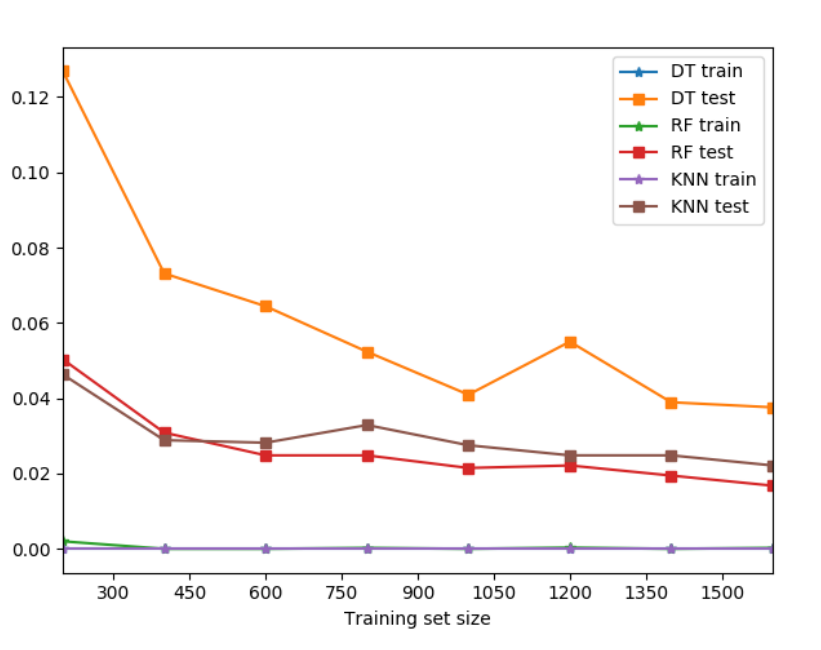
If new technology is reviewed The Internet of Things (IoT) can be cosnidered as well alternative which enables to collect data from daily routine acitivites of people then which can make decision towards improving advice via smart watch , mobile phones.That means integration of data into healthcare industry aims to meet healthcare demands.As amount of data increases data as a term is evolving to big data so there has been a rapid emergence of [big data analytics](https://www.sciencedirect.com/topics/computer-science/big-data-analytics) which is a key initiative within the IoT field since data is increasingly moving at an uncontrollable scale also the proliferation of sensors and smart devices (T. Ojha, et al. 2019). Moreover, it is approximated that in the year 2025, IoT might generate a significant annual effect of about $2.7 trillion to $6.4 trillion, thus by 2030 about 8 billion and about 25 billion active smart gadgets are targeted to be interlinked and woven through a single massive network of information (OECD,2018).

With the potential to significantly improve patient care, diagnosis, and treatment, machine learning and data analytics in healthcare have a bright future. Predictive analytics, which makes use of enormous volumes of medical data, can detect disease outbreaks, tailor treatment regimens, and foresee patient deterioration—all of which improve proactive care. Because machine learning algorithms can identify trends in large, complicated datasets, they will make it possible to use genetics and sophisticated imaging to diagnose diseases like cancer early. AI integration in healthcare systems will also simplify administrative duties, lower expenses, and boost operational effectiveness. As these technologies develop, it will be crucial to take data security, fair access, and ethics into account to make sure that everyone can benefit from these advancements.

# **2. RELATED WORKS**

In the paper 5 different ML algorithms are tested also there are available another studies regarding to performance results as well as comparision of various methods.

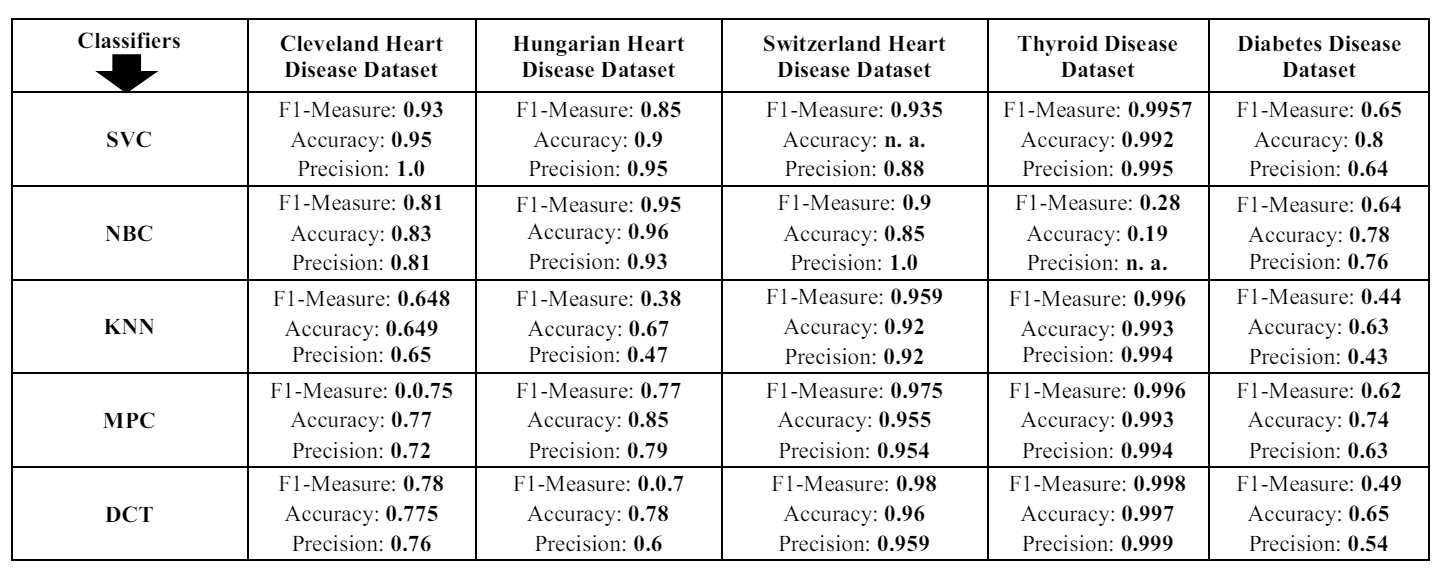
Silvestrin’s (2019) study at Vrije Universiteit Amsterdam has been observed Decision Tree,Random Forest,K-Nearest Neighbor on the basis of perfomance results across Hydraulic System dataset.

Fig. 2.1 The reported classification error on the test set for the other machine learning algorithms after training with datasets of varied sizes (Slivestrin 2019:6)

* As baseline techniques, Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN) are assessed.
* In every situation, Random Forest exhibits less inaccuracy than single trees.
* With 400 training instances, KNN and Random Forest both perform comparably, with an error of about 3%.
* More than RF and KNN, DT has a tendency to overfit.
* All algorithms exhibit decreasing variability in their output as the size of the training set grows.

Silvestrin’s study has been demosntrated the performance results across different case but Singhs (2019) paper includes dataset about heart disease,thyroid and diabetes then which compare performance results ML algorithms according to the data.

Utizied algorithm for this paper are Support Vector Machine, Naïve Bayesian Classifier, K-Nearest Neighbors Classifier, Multilayer Perceptron Classifier, and Decision Tree Classifier.However this study is interested in KNN , SVC and DT.

Fig. 2.1 Benchmark Score of 5 Machine Learning Algorithms for 5 Healthcare Dataset (Singhs 2019:5)

* For issues related to heart,support vector classifier is able to demonstrate better performance results than KNN and DCT at 3 different medical dataset towards to heart problems.
* For thyroid disease all of 3 algorithms can give performance as much as possible thoguh there are very little differences.
* Contrastly as a chronic problem for diabetes the situation differentates from other 2 disease categories because performance score goes down significantly across 3 ML methods but still SVC is the best among them.

Silvestrin (2019) and Singh (2019) have published studies that highlight significant discoveries on machine learning algorithms for health datasets. On a Hydraulic System dataset, Silvestrin discovered that Random Forest (RF) performed better than Decision Tree (DT) and K-Nearest Neighbors (KNN), with RF exhibiting the least amount of inaccuracy. According to Singh's research, all algorithms performed similarly for thyroid disease, but Support Vector Machine (SVM) fared best for heart disease and remained stable despite a decline in performance in diabetes datasets. These results highlight the high performance of SVM and the robustness of RF in particular medical applications, highlighting the significance of selecting the appropriate method depending on the dataset and situation.

# **3. METHODOLOGY**

The dataset was taken from Kaggle in 2024 and which was built up based on 10 features such as disease,fever,cough,fatigue,difficulty breathing,age,gender,blood pressure cholesterol level and outcome variable.Moreover the dataset contains 349 patients acroos these features.

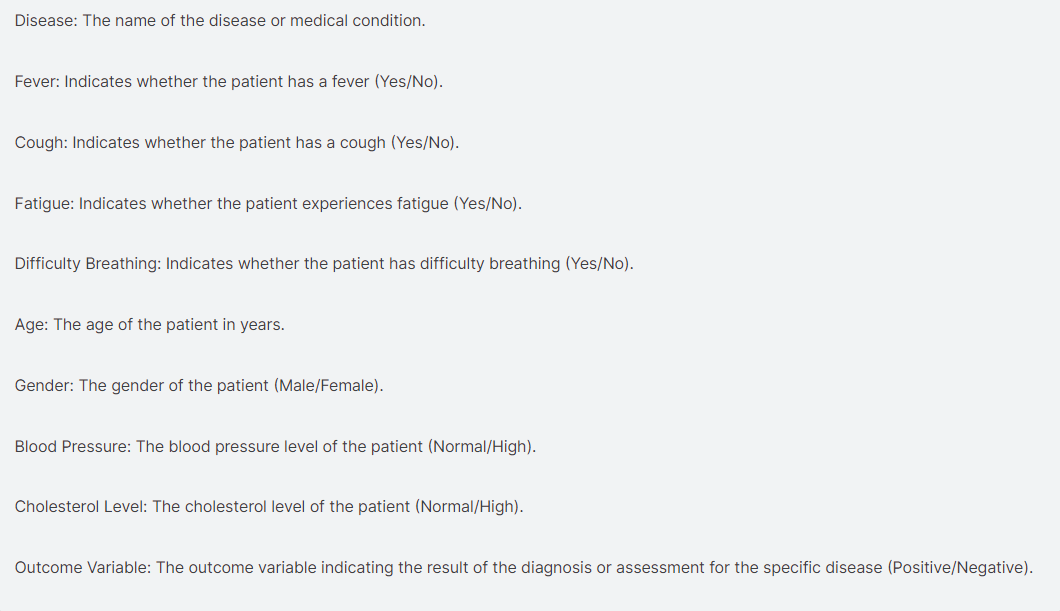


Figure 1.1 Details of Attributes in the Dataset

Several stakeholders are able to use this dataset, including:

Healthcare Professionals: This dataset can be used for clinical analysis, research studies, and epidemiological investigations pertaining to various diseases by doctors, medical practitioners, and researchers. It might help to comprehend the frequency and symptom patterns among people with particular medical disorders.

Medical Researchers: The dataset can be utilized by researchers who are interested in certain diseases or conditions to investigate correlations between symptoms, age, gender, and other factors. The development of fresh perspectives, therapeutic approaches, and preventative measures can benefit from this data.

Healthcare Technology Companies: This dataset can be used by companies creating AI algorithms, diagnostic tools, or healthcare applications to train and validate their models. The information can help with the creation of prediction models for tracking or diagnosing diseases.

In machine learning world basically there are two types of algortihm.These algorithms are called as supervised,unsupervised and reinforcement learning..

Supervised learning involves training a model that uses a labeled dataset to map input data to corresponding output labels.The algorithm adjusts parameters during training to minimize the difference between the predicted output and the actual labeled output.This approach is often used in classification and regression tasks where metrics such as accuracy and precision are used to evaluate model performance.

On the other side, unsupervised learning deals with unlabeled data and aims to discover patterns and structures in the data without explicit guidance about the correct output.Unsupervised learning algorithms, such as clustering and dimensionality reduction, explore the inherent structure of data. Actually we can comparision these two algorithms like that unsupervised learning aims to find exit of maze by trial and error but supervised learning achieve that by using map.

Reinforcement learning allows machines to determine automatically its behavior within a specific context to maximize its performance. Simple reward feedback is required to learn its behavior known as reinforcement signal. This learning occurs when you present the algorithm with examples that lack labels, as in unsupervised learning. However, you can accompany an example with positive or negative feedback per the solution the algorithm proposes (Subbu R. ,2017).

Moreover the dataset have labels as a result of this condition we need to select supervised learning algorithms.Moreover which includes Logistic Regression,K-Nearest Neighbour,Support Vcctor Machine,Decision Tree and Random Forest.

The dataset is handled on Python by Anaconda interface since the primary language is Python, and the entire suite of tools included with Anaconda includes R, Python, and data science tools. Because of its pre-configured environment and package management capabilities, Anaconda might be a better option if you're working on data analysis, machine learning, or scientific computing.

About technical equipment the system has Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz with 8GB RAM as well as NIVIDA GEFORCE GTX 1050Ti 4GB GPU but accelareted GPU features wasn’t used for paralell computation.

# **4. ANALYSING of THE DATASET**

The data process contains basically 4 steps these stages are cleaning the dataset from either missing points or noisy data , visualization of the dataset across attributes as well as extract statistical analysis and apply various ML algorithms onto the dataset then compare benchmark results to decide best method for selected dataset.

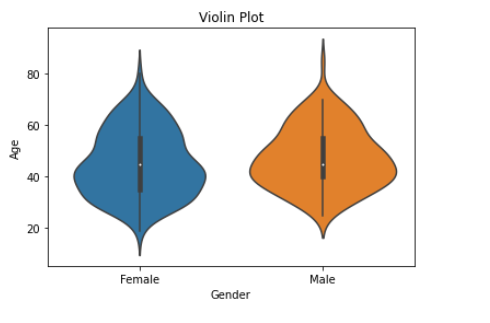
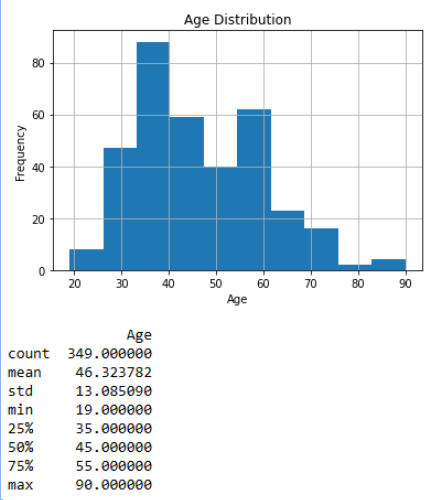
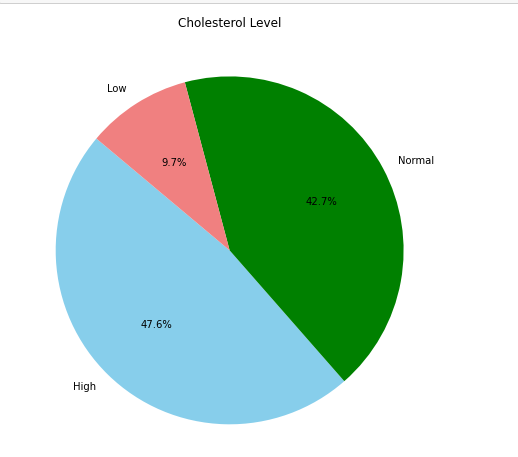
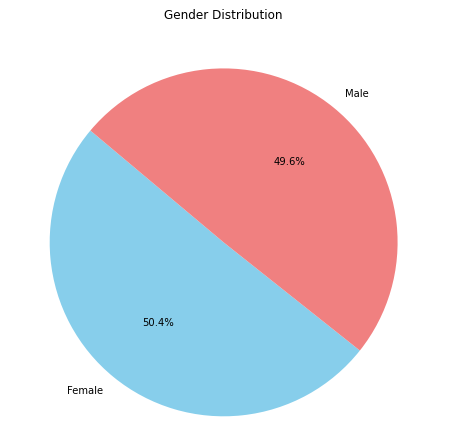


Figure 2.1 Statistical Analysis of Ages Based on General and Gender



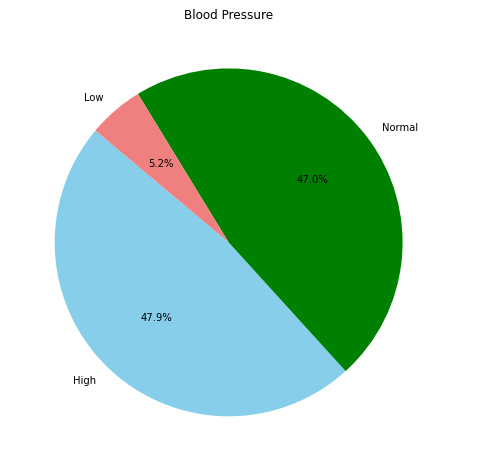
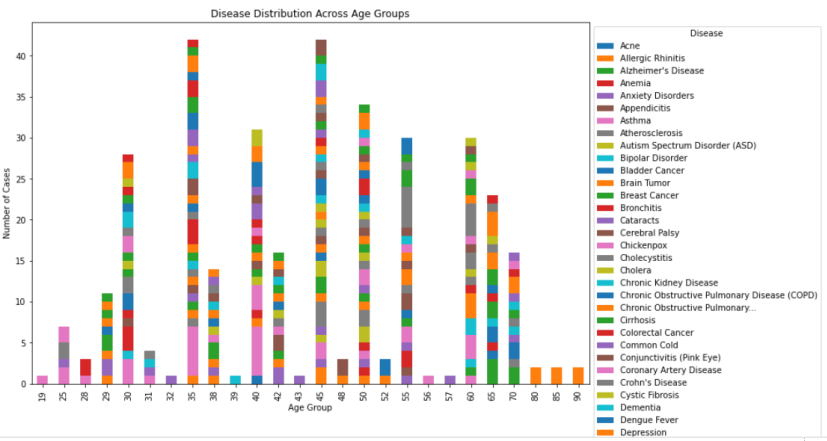


Figure 2.2 Distribution of Gender - Cholesterol Level – Blood Pressure Across Dataset

* Average age of people in the dataset is around 46 years old and there is no observable difference between genders.
* Homogeneous of gender is perfect because percentage of male-female %49.6 and %50.4 respectively so extracted outcome from the data migh be more compherensive as well as meaningful.
* Roughly half of the population have normal level of blood pressure and cholesterol level.Majority of remaining show high level regarding to these health problems.



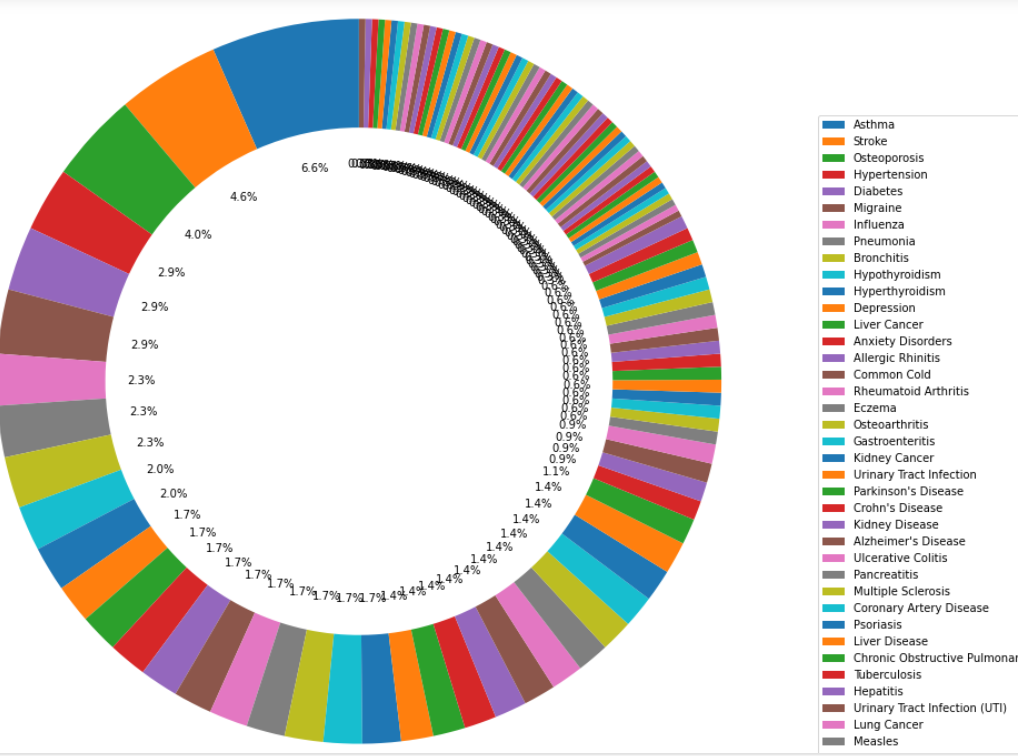
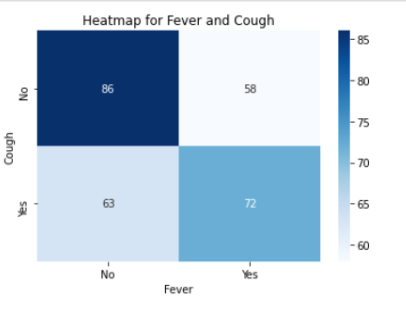
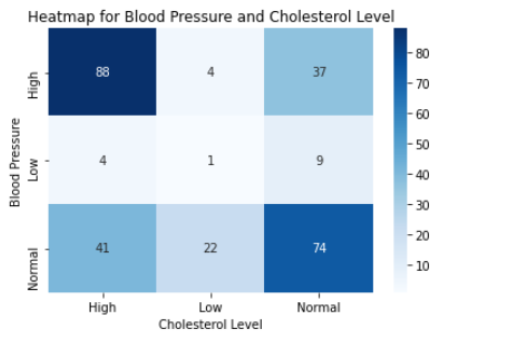
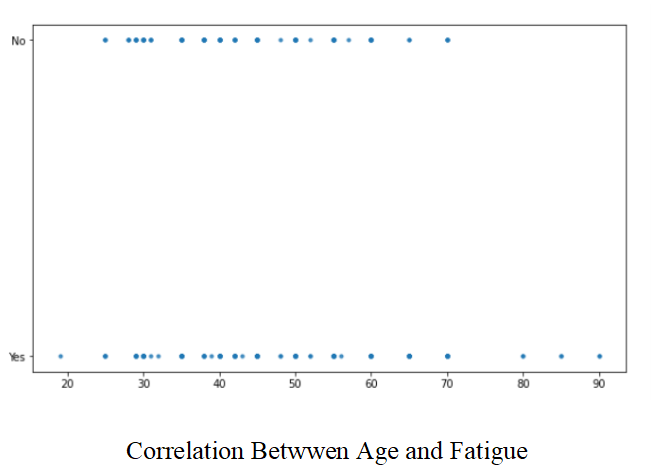
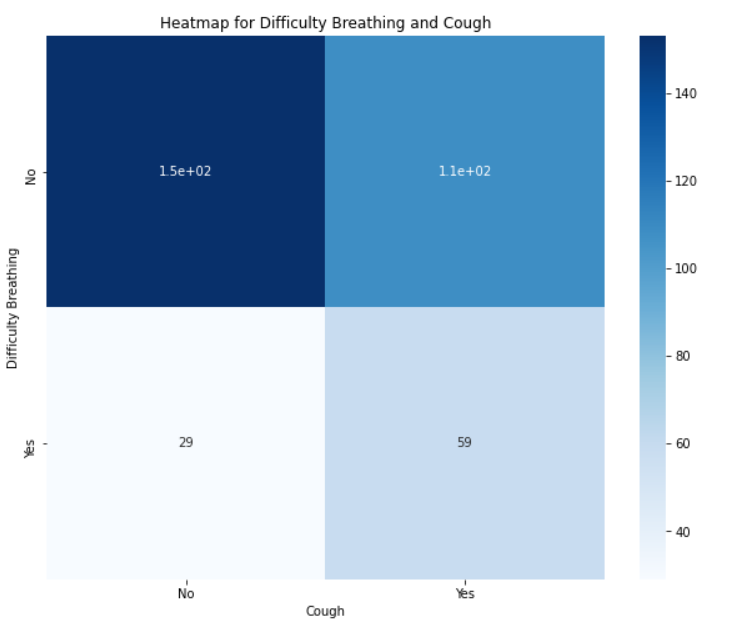


Figure 2.3 Disease Distrubition Across Age and General Review

* After age 35 general situation demonstrates there is increasing about vairous diesases.Also around 35-45 years old which approachs its peak level based on number of kinds health issues.
* Majority of illnesses belongs to chronicle category such as asthama,stroke,hypertension,diabetes,migraine also this insight supports previous comment (highlights middle age) as the population ages, so does the prevalence of chronic disease (Hacker K. , 2024).





* According to Ulfah et al there is no strong correlation between cholesterol level – blood pressure (2017:28). However in our dataset this situation is different than this research because there is a positive correlation between of them. Most probably this case is related to higher average age of patients.These people are able to have classical middle age health problems such as cholesterol and blood pressure independently from each other that’s why there seems a positive correlation.
* Most of the time if any person has cough this problem continue fever due to existence of infections in blood such as COVID-19 , RSV or other seasonal influenza.Mary C. suggest this opinion by sample COVID cases (2022:2).Also in the dataset there is positive correlation between these two attributes.
* There is no seems meaningful relatiosnhip among cough and breathing.
* Interestingly , fatigue is widespread throughout all ages scale , becuase a wide range of medical problems and decompensation processes of underlying diseases can cause fatigue. Sleep disorders, sleep-related respiratory disorders, depression, and psychosocial stress are the main causes of persistent fatigue (Maisel P. et al. , 2021)..

# **5. COMPARISION of PERFORMANCE RESULTS**

## **5.1 Overview of ML Algorithms**

Boateng’s study (2020) which descirbes K-Nearest , Support Vector Machine and Random

Forest as follows;

In pattern recognition, the KNN algorithm is an instance based learning method used to classify objects based on their closest training examples in the feature space. An object is classified by a majority vote of its neighbors, that is, the object is assigned to the class that is most common amongst its k-nearest neighbors.

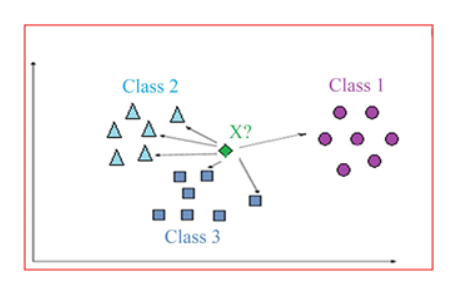


Figure 5.1 A simple pictorial overview of the K-Nearest Neighbor (KNN) algorithm

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The SVM algorithm tries to find the optimal hyperplane in n-dimensional classification space with the highest margin between classes.

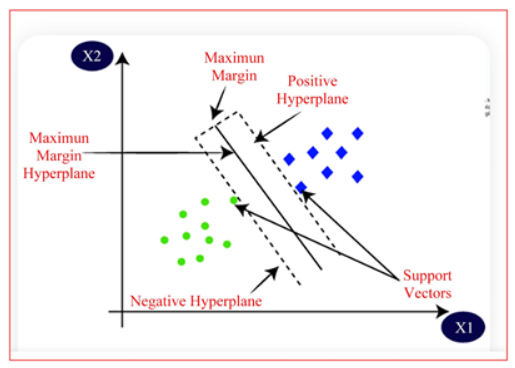


Figure 5.2 A simple illustration of the Support Vector Machine (SVM) algorithm in 2-dimensions

An Random Forest classifier consists of a number of trees, with each tree grown using some form of randomization. The leaf nodes of each tree are labeled by estimates of the posterior distribution over the image classes. An image is classified by sending it down every tree and aggregating the reached leaf distributions. Randomness can be injected at two points during training: in sub-sampling the training data so that each tree is grown using a different subset, and in selecting the node tests.

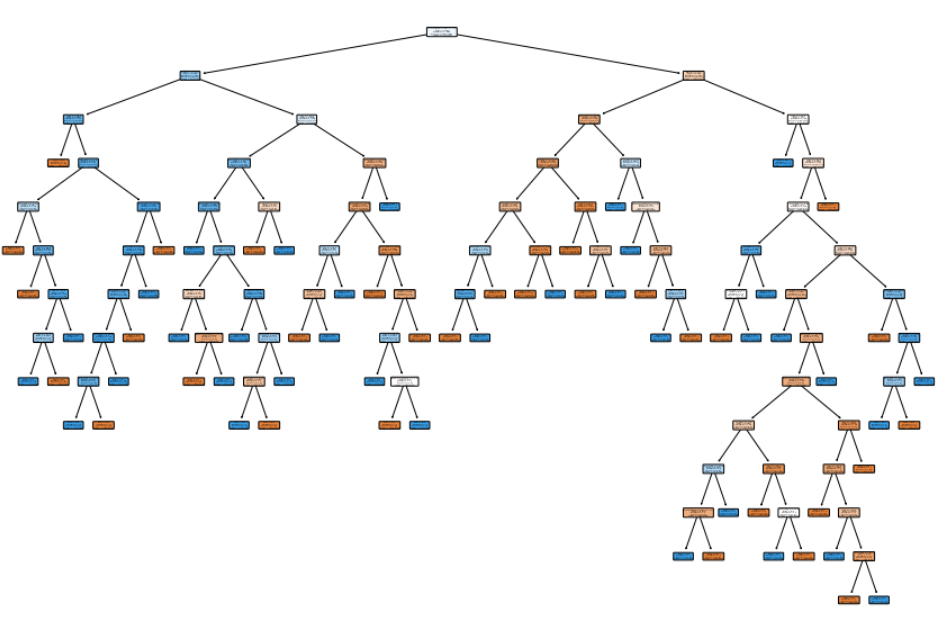


Figure 5.3 Real Visualization of Decision Tree by The Dataset

Logistic regression uses a sigmoid function to model the probability that an instance belongs to a particular class.The training process involves adjusting model parameters to maximize the probability of observed data and minimize the associated log loss.The decision boundary is determined by a threshold applied to the predicted probabilities.Logistic regression is a fundamental algorithm for binary classification tasks and is widely used due to its simplicity and ease of interpretation.Furthermore sigmoid function can be defined as .

The logic of a random forest involves building an ensemble of decision trees through bootstrap sampling and random feature selection, and predictions are made by aggregating the predictions of the individual trees.This ensemble approach improves model performance and reduces the risk of overfitting.

## **5.2 Comparision of Benchmark Results**

Evaluating and comprehending the effectiveness, dependability, and performance of various models on benchmark tasks requires comparing ML benchmark outcomes. By giving testing and result verification a common foundation, it guarantees transparency and reproducibility. Furthermore, benchmarking monitors development over time, promoting innovation, enhancing industry best practices, and assisting in the identification of the most successful and efficient models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | K-Nearest Neighbor | Support Vector Ma. | Decision Tree | Random Forest |
| Accuracy | 0.63 | 0.69 | 0.51 | 0.68 | 0.70 |
| Confusion Matrix | [15 15]  [16 24] | [7 6]  [5 17] | [15 15]  [19 21] | [21 17]  [11 39] | [24 14]  [15 35] |
| Precision Negative / Positive | 0.48 / 0.62 | 0.58 / 0.74 | 0.44 / 0.58 | 0.66 / 0.70 | 0.66 / 0.74 |
| Recall Negative / Positive | 0.50 / 0.60 | 0.54 / 0.77 | 0.50 / 0.53 | 0.55 / 0.78 | 0.66 / 0.75 |
| F-1 Negative / Positive | 0.49 / 0.61 | 0.56 / 0.76 | 0.47 / 0.55 | 0.60 / 0.74 | 0.66 / 0.74 |
| Precision Macro Avg / Weight Avg | 0.55 / 0.56 | 0.66 / 0.68 | 0.51 / 0.52 | 0.68 / 0.68 | 0.68 / 0.72 |
| Recall Macro Avg / Weight Avg | 0.55 / 0.56 | 0.66 / 0.69 | 0.51 / 0.51 | 0.67 / 0.68 | 0.69 / 0.69 |
| F-1 Macro Avg / Weight Avg | 0.55 / 0.56 | 0.66 / 0.68 | 0.51 / 0.52 | 0.67 / 0.68 | 0.70 / 0.70 |

* The models that perform best overall and with the highest accuracy are Random Forest and K-Nearest Neighbor.
* The fact that Support Vector Machine has the lowest performance metrics suggests that it might not be the best choice for this particular healthcare dataset.
* In general, Decision Tree outperforms Logistic Regression in terms of precision, recall, and F1-score balance.

Given its superior performance, Random Forest is a top choice for this healthcare dataset.

These findings emphasize how crucial model selection is in machine learning, both dataset characteristics and the technique employed can have a big impact on performance.

# **5. RECOMMENDATIONS and POSSIBLE IMPROVING**

Machine learning algorithms are fundamental instruments for information driven bits of knowledge, yet experiencing low execution is a common issue. Factors, for example, dataset quality, highlight designing, and hyperparameter tuning essentially impact calculation viability. This concise outline investigates the purposes for sub-standard outcomes, giving a brief look into methodologies to further developing execution in that field.Moreover overcome to this problem we may follow some solutions.

## **5.1 BootsStrapping**

Bootstrapping can be described as like this.Assume that we need to predict price of a car out of 100 sample but the accuracy might be low due to variance so for removing this problem we are able to split dataset into 2 groups by randomly in that case variance will be decreased dramatically and estimation has high reliability.

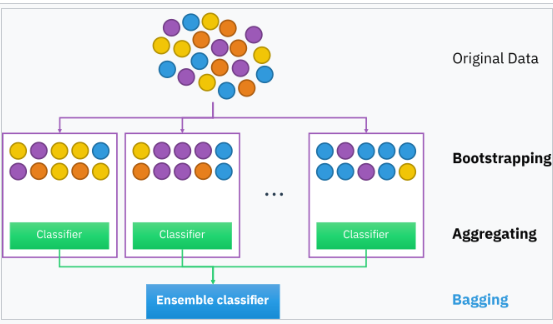


Figure 5.1 Illustration of Bootstrapping (Shinde 2014:42)

In that solution , decision trees are grown deep not pruned so high variance occurs and low bias occurs.Averaging these trees which decrease variance dramaticially.

## **5.2 Error Estimation**

Out-of-Bag Error Estimation is a straightforward approach to estimating the test error of a bagged model without requiring cross-validation. The perceptions not used to fit a given stowed tree are alluded to as the out-of-pack (OOB) perceptions. We can just anticipate the reaction for the ith perception involving every one of the trees wherein that perception was OOB. We normal those anticipated reactions, or take a greater part vote, in the event that assuming that the reaction is quantitative or subjective. It is possible to calculate a classification error rate or overall OOB MSE (mean squared error). This is an adequate test mistake rate in light of the fact that the forecasts depend on just the trees that were not fit utilizing that perception.

## **5.3 Feature Importance**

After obtain computing results we can’t interpret them easily at every time and not clear which variable is more important to relationship.Therefore, in this method you can detect which variable at each split point has higher influence on computing results than others. In regression problems this might be the decrease in residual sum of squares and in classification this might be the Gini score.As we mentioned it on decisison tree algorithm before the highest influence within labels belongs to diseases.

## **5.4 Gradient Boosting**

In gradient boosting, an outfit of weak learners is utilized to make strides the execution of a machine learning demonstrate. The weak learners are ordinarily decision trees. Combined, their yield comes about in superior models.

In case of regression, the ultimate result is produced from the normal of all weak learners. With classification, the ultimate result can be computed as the lesson with the biggest share of votes from weak learners.

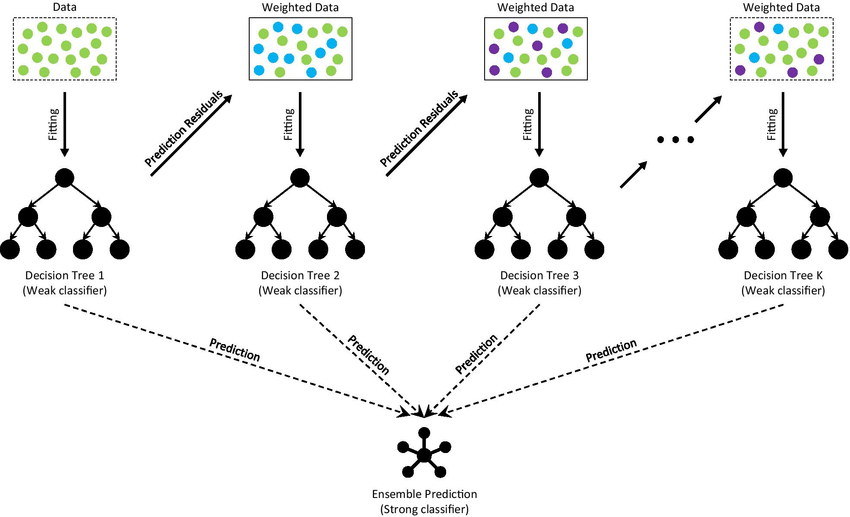


Figure 5.2 Illustration of Gradient Boosting (Haowen 2021:4)

In gradient boosting, weak learners work successively. Each show tries to improve on the blunder from the past model. Usually distinctive from the bagging method, where a few models are fitted on subsets of the information in a parallel way. These subsets are as a rule drawn arbitrarily with substitution. A extraordinary case of sacking is in random forests.

# **CONCLUDING REMAKRS**

Within the quickly advancing scene of healthcare, the integration of information analytics has risen as a effective device for changing and making strides clinical care. With an phenomenal volume of information produced day by day and headways in innovation, the healthcare division has the opportunity to use data-driven approaches to upgrade understanding care. The venture, titled "Comparision of Machine Learning Algorithms from Sample Healthcare Dataset," dives into the central part of information analytics in reshaping healthcare benefit conveyance, proficiency, and quality.

The investigation includes different viewpoints of information examination, counting prescient modeling, manufactured insights, information visualization, and design acknowledgment. The extend looks for to reveal covered up designs, connections, and patterns inside clinical datasets, giving important bits of knowledge to illuminate evidence-based hones and drive changes in understanding care.

Recognizing the significance of moral contemplations, information protection, and security in healthcare, the extend emphasizes the require for a system that guarantees capable and secure taking care of of delicate persistent data, following to administrative standards and keeping up believe within the healthcare framework.

In conclusion, the venture grandstands the potential of information analytics in revolutionizing healthcare. Through cautious investigation, visualization, and the execution of machine learning calculations, the venture seeks to contribute to the progression of restorative care, eventually driving to way better understanding results and more educated decision-making by healthcare experts.

# BIBLIOGRAPHY

[1] Akhondzadeh, S. (2014). Personalized medicine: a tailor-made medicine. *Avicenna Journal of Medical Biotechnology*, 6(4), p.191. PMID: 25414780; PMCID: PMC4224657.

[2] Boateng, E.Y., Otoo, J. and Abaye, D.A. (2020). Basic tenets of classification algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A review.

[3] Dietrich, C.F. and Riemer-Hommel, P. (2012). 'Challenges for the German Health Care System', *Zeitschrift für Gastroenterologie*, 50(6), pp. 557-572. doi: 10.1055/s-0032-1312742.

[4] Fernando, K., Menon, S., Jansen, K., Naik, P., Nucci, G., Roberts, J., Wu, S.S. and Dolsten, M. (2022). Achieving end-to-end success in the clinic: Pfizer's learnings on R&D productivity. *Drug Discovery Today*, 27(3), pp. 697-704. doi: 10.1016/j.drudis.2021.12.010. Epub 2021 Dec 15. PMID: 34922020; PMCID: PMC8719639.

[5] Hacker, K. (2024). The Burden of Chronic Disease. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, 8(1), pp. 112-119.

[6] Haowen, D., Youyou, Z., Lin, W. and Cheng, Z. (2021). ‘Ensemble Learning for The Early Prediction of Neonatal Jaundice with Genetic Features.’ *Journal BMC Medical Informatics and Decision Making*, 4.

[7] Maisel, P., Baum, E. and Donner-Banzhoff, N. (2021). Fatigue as the Chief Complaint–Epidemiology, Causes, Diagnosis, and Treatment. *Deutsches Ärzteblatt International*, 118(33-34), pp. 566-576.

[8] OECD (2018). Achieving Inclusive Growth in the Face of Digital Transformation and the Future of Work. Report to G-20 Finance Ministers, 19 March.

[9] Ojha, T., Misra, S., Raghuwanshi, N.S. and Poddar, H. (2019). DVSP: Dynamic Virtual Sensor Provisioning in Sensor–Cloud-Based Internet of Things. *IEEE Internet of Things Journal*, 6(3), pp. 5265-5272.

[10] Reinsel, D., Gantz, J., and Rydning, J. (2018). The Digitization of the World: From Edge to Core. An IDC White Paper, Sponsored by Seagate. Available at: [<https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pd f>] (Accessed: 25.05.2024).

[11] Silvestrin, L.P., Hoogendoorn, M. and Koole, G. (2019). A Comparative Study of State-of-the-Art Machine Learning Algorithms for Predictive Maintenance. 2019 *IEEE Symposium Series on Computational Intelligence* (SSCI), Xiamen, China, pp. 760-767.

[12] Singh, A.K. (2019). A comparative study on disease classification using machine learning algorithms. In: 2nd International Conference on Advanced Computing and Software Engineering (ICACSE-2019).

[13] Subbu, R.S. (2017). Brief study of classification algorithms in machine learning. CUNY City College.