Disease Prediction Using Machine Learning

Vishnukant Shukla, Piyush Varshney, Harsh Sharma

GLA University, Mathura

Abstract – With current developments in artificial intelligence, the integration of AI and Machine Learning into healthcare systems has shown promising results in disease prediction, diagnosis, and prognosis. This paper presents a comprehensive review of the methodologies, challenges, and advancements in prediction of diseases using machine learning algorithms. The main foundation of paper is on the application of various machine learning models such as Support Vector Machines, Decision Trees, Logistic Regression, K Nearest Neighbours and Naïve Bayes, in predicting diseases across different medical domains.

The paper begins by outlining the significance of disease prediction in early diagnosis and emphasizing the potential of machine learning to enhance accuracy and clinical outcomes. It then discusses the process of data acquisition, preprocessing, and feature selection, highlighting the importance of high-quality data and feature representation in building robust predictive models. Furthermore, it reviews different evaluation metrics and validation techniques commonly employed to compare and analyse the accuracy of models in disease prediction.

The main research delves into a comprehensive analysis of machine learning techniques applied to various diseases, including Heart Attack, Malaria, Diabetes, Fungal Infection, Typhoid, Jaundice, Pneumoina and others. The paper examines the specific challenges, datasets, features, and algorithms utilized, along with the reported accuracies and limitations of existing models.

Finally, the paper concludes with accuracy scores of K Nearest Neighbours (KNN), Decision Trees, Logistic Regression, Support Vector Machines (SVC), and Naïve Bayes algorithms. The model aims to handle overfitting and provide a good fit to handle different scenarios.

Keywords: Machine Learning, K Nearest Neighbours (KNN), Decision Trees, Logistic Regression, Support Vector Machines (SVC), Naïve Bayes.

Introduction

Recent years have witnessed remarkable progress at the nexus of healthcare and machine learning, driving substantial advancements in disease prediction. By harnessing extensive datasets, machine learning algorithms exhibit exceptional proficiency in deciphering intricate patterns and forecasting disease trajectories with unprecedented precision. This paper seeks to offer a thorough exploration of machine learning applications in disease prediction, underscoring their transformative potential in healthcare delivery and patient welfare enhancement.

Machine learning (ML) holds immense potential in disease prediction for several reasons:

1. Early Detection: ML algorithms can analyse vast amounts of patient data, including demographic information, medical history, genetic factors, and lifestyle choices, to identify patterns and risk factors associated with various diseases.

2. Risk Management: ML models can categorise individuals into different risk levels based on their possibility of developing certain diseases.
3. Multi-factorial Analysis: ML techniques excel at handling complex, multi-dimensional data from diverse sources.
4. Public Health Surveillance: ML algorithms can analyse population-level data from sources such as health surveys, disease registries, social media, and internet searches to monitor disease trends and outbreaks in real-time. This early warning system allows public health authorities to implement timely interventions and containment measures to prevent the spread of diseases.
5. Low Cost / No Cost Service: ML Models once developed can easily provide no cost disease prediction service to a large number of people hence reducing the workload on hospitals, preventing overcrowding and spread of diseases.

Overall, ML-driven prediction of diseases has an unimaginable amount of potential for improving and modernizing disease treatment, minimizing healthcare costs, and upgrading our known knowledge of disease biology. However, lack of data, data quality, model interpretability, ethical considerations and privacy concerns possess a critical challenge to realization of the full potential of Machine Learning in disease prediction.

Literature Review

According to Quinlan et al. [7] Decision Trees belong to the category of supervised learning algorithm that can be used both for regression and classification purposes.

This technique involves splitting of dataset into subsets recursively based on the most significant feature that separates the data optimally.

Throughout the entire splitting process, the algorithm chooses the feature that minimizes impurity (Gini Index Method) or maximises the information gain (ID3 Method).

Decision tree Algorithm is an easy and intuitive method to interpret and this makes it suitable for tasks that involve explainability and transparency.

However, decision trees may be prone to overfitting, especially with complex datasets or when the tree grows too deep.

According to Hosmer Jr et al [8], Logistic regression is a statistical method that is used mostly for binary classification problems,where the result has only two possible outcomes.

This technique evaluates the probability of a given input belonging to a particular class using the sigmoidal logistic function, which maps the input space to the range [0, 1].

Logistic regression algorithm calculates the coefficients of the input features in order to maximize the likelihood of the data observed.

In contrast to its name, Logistic Regression is a linear model and works well when there is a linear relationship between the input features and the target variable or it can be transformed to linear.

Logistic regression is interpretable, computationally efficient, and robust to noise, but it may struggle with non-linear relationships between features and target.

Rish et al [9], Based on Bayes Theorem, Naive Bayes is a probabilistic classifier which assumes the independence among predictors (hence "naive").

It estimates the probability of a given instance belonging to each class and then selects the class with the maximum probability.

Despite its strong independence assumption and simplicity, Naive Bayes’ performance is often surprisingly well, especially in spam filtering and test classification jobs.

Naive Bayes classifiers are easy to implement, computationally efficient, and work pretty well with high-dimensional data.

However, the independence assumption may not hold true in many real-world scenarios, which can lead to suboptimal performance.

Altman et al [10], K Nearest Neighbours is an instance-based, non-parametric learning algorithm that is used for classification as well as regression problems.
In this method, a new instance is classified as the majority class label among its k nearest neighbours in the training dataset, where k is a user-defined variable.
KNN relies on the notion that similar instances tend to belong to the same class so it is intuitive and easy to implement.
KNN does not require training a model explicitly, making it particularly useful for lazy learning and for handling non-linear decision boundaries.
However, using KNN on large datasets can be computation costly and may also face the curse of dimensionality.

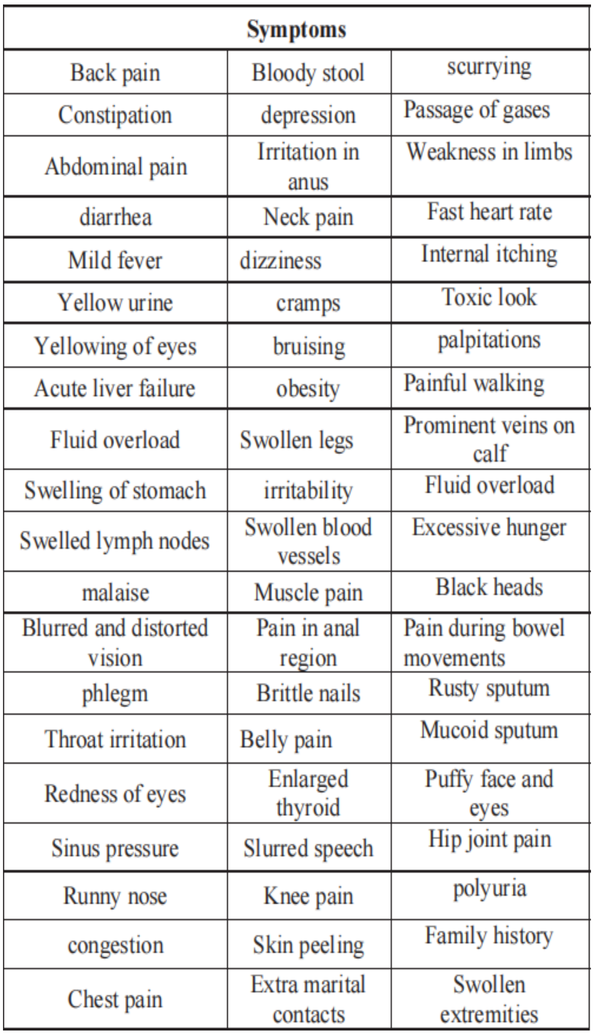
Burges et al [11], Support Vector Machine is a powerful and efficient supervised learning algorithm used not only for regression but also classification tasks.
It finds the best hyperplane that optimally separates the data into different classes in such a way that the margin between the classes is maximum.
SVM uses a variety of kernel functions to handle linear as well as non-linear decision boundaries and map the input data to higher dimensional feature spaces.
SVMs are also effective with datasets of higher dimensions with a small number of training samples.
The choice of kernel and parameters is critical to SVMs, and training time can be significant for large datasets.

Methodology

Dataset

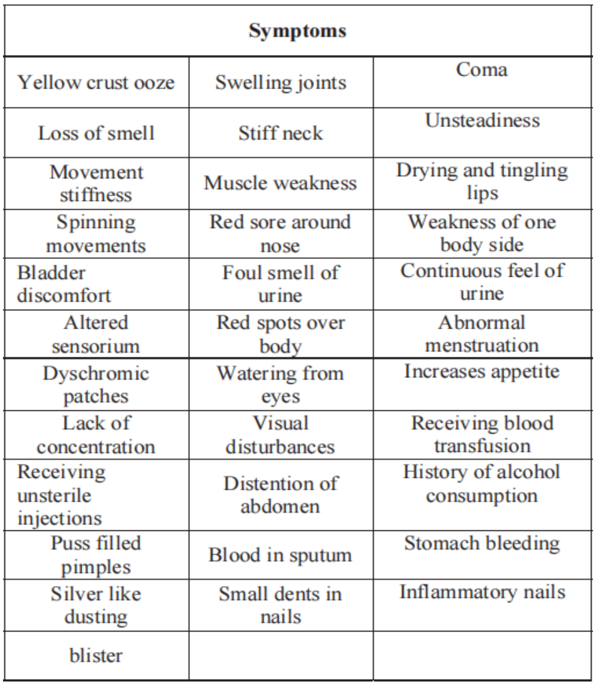
The dataset we have taken consists of 4920 Rows. Each row of data has some symptoms and corresponding disease. There are 41 unique diseases and 131 unique symptoms in this dataset.

The symptoms include:



Symptoms		
Back pain	Bloody stool	scurrying
Constipation	depression	Passage of gases
Abdominal pain	Irritation in anus	Weakness in limbs
diarrhea	Neck pain	Fast heart rate
Mild fever	dizziness	Internal itching
Yellow urine	cramps	Toxic look
Yellowing of eyes	bruising	palpitations
Acute liver failure	obesity	Painful walking
Fluid overload	Swollen legs	Prominent veins on calf
Swelling of stomach	irritability	Fluid overload
Swelled lymph nodes	Swollen blood vessels	Excessive hunger
malaise	Muscle pain	Black heads
Blurred and distorted vision	Pain in anal region	Pain during bowel movements
phlegm	Brittle nails	Rusty sputum
Throat irritation	Belly pain	Mucoid sputum
Redness of eyes	Enlarged thyroid	Puffy face and eyes
Sinus pressure	Slurred speech	Hip joint pain
Runny nose	Knee pain	polyuria
congestion	Skin peeling	Family history
Chest pain	Extra marital contacts	Swollen extremities

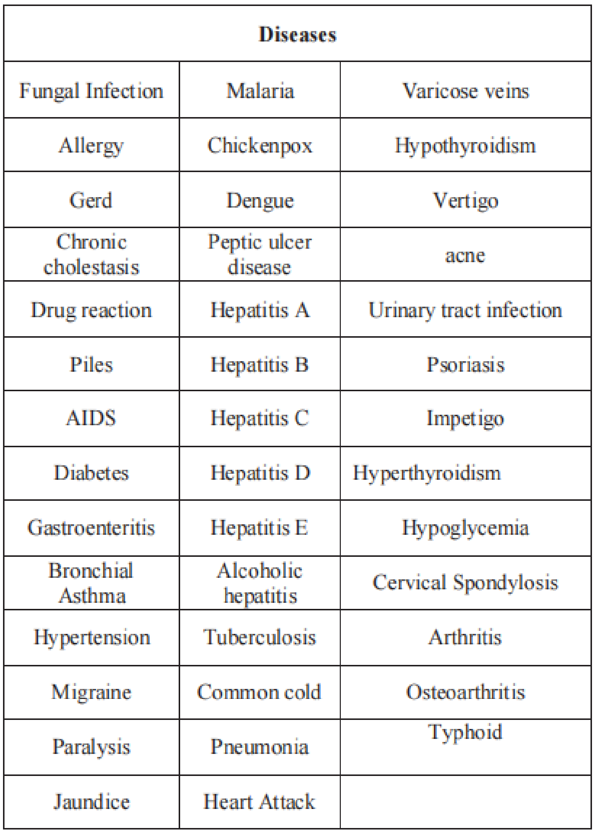
Fig. 1.1 Sneha et al [1] Symptoms Table



Symptoms		
Yellow crust ooze	Swelling joints	Coma
Loss of smell	Stiff neck	Unsteadiness
Movement stiffness	Muscle weakness	Drying and tingling lips
Spinning movements	Red sore around nose	Weakness of one body side
Bladder discomfort	Foul smell of urine	Continuous feel of urine
Altered sensorium	Red spots over body	Abnormal menstruation
Dyschromic patches	Watering from eyes	Increases appetite
Lack of concentration	Visual disturbances	Receiving blood transfusion
Receiving unsterile injections	Distention of abdomen	History of alcohol consumption
Puss filled pimples	Blood in sputum	Stomach bleeding
Silver like dusting	Small dents in nails	Inflammatory nails
blister		

Fig. 1.2 Sneha et al [1] Symptoms Table

The diseases include:



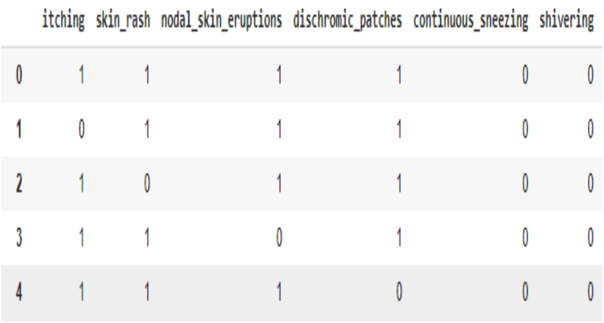
Diseases		
Fungal Infection	Malaria	Varicose veins
Allergy	Chickenpox	Hypothyroidism
Gerd	Dengue	Vertigo
Chronic cholestasis	Peptic ulcer disease	acne
Drug reaction	Hepatitis A	Urinary tract infection
Piles	Hepatitis B	Psoriasis
AIDS	Hepatitis C	Impetigo
Diabetes	Hepatitis D	Hyperthyroidism
Gastroenteritis	Hepatitis E	Hypoglycemia
Bronchial Asthma	Alcoholic hepatitis	Cervical Spondylosis
Hypertension	Tuberculosis	Arthritis
Migraine	Common cold	Osteoarthritis
Paralysis	Pneumonia	Typhoid
Jaundice	Heart Attack	

Fig. 2 Sneha et al [1] Diseases Table

Data Pre-Processing: To Train a ML model we need to clean and process the data. Convert Categorical data to numerical data such that our model can be trained on it. Remove or fill NaN values.

Our Dataset has several diseases and symptoms, we need to make these symptoms as columns and numerical value to represent their existence (1 for present , 0 for absent).

Data after pre processing



	itching	skin_rash	nodal_skin_eruptions	dischromic _patches	continuous_sneezing	shivering
0	1	1	1	1	0	0
1	0	1	1	1	0	0
2	1	0	1	1	0	0
3	1	1	0	1	0	0
4	1	1	1	0	0	0

Fig. 3 Data after pre processing

Models Selected

1. Decision Tree

Decision trees serve as hierarchical constructs utilized in machine learning for regression and classification purposes. They segment the input data into smaller partitions using decision rules, fostering a straightforward comprehension of decision-making mechanisms. Renowned for their simplicity and transparency, decision trees find extensive application across diverse domains, providing valuable insights into predictive modelling and streamlining data-informed decision-making processes.

2. Logistic Regression

Logistic regression, a statistical technique frequently employed in machine learning, is designed for binary classification tasks. By modelling the correlation between independent variables and a binary outcome, it estimates the probability of occurrence. Despite its straightforwardness, logistic regression is valued for its reliability and ease of interpretation, rendering it a favoured approach for predictive modelling across diverse sectors, including healthcare, finance, and marketing.

3. Naïve Bayes

Naive Bayes is a classification algorithm based on Bayes' theorem, assuming independence among predictors. Despite its simplicity and the "naive" assumption, Naive Bayes frequently demonstrates strong performance, particularly in text classification and other tasks involving high-dimensional data. Its effectiveness and efficiency have made it a popular choice for sentiment analysis, spam filtering, and document categorization.

4. K Nearest Neighbours

K-Nearest Neighbours (KNN) is a non-parametric classification algorithm, determining a data point's class label by considering the majority class among its k nearest neighbours. Despite its straightforward approach, KNN proves highly effective in both classification and regression scenarios, especially when confronted with noisy or non-linear data. Renowned for its intuitive operation and adaptability, KNN finds widespread application across diverse domains such as anomaly detection, recommendation systems, and pattern detection.

5. Support Vector Machine

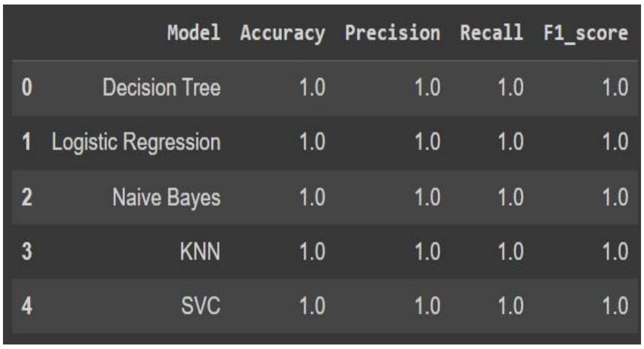
Support Vector Machine (SVM) is a robust supervised learning technique utilized for regression and classification objectives. It identifies the optimal hyperplane that effectively separates classes or accurately fits the data, ultimately minimizing classification errors. Renowned for their adeptness in handling high-dimensional data and nonlinear associations, SVMs find extensive application across diverse fields such as image classification, text categorization, and bioinformatics.

Implementation

Models for KNN, SVC, Decision Tree, Naïve Bayes and Logistic Regression are trained on the processed data. These models are used to get predictions for test data. The predictions are checked with actual data to get accuracy score, precision, recall and f score. All the values are compared to get a comparative analysis of all these algorithms.

Results

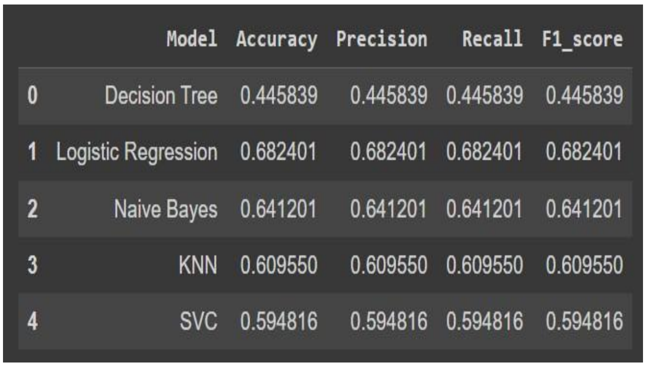
The models trained on the dataset provided us an accuracy of 100% on testing. This suggested that there must be high chance of overfitting.



	Model	Accuracy	Precision	Recall	F1_score
0	Decision Tree	1.0	1.0	1.0	1.0
1	Logistic Regression	1.0	1.0	1.0	1.0
2	Naive Bayes	1.0	1.0	1.0	1.0
3	KNN	1.0	1.0	1.0	1.0
4	SVC	1.0	1.0	1.0	1.0

Fig. 4 Testing on trained data

So, we generated some data for testing andwhen we tested our model on this data, accuracy dropped a lot.



	Model	Accuracy	Precision	Recall	F1_score
0	Decision Tree	0.445839	0.445839	0.445839	0.445839
1	Logistic Regression	0.682401	0.682401	0.682401	0.682401
2	Naive Bayes	0.641201	0.641201	0.641201	0.641201
3	KNN	0.609550	0.609550	0.609550	0.609550
4	SVC	0.594816	0.594816	0.594816	0.594816

Fig. 5 Testing on Test Data

Limitation

The models memorize the training data but fail on testing. There is overfitting.

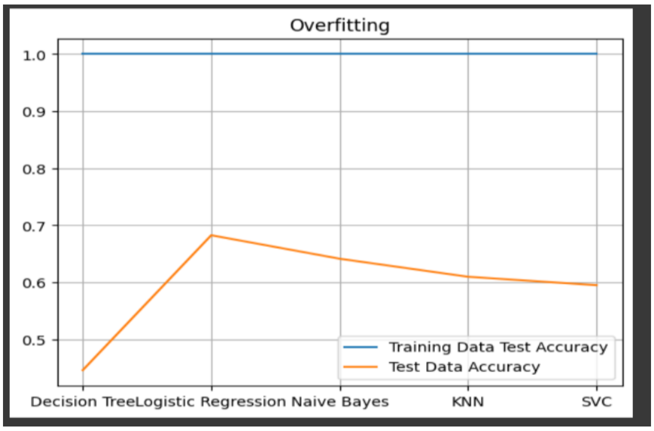
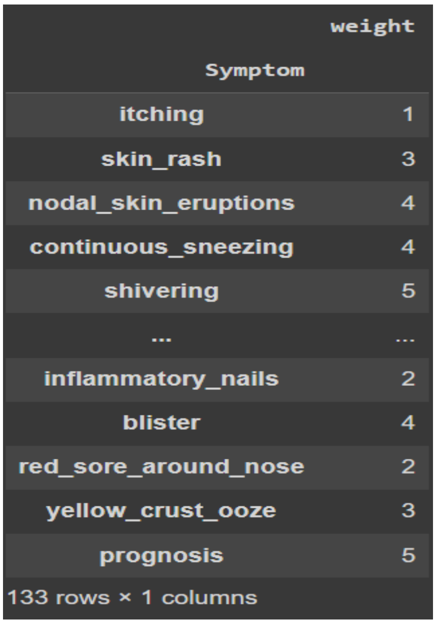


Fig. 6 Overfitting

Solution

Several symptoms have different degree of weightage so we used these weights as their values. Further we generated more of data from given dataset using their combinations and permutations to better train our model.



	weight
Symptom	
itching	1
skin_rash	3
nodal_skin_eruptions	4
continuous_sneezing	4
shivering	5
...	...
inflammatory_nails	2
blister	4
red_sore_around_nose	2
yellow_crust_ooze	3
prognosis	5

133 rows × 1 columns

Fig. 7 Symptom Weight Table

Model Trained on updated data provided good fit.

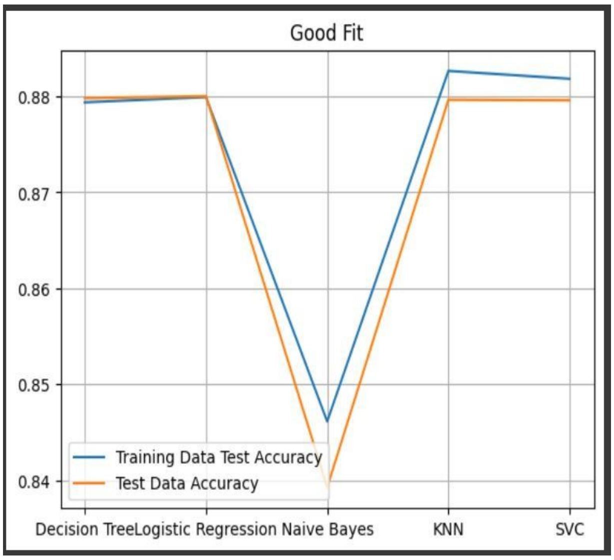
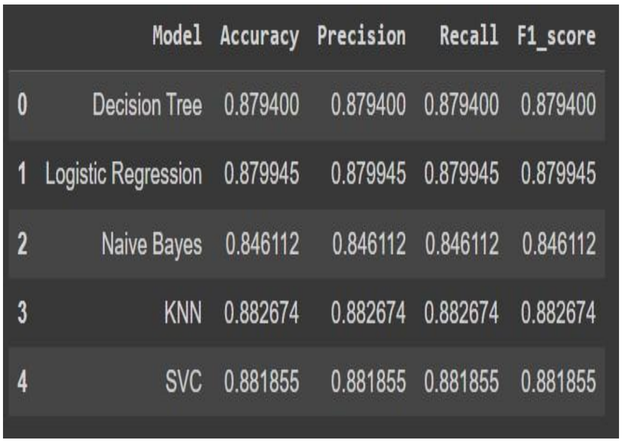


Fig. 8 Good Fit Model

Result: Final Accuracy Achieved



	Model	Accuracy	Precision	Recall	F1_score
0	Decision Tree	0.879400	0.879400	0.879400	0.879400
1	Logistic Regression	0.879945	0.879945	0.879945	0.879945
2	Naive Bayes	0.846112	0.846112	0.846112	0.846112
3	KNN	0.882674	0.882674	0.882674	0.882674
4	SVC	0.881855	0.881855	0.881855	0.881855

Fig. 9 Final Model Accuracy, Precision, Recall and F score

Analysis

Logistic Regression and Decision Tree both provide an accuracy score of 87%. Naïve Bayes has the lowest accuracy of 84%. KNN and SVC have the highest accuracy of 88%.

These results are close for training as well as for testing data. Thus, the models are neither overfitting nor underfitting. A good fit is achieved. This also suggests that by increasing our data we can further improve our models and increase accuracy, precision, recall and f score.

Conclusion

This research proved that there is high problem of overfitting and lack of sufficient amount of data to train models accurately and efficiently. We initially got 100% accuracy from all our models on the dataset. After trying to overcome the problem of overfitting, we achieved an accuracy of 84% to 88% from various models. KNN proved to be the most accurate among all the models.

References

1. S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9154130.

2. D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

3. Kanchanamala, P. & Das, Smritilekha & Neelima, G.. (2022). Symptoms-Based Disease Prediction Using Big data Analytics. 10.1007/978-981-16-8987-1_36.

4. Talasila, Bhanuteja & Kolli, Saipoornachand & Kumar, Kilaru & Anudeep, Poonati & Ashish, Chennupati. (2021). Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach. International Journal of Innovative Technology and Exploring Engineering. 10. 67-72. 10.35940/ijitee.I9364.0710921.

5. Keniya, Rinkal and Khakharia, Aman and Shah, Vruddhi and Gada, Vrushabh and Manjalkar, Ruchi and Thaker, Tirth and Warang, Mahesh and Mehendale, Ninad, Disease Prediction From Various Symptoms Using Machine Learning (July 27, 2020). Available at SSRN: https://ssrn.com/abstract=3661426 or http://dx.doi.org/10.2139/ssrn.3661426

6. P. Hamsagayathri and S. Vigneshwaran, "Symptoms Based Disease Prediction Using Machine Learning Techniques," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 747-752, doi: 10.1109/ICICV50876.2021.9388603.

7. Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

8. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

9. Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

10. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.

11. Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

12. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.