International Journal of Advance Research in Science and Engineering
Volume No. 13, Issue No. 06, June 2024
www.ijarse.com
IJARSE
ISSN 2319 - 8354

# ANALYSIS OF CANCER DISEASE BY USING DIFFERENT LEARNING ALGORITHMS

Dr. S. PAVAN, M.Tech,.,Ph.D,[1]
S VENU GOPAL REDDY[2], K SRI KRISHNA PRASAD[3], K SAISUSMITHA[4],
M THARUN CHANDRA[5], K TEJESH[6]
[1]Associate Professor, Department of Electronics and Communication Engineering, Tirumala Engineering College, [2] [3] [4] [5] [6] Department of Electronics and Communication Engineering, Sai Tirumala NVR Engineering College, JNTU KAKINADA, India

**ABSTRACT:**

**The early detection and prediction of cancer is crucial for effective intervention and improved patient outcomes. This study presents a comprehensive analysis of cancer prediction using various machine learning algorithms. The research utilizes a diverse dataset comprising clinical, genetic, and demographic information of patients to train and evaluate the performance of different algorithms. The primary objective is to develop a predictive model capable of accurately identifying cancer patients and distinguishing between different cancer types. This study explores the application of ML techniques in predicting cancer, focusing on model training, evaluation and feature selection . The experimental results demonstrate the efficiency of machine learning algorithms in predicting cancer disease, with certain models exhibiting superior performance compared to others. The findings highlight the importance of feature selection, data preprocessing, and algorithms selection in building robust and accurate predictive models for cancer diagnosis and prognosis.**

***Keywords – Student prediction, Datasets, feature Selection.***

## I . INTRODUCTION

The development of cancer typically involves a combination of genetic, environmental, and lifestyle factors. While certain genetic mutations may predispose individuals to cancer, environmental exposures such as tobacco smoke, ultraviolet radiation, and certain chemicals can also increase the risk. Additionally, lifestyle factors including diet,physical activity, and alcohol consumption can influence cancer risk.

Due to the development of advance healthcare systems, lots of patient data are nowadays available (i.e. Big Data in Electronic Health Record System) which can be used for designing predictive models for Cancer diseases. Data mining or machine learning is a discovery method for analyzing big data from an assorted perspective and encapsulating it into useful information. “Data Mining is a non-trivial extraction of implicit, previously unknown and potentially useful information about data”.

Nowadays, a huge amount of data pertaining to disease diagnosis, patients etc. aregenerated by healthcare industries. Data mining provides a number of techniques which discover hidden patterns or similarities from data.

Therefore, in this paper, a machine learning algorithm is proposed for the implementation of a cancer disease prediction system which was validated on two open access cancer disease prediction datasets. Data mining is the computer- based process extracting useful information from enormous sets of databases. Data mining is most helpful in an explorative analysis because of nontrivial information from large volumes of evidence.Medical data mining has great potential for exploring the cryptic patterns in the data sets ofthe clinical domain.

These patterns can be utilized for healthcare diagnosis. However, the available raw medical data are widely distributed, voluminous and heterogeneous in nature. This data needsto be collected in an organized form. This collected data can be then integrated to form a medical information system.

Data mining provides a user-oriented approach to novel and hidden patterns in the Data The data mining tools are useful for answering business questions and techniques for predicting the various diseases in the healthcare field.

International Journal of Advance Research in Science and Engineering
Volume No. 13, Issue No. 06, June 2024
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Disease prediction plays a significantrole in data mining. This paper analyzes the cancer disease predictions using classification algorithms. These invisible patterns can be utilized for health diagnosis in healthcare data.

Data mining technology affords an efficient approach to the latest and indefinite patterns in the data. The information which is identified can be used by the healthcare administrator get better services. Cancer disease was the most crucial reason for victims in the countries like India, United States.

In this project we are predicting the cancer disease using classification algorithms.Machine learning techniques like Classification algorithms such as Random Forest, Support vector machine, Decision tree and K-nearest neighbor are used to explore different kinds of cancer-based problems.

## II . LITERATURE REVIEW

The literature survey encompasses a diverse range of studies that delve into various aspects of predictive modeling,including algorithm selection, feature engineering, model evaluation, and practical applications in educational settings. These studies contribute to a comprehensive understanding of the capabilities and limitations of learning algorithms in predicting student success.

S. Acharya, P. S. Suresh, K. P. Rao, and S. V. R. Rao (2016) Introduced a "A survey of data mining techniques for cancer detection and prediction". This survey paper provides an overview of various data mining techniques, including machine learning algorithms, usedfor cancer detection and prediction based on different types of data such as gene expression data, medical imaging, and clinical data. R. K. Y. Lee, G. A. R. Yap, and R. Y. K. Kuo (2017) Introduced a "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study on prostate cancer". This paper presents a case study on predicting cancer susceptibility using single-nucleotide polymorphism (SNP) data, focusing on prostate cancer as an example. It discusses the challenges and potential approaches for using genetic data in cancer prediction. J. M. Kim, Y. Kim, and H. Kim (2019) Introduced a "Predicting lung cancer incidence from air pollution data using machine learning techniques". This paper investigates the use of machine learning techniques to predict lung cancer incidence based on air pollution data, highlighting the potential of environmental factors in cancer prediction models. P. Vijayarani and S. G. S. Raj (2021) Introduced a "Prediction of breast cancer using associative classification mining". This paper proposes an associative classification mining approach for predicting breast cancer based on patient data, demonstrating the application ofdata mining techniques in cancer prediction.

## III . EXISTING METHODS

In exploring the existing systems or approaches that contribute to academic success, it's important to recognize the multifaceted nature of this endeavor. One existing system involves traditional educational structures, including curriculum design, teaching methodologies, and assessment practices. These systems provide a framework for delivering education and evaluating student performance. However, they may vary widely across different educational institutions and contexts, leading to disparities in academic outcomes. Another existing system involves academic support services provided by schools, colleges, and universities. These services may include tutoring programs, writing centers, academic advising, counseling services, and peer mentoring initiatives. These support systems aim to address students' diverse learning needs, provide guidance and assistance, and promote academic success. However, their effectiveness may depend on factors such as accessibility, availability, and quality of support. Furthermore, technological advancements have led to the emergence of online learning platforms, digital resources, and educational technologies that supplement traditional educational systems. These tools offer opportunities for personalized learning, flexibility, and access to educational content. However, they also present challenges related to digital literacy, equitable access, and the quality of online learning experiences.

## IV . METHODOLOGY

proposed system involves implementing a comprehensive framework that integrates various strategies, support mechanisms, and interventions to support students' academic journey effectively. This

International Journal of Advance Research in Science and Engineering
Volume No. 13, Issue No. 06, June 2024
www.ijarse.com

SE
- 8354

proposed system encompasses the following key components: Implementing personalized support services tailored to students' individual needs and circumstances can help address inequities and provide targeted assistance where it's most needed. This may include offering academic tutoring, counseling, mentoring, and academic advising services that are accessible, culturally responsive, and inclusive.

## V. BLOCK DIAGRAM

The block of prediction of student performance using learning algorithms. It have five blocks those are data collection , data preprocessing, learning algorithms, New Data and predicted output.

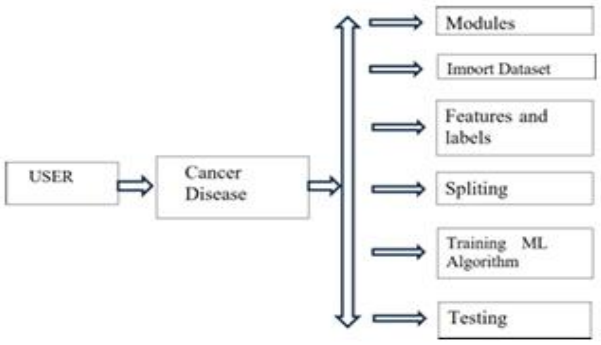



**Fig V : Use case Diagram**

a) Data collection

Data collection refers to the systematic process of gathering and measuring information on variables of interest in a structured and organized manner.

Selected features*:*

1) Academic Records
2) Attendance Patterns
3) Student Demographics
4) Extra Curricular Activities

b) Data preprocessing And Labelling

Data preprocessing involves cleaning and transforming raw data into a format suitable for analysis or model training. In this Data preprocessing we are using normalization method.

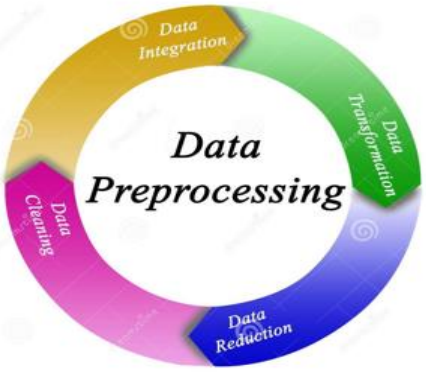



Figure b) : Data Processing And Labelling

C) Data Labelling

Data labeling refers to the process of assigning predefined categories or class labels to the data points in a dataset. In a classification task, labels represent the categories or classes that the model aims to predict.

d) Learning Algorithms

The learning algorithms are used to predict the student performance accurately.

Those are :

1) SVM ( Support Vector Machine ) Classifier

Support Vector Machines (SVM) are powerful supervised learning models used for both classification and regression tasks. The core idea behind SVM is to find the optimal hyperplane that separates data points of different classes with the maximum margin. In the case of linearly separable data, this hyperplane is a straight line in two dimensions or a hyperplane in higher dimensions. SVM works by transforming the input data into a higher-dimensional space using a kernel function, which enables it to find a linear decision boundary that wasn't possible in the original feature space.

One of the key strengths of SVM is its ability to work well in high-dimensional spaces, making it effective for tasks with a large number of features. Additionally, SVM is less affected by overfitting, especially in cases where the margin is properly regularized using techniques like soft margin SVM (allowing for some misclassification) or using appropriate kernel parameters.

International Journal of Advance Research in Science and Engineering
Volume No. 13, Issue No. 06, June 2024
www.ijarse.com
IJARSE
ISSN 2319 - 8354

2) Decision Tree Classification

A decision tree is a standalone machine learning model that operates by splitting the data into subsets based on features. Unlike Random Forest, it doesn't involve combining multiple trees. However, it can still introduce randomness by considering only a random subset of features at each split, reducing overfitting.

3) Random Forest Classification

Random Forest is an ensemble learning technique that combines multiple decision trees to make predictions. Each decision tree in the Random Forest is built independently and operates by splitting the data into subsets based on features. However, unlike traditional decision trees, Random Forest introduces randomness both in the selection of data points and features during the tree-building process. During training, each tree is constructed using a bootstrap sample (randomly selected with replacement) from the training data, which introduces diversity among the trees. Additionally, at each split in a tree, only a random subset of features is considered, further enhancing diversity and reducing overfitting.

4) K-Nearest Neighbor Classification

The k-Nearest Neighbors (KNN) algorithm is a versatile supervised learning method used for both classification and regression tasks. During training, it memorizes the entire training dataset. When predicting the label or value for a new data point, KNN calculates the distance between that point and all training points, typically using metrics like Euclidean distance or cosine similarity. It then selects the k nearest neighbors based on these distances. For classification, it employs a majority vote among the k neighbors to assign the class label to the new point, while for regression, it predicts the average value of the target variable. KNN's simplicity and effectiveness make it a popular choice, although optimal parameter selection and computational cost considerations are important for its performance.

Advantages

Firstly, these algorithms can analyze vast amounts of data, including historical academic records, demographic information, and even behavioral patterns, to identify key factors influencing student outcomes and it can detect early warning signs of academic challenges or potential dropout risks.

Applications

The applications of predicting student performance using learning algorithms are vast and have numerous implications for education and student support. Here are several specific applications:

1) Early Intervention and Support
2) Personalized Learning
3) Curriculum Enhancement
4) Institutional Decision-Making

## V . RESULTS

The results of predicting student performance using learning algorithms can have a significant impact on various aspects of education and student outcomes. Here are some key results and outcomes that can be achieved through the application of learning algorithms in predicting student performance. The accuracy table drawn below

Table -1:Confusion metrics of Support Vector Machine

| | CP | HP | TOTAL |
|---|---|---|---|
| CP | 78 | 0 | 78 |
| HP | 0 | 172 | 172 |
| TOTAL | 78 | 172 | 250 |

Table-2: Confusion metrics of Decision Tree

| | CP | HP | TOTAL |
|---|---|---|---|
| CP | 78 | 0 | 78 |
| HP | 0 | 172 | 172 |
| TOTAL | 78 | 172 | 250 |

International Journal of Advance Research in Science and Engineering
Volume No. 13, Issue No. 06, June 2024
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Table -3 :Confusion metrics of Random Forest Classifier

| | CP | HP | TOTAL |
|---|---|---|---|
| CP | 78 | 0 | 78 |
| HP | 0 | 172 | 172 |
| TOTAL | 78 | 172 | 250 |

Table -4:Confusion metrics of K- Nearest Neighbor

| | CP | HP | TOTAL |
|---|---|---|---|
| CP | 78 | 0 | 78 |
| HP | 0 | 172 | 172 |
| TOTAL | 78 | 172 | 250 |

Table -5: Accuracy Report of Learning Algorithms

| S. No | Algorithm | Accuracy | Precision | Recall |
|---|---|---|---|---|
| 1. | SVM | 0.98 | 1.0 | 1.0 |
| 2. | DTC | 1.0 | 0.98 | 1.0 |
| 3. | RFC | 1.0 | 0.99 | 1.0 |
| 4. | K-NN | 0.99 | 1.0 | 1.0 |

Figure -: V.1 First Patient Details

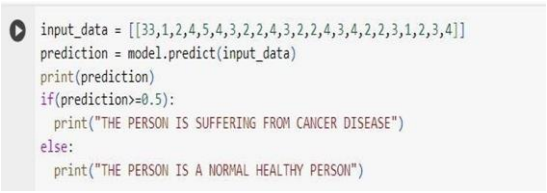
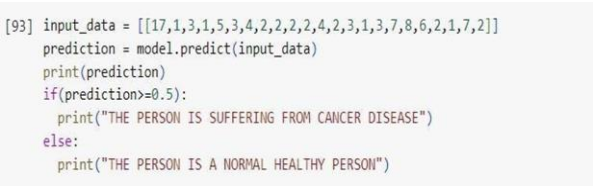



Figure -: V.2 Second Patient Details



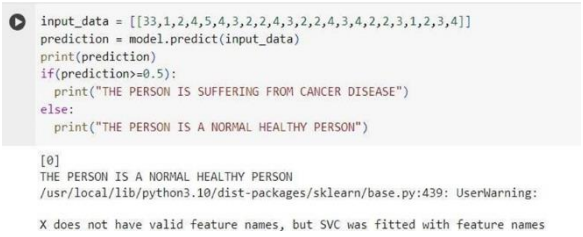




Figure -: V.3 First Patient is a Healthy Person

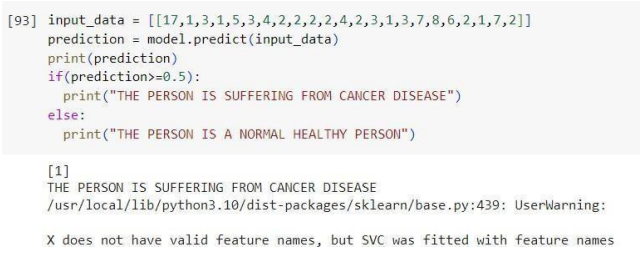



Figure V 4: Second Patient was a Cancered Patient

## VI . CONCLUSION

In conclusion, the prediction of student performance using learning algorithms offers several advantages. It enables early intervention and personalized learning, helping educators identify struggling students and tailor educational content to their needs. Learning algorithms also optimize resource allocation, support data-driven decision making, and provide continuous feedback for improvement. These applications demonstrate the potential for technology to positively impact education by enhancing student outcomes and promoting individualized learning experiences. And to see how learning algorithms can contribute to a more effective and inclusive educational system with the help of learning algorithms, educators can gain valuable insights into student performance. By analyzing data such as academic records, test scores, attendance, and even social factors, these algorithms can identify patterns and trends that may impact student success. This information can then be used to provide targeted interventions and support to students who may be at risk of falling behind. Additionally, learning algorithms can help in identifying areas where teaching methods and curriculum can be improved, leading to more effective instruction and better student outcomes. It's amazing how technology can assist educators in better understanding and supporting their students' academic journey

**Reference :**

1. Seyfried, T.N. and Shelton, L.M. Cancer as a metabolic disease. *Nutrition & metabolism*, *7*, pp.1-22, 2010.
2. Hornberg, J.J., Bruggeman, F.J., Westerhoff, H.V. and Lankelma, J. Cancer: a systems biology disease. *Biosystems*, *83*(2-3), pp.81-90, 2006.
3. Ravasco, P., Monteiro-Grillo, I., Vidal, P.M. and Camilo, M.E. Cancer: disease and nutrition are key determinants of patients' quality of life. *Supportive Care in Cancer*, *12*, pp.246-252, 2004.
4. Weedon, D.D., Shorter, R.G., Ilstrup, D.M., Huizenga, K.A. and Taylor, W.F. Crohn's disease and cancer. *New England Journal of Medicine*, *289*(21), pp.1099- 1103, 1973.
5. Xie, J. and Itzkowitz, S.H. Cancer in inflammatory bowel disease. *World journal of gastroenterology: WJG*, *14*(3), p.378, 2008.
6. Koene, R.J., Prizment, A.E., Blaes, A. and Konety, S.H.Shared risk factors in cardiovascular disease and cancer. *Circulation*, *133*(11), pp.1104-1114, 2016.
7. Folkman, J. and Kalluri, R. Cancer without disease. *Nature*, *427*(6977), pp.787-787, 2004.
8. Naccache, J.M., Gibiot, Q., Monnet, I., Antoine, M., Wislez, M., Chouaid, C. and Cadranel, J.Lung cancer and interstitial lung disease: a literature review. *Journal of thoracic disease*, *10*(6), p.3829, 2018.
9. Pallis, A.G. and Syrigos, K.N. Lung cancer in never smokers: disease characteristicsand risk factors. *Critical reviews in oncology/hematology*, *88*(3), pp.494-503, 2013.
10. Gridelli, C., Rossi, A., Carbone, D.P., Guarize, J., Karachaliou, N., Mok, T., Petrella,F., Spaggiari, L. and Rosell, R. Non-small-cell lung cancer. *Nature reviews Diseaseprimers*, *1*(1), pp.1-16, 2015.
11. Durham, A.L. and Adcock, I.M. The relationship between COPD and lung cancer. *Lung cancer*, *90*(2), pp.121-127,2015.
12. Schabath, M.B. and Cote, M.L.Cancer progress and priorities: lung cancer. *Cancer epidemiology, biomarkers & prevention*, *28*(10), pp.1563-1579,2019.
13. de-Torres, J.P., Wilson, D.O., Sanchez-Salcedo, P., Weissfeld, J.L., Berto, J., Campo, A., Alcaide, A.B., García-Granero, M., Celli, B.R. and Zulueta, J.J., 2015.
14. Woodman, C., Vundu, G., George, A. and Wilson, C.M., February. Applications and strategies in nanodiagnosis and nanotherapy in lung cancer. In *Seminars in cancer biology* (Vol. 69, pp. 349-364). Academic Press, 2021