IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 5, Issue 1, January 2025

Discovering Practical Insights with Supervised Machine Learning

Mr. Parvej Ikbal Shaikh[1] and Mrs. Shradha Balasaheb Linge[2]
Student, MIT Arts, Commerce & Science College, Alandi, India[1]
Assistant Professor, MIT Arts, Commerce & Science College, Alandi, India[2]
parvejshaikh7887@gmail.com and sblinge@mitacsc.ac.in

Abstract: *Machine Learning Algorithms are subset of artificial intelligence that enable computer to learn from data without being explicitly programmed. Supervised Learning Algorithm is a category of Machine Learning that uses labelled dataset to train algorithms to predict outcomes and recognize pattern. Types of Supervised Machine Learning Algorithm are Linear Regression, Softmax Regression, K-Nearest Neighbour, Decision Tree, Random Forest, Logistic Regression, Support Vector Machines, etc. We can enhance decision-making processes, optimize resource allocation and unlock new predictive modelling and automation. Cancer is a complex disease with diverse etiology and outcomes. Early detection and accurate prediction of cancer can reduce the risk and is helpful to cure disease by giving effective treatment and improved patient outcomes. In recent years, machine learning algorithms have shown promising results in cancer prediction by analysing various biomedical data sources such as genetic, clinical, and imaging data. This study aims to develop a predictive model for cancer occurrence using machine learning techniques.*

Keywords: Machine learning, Healthcare, Logistic Regression, Softmax Regression, Decision Tree, Breast Cancer

I. INTRODUCTION

Cancer continues to be one of the most significant health challenges worldwide, with its incidence steadily increasing over the years[5]. Early detection and accurate prediction of cancer risk are critical for timely intervention and improved patient outcomes. Traditional methods for cancer prediction rely heavily on manual interpretation of clinical and histopathological data, which can be time-consuming and subjective. Moreover, these methods often lack the precision and reliability needed for effective cancer screening and diagnosis.

The recent advancement of machine learning (ML) techniques has revolutionized the field of healthcare by providing powerful tools for data analysis and prediction. ML algorithms can effectively analyse large and complex datasets, including genetic profiles, clinical records, imaging data, and other biomarkers, to identify patterns and trends that may not be apparent to human observers[8].By leveraging these data-driven approaches, researchers and clinicians can develop predictive models for cancer occurrence that offer improved accuracy and efficiency compared to traditional methods.

There are four types of machine learning algorithm they are supervised, Unsupervised, Semi-supervised and Reinforcement Machine Learning.Supervised machine learning is a type of machine learning that represents the relationship between input variable x and output variable y. Types of supervised machine learning are softmax regression, linear regression, decision tree, random forest, logistic regression etc[2].

Softmax regression, also known as multinomial logistic regression, is a type of regression analysis used for classification tasks where the output variable consists of multiple classes. It is an extension of logistic regression, which is used for binary classification[9].In softmax regression, the goal is to predict the probability distribution of a categorical dependent variable (target variable) given a set of independent variables (features). It models the probabilities of the different classes as a multinomial distribution, which is suitable for classification problems with more than two classes.

Logistic regression is a type of supervised learning algorithm. It is commonly used for solving classification issues in machine learning [7].Logistic regression is a statistical model used for binary classification, where the outcome or

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 5, Issue 1, January 2025

dependent variable is categorical and has only two possible values, often represented as 0 and 1. It's named "regression" because it's based on the logistic function, also known as the sigmoid function, which is used to model the probability that a given input belongs to one of the two classes. The vectorize form of Logistic regression model that estimate the probabilities is $h_\Theta(x)=ð(x^T\Theta)$[9].

A decision tree is a popular machine learning algorithm used for both classification and regression tasks.[2] It operates by recursively partitioning the input space into regions, with each partition corresponding to a decision based on the value of one of the input features. At each step, the algorithm selects the feature that best splits the data into homogeneous subsets according to some criterion and information gain.

Furthermore, we will explore the potential clinical implications of our predictive model, including its utility in early detection, risk stratification, and personalized treatment planning. By providing clinicians with a reliable tool for cancer prediction, we aim to improve patient outcomes and reduce the burden of this devastating disease.In this project we study underscores the transformative potential of ML algorithms in cancer prediction and highlights the importance of leveraging diverse data sources for improved accuracy and reliability[10].By harnessing the power of data-driven approaches, we can advance our understanding of cancer biology and pave the way for more effective strategies for cancer prevention, diagnosis, and treatment[4].

II. LITERATURE REVIEW

The abstract discusses the use of decision trees to represent unconscious controllers and highlights the advantages of decision trees in terms of size and interpretation compared to other methods such as the table view and binary decision graph. This paper introduces the dtControl toolkit, which supports a variety of decision tree learning algorithms, including two new technologies that improve decision trees. Evaluation of these techniques across ten studies demonstrates the effectiveness and simplicity of decision trees, especially when using methods that generate very large trees[1].

Decision tree classifiers are widely used in many fields such as medical diagnosis, text classification, image recognition. Many researchers are looking for ways to improve decision tree algorithms, focusing on their implementation and accuracy. This review examines various studies, benchmark algorithms, datasets, and results to demonstrate the effectiveness of decision trees in task classification. This analysis shows the evolution of decision trees for different data types and suggests further advances to improve their accuracy and efficiency in real-world situations[2].

The contents show the importance of softmax function to build classifiers in LLM and to provide assistance in training models. Also, the work of Cohen, Lee, Song and Hom is mentioned and the role of softmax in convex optimization, especially in stability control and success is highlighted. The proposed softmax regression problem solved by greedy algorithm is presented as a new contribution, and theoretical results support the effectiveness of this method in training softmax function[3].

An overview of breast cancer incidence and mortality in the United States, noting that incidence has been steadily increasing but mortality has been decreasing since 1989. Again, survival is lower for women. This gap is particularly prevalent in younger women and for breast cancer. The results suggest that closing this disparity requires improving access to screening and quality care, particularly through Medicaid expansion and community engagement[4].

Breast cancer is the most common cancer worldwide and is associated with multiple risk factors, including genetic predisposition. The disease is highly prevalent and treatment strategies are based on molecular features such as HER2 activation, hormone receptors (ER and PR), genetic mutations (e.g. BRCA1/2, PIK3CA) and immune response markers (TIL, PD-L1). Early breast cancer can often be treated with surgery and radiation therapy, and if necessary, chemotherapy. In severe cases, which are often incurable, medical therapy remains the mainstay of treatment[5].

This abstract discusses the use of ridge regression in logistic regression to improve estimators and reduce estimation error by focusing on the selection of parameters from competition. Three error measures are examined: standard deviation, squared error, and negative log-likelihood. This method is used to estimate two-year survival of ovarian cancer patients and overcomes the problem of overfitting due to high covariate-to-observation ratios. The model demonstrates the clinical utility of ridge regression by improving interpretation and estimation accuracy by parameterizing the parameters[7].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 5, Issue 1, January 2025

Machine learning (ML) literature frequently explores its important role in many areas such as data mining, image processing, and predictive analytics. Research shows how machine learning algorithms can perform tasks by learning from data, making them important in applications such as web search. Researchers emphasize machine learning's ability to process large amounts of data, improve decision-making, and adapt to new information. Future research should focus on improving algorithm efficiency, explanation, and applications in new areas such as intelligence-driven automation and personalized services[8].

Logistic regression is a classification technique used in machine learning that uses a linear model developed by a logistic function to divide data into two groups. The logistic function is well suited to binary distribution because it can represent true values with probability 0 and 1, thus supporting good decision making. The logistic regression literature explores its theoretical basis, including the basis of the logistic function, and tests its computational effectiveness with theoretical analysis and experimental evidence[9].

The threat landscape of cyberattacks continues to expand, prompting companies to develop advanced intelligence tools to analyze data security and investigate threats. Machine learning (ML) plays a key role here to improve security by reducing high maintenance costs and vulnerabilities in security service centers (SOCs). Logistic regression is a widely used statistical approach, but it can be problematic in large-scale applications. Choosing the right machine learning methodology remains a significant challenge in identifying cyber threats[10].

III. OBJECTIVE OF STUDY

- Understanding Core Concepts: Understanding Core Concepts of supervised machine learning, including algorithms, techniques, and methodologies.
- Exploration of Applications: Identify and explore various real-world applications of supervised machine learning across domains of healthcare.
- Performance Evaluation: Evaluate the performance metrics and criteria for assessing the effectiveness of supervised learning models.
- Model Selection and Tuning: Compare and contrast different supervised learning algorithms and hyper-parameter tuning methods to achieve optimal model performance.
- Scalability and Efficiency: Explore strategies for scaling up supervised learning algorithms to handle large datasets efficiently.
- Risk Stratification: Create models to stratify individuals based on their risk of developing cancer, allowing for targeted interventions and personalized treatment plans.
- Prediction of Subtypes: Develop models capable of predicting cancer subtypes or specific characteristics (e.g., tumor aggressiveness) to guide treatment decisions.
- Integration of Multiple Data Sources: Explore methods for integrating heterogeneous data sources to improve prediction accuracy.
- Validation and Generalization: Validate the performance of the developed models on independent datasets to ensure their generalizability and reliability in real-world settings.

IV. METHODOLOGY

The Machine learning is a subset if artificial intelligence. The ML algorithms is use for study is to develop a predictive model for cancer occurrence. For this purpose we use the Breast Cancer data set[6].

The Machine Learning Algorithm such as logistic regression, softmax regression and decision tree are use to compare prediction of cancer. Here by using python language we fit the Machine learning algorithms. We need matplot, seaborn, sklearn and pandas etc libraries to import dataset and to fit various algorithms.

Process of the fitting model:

We get data set from Kaggle and import it in Python using pandas library. We use Breast Cancer Dataset for prediction using Machine Learning Algorithm. This data set is created byDr. William H. Wolberg, General Surgery Dept., University of Wisconsin; W. Nick Street, Computer Sciences Dept., University of Wisconsin and Olvi L.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 5, Issue 1, January 2025

Mangasarian, Computer Sciences Dept., University of Wisconsin[6]. The number of instances in the data set is 569 and number of attributes are 32 (ID, diagnosis, 30 real-valued input features) from which we use only six attributes they are radius (mean of distances from centre to points on the perimeter, texture (standard deviation of gray-scale values, perimeter, area, smoothness (local variation in radius lengths) and diagnosis. We have to detect whether the cancer is M=malignant means cancerous which is denoted by 0 in or B=benign means not cancerous which is denoted by 1 in the dataset which is modified for prediction purpose[6].

```
import pandas as pd
cancer=pd.read_csv("Breast_cancer_data.csv")
cancer
```

	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
0	17.99	10.38	122.80	1001.0	0.11840	0
1	20.57	17.77	132.90	1326.0	0.08474	0
2	19.69	21.25	130.00	1203.0	0.10960	0
3	11.42	20.38	77.58	386.1	0.14250	0
4	20.29	14.34	135.10	1297.0	0.10030	0
...	...	...	...	...	...	...
564	21.56	22.39	142.00	1479.0	0.11100	0
565	20.13	28.25	131.20	1261.0	0.09780	0
566	16.60	28.08	108.30	858.1	0.08455	0
567	20.60	29.33	140.10	1265.0	0.11780	0
568	7.76	24.54	47.92	181.0	0.05263	1

569 rows × 6 columns

Our data set contains 569 rows and 6 columns in which radius, texture, perimeter, area, smoothness and diagnosis are features.

To diagnosis the counts of malignant cases which is indicated by 0 and benign cases which is indicated by 1 in our data set we use the function.

In our data set we have 357 malignant and 212 benign cases which are predicted using various parameters.

```
cancer["diagnosis"].value_counts()
```

```
diagnosis
1    357
0    212
Name: count, dtype: int64
```

Using seaborn library we draw the plot of malignant and benign cases.

The above plot shows the graphical view of diagnosis to count. The blue and orange region shows benign and malignant case respectively. Clearly we see that benign cases are less as compare to malignant cases.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 5, Issue 1, January 2025

```
cancer.describe()
```

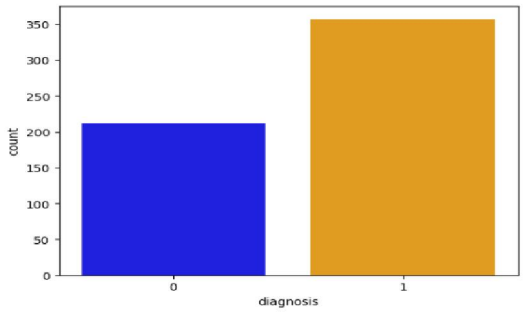
	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.627417
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.483918
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.000000
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.000000
50%	13.370000	18.840000	86.240000	551.100000	0.095870	1.000000
75%	15.780000	21.800000	104.100000	782.700000	0.105300	1.000000
max	28.110000	39.280000	188.500000	2501.000000	0.163400	1.000000

The describe() function gives overall description of data set.
We count the statistical summary such as mean, standard deviation, minimum, maximum, quartiles of our data respective to each parameter.
To fit the model we consider x as an input features and y as a predicted value for inputs. In this case diagnosis is dependent variable and other parameters are independent variable by using we will able to predict dependent variable.

```
x=cancer[["mean_radius","mean_texture","mean_perimeter","mean_area","mean_smoothness"]].values
y=cancer[["diagnosis"]].values
```

Softmax Regression is a generalization of Logistic Regression to the case where we want to handle multiple classes[3].
Now our we are ready to fit various Machine Learning model. First we will fit Softmax Regression for this we need 'LogisticRegression' class from scikit-learn library. The fit() function is used to fit independent variable and dependent variable.

```
import seaborn as sns
sns.countplot(x="diagnosis", data=cancer, palette=['blue', 'orange'])
C:\Users\lenovo 5fin\AppData\Local\Temp\ipykernel_18876\2157231811.py:1: FutureWarning:
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. A
  sns.countplot(x="diagnosis", data=cancer, palette=['blue', 'orange'])
<Axes: xlabel='diagnosis', ylabel='count'>
```



Logistic Regression is a type of supervised machine learning techniques which is used for predicting the categorical dependent variable using the set of independent variable.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 5, Issue 1, January 2025

We fit Logistic Regression which is approximately similar to softmax regression i.e. softmax is advancement of logistic regression.

Decision Tree is a supervised machine learning algorithm use for classification and prediction but, mostly used for classification.

```
from sklearn.linear_model import LogisticRegression
model1=LogisticRegression()
model1.fit(x,y)
```

```
C:\Users\lenovo 5fin\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\util
s\validation.py:1300: DataConversionWarning: A column-vector y was passed when a 1d array w
as expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

```
LogisticRegression
LogisticRegression()
```

The third algorithm we use is Decision Tree. To fit this model at first we need to separate diagnosis part named target and remaining attributes named data. After that we split the dataset into test and train set by using 'model_selection' from sklearn library and at last we fit the model.

```
target=cancer['diagnosis']
```

```
data=cancer.drop(['diagnosis'],axis='columns')
```

```
from sklearn.model_selection import train_test_split
x_train , x_test , y_train , y_test = train_test_split(data,target,stratify=target,random_state=0)
```

```
from sklearn.tree import DecisionTreeClassifier
tree=DecisionTreeClassifier(random_state=0)
tree.fit(x_train,y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=0)
```

```
model.predict([[19.99,115.38,152.80,1111.01,1.11840]])
```

```
array([0], dtype=int64)
```

```
model.predict([[7.59,14.6,45.36,195.2,0.0452]])
```

```
array([1], dtype=int64)
```

```
model.predict([[8.23,25.0,13.56,225.7,2.99]])
```

```
array([1], dtype=int64)
```

```
model.predict([[15.71,55.45,122.25,951.0,1.0000]])
```

```
array([0], dtype=int64)
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 5, Issue 1, January 2025

V. RESULT

We fit various ML Algorithms for prediction to evaluate the performance score() function is used We predict thescore between independent variable and dependent variable by using the model which we will fit

Softmax regression:

```
model.score(x,y)
```

```
0.9173989455184535
```

Softmax Regression model give 91.73% accurate result. Now to predict whether the tumor is malignant or benign, we give some input features.

```
tree.score(x_test,y_test)
```

```
0.9020979020979021
```

```
tree.score(x_train,y_train)
```

```
1.0
```

Logistic regression:

Logistic Regression model give 91.21% accurate result. Now to predict whether the tumor is malignant or benign, we give some input features.

```
model.predict([[19.99,115.38,152.80,1111.01,1.11840]])
```

```
array([0], dtype=int64)
```

```
model.predict([[7.59,14.6,45.36,195.2,0.0452]])
```

```
array([1], dtype=int64)
```

```
model.predict([[8.23,25.0,13.56,225.7,2.99]])
```

```
array([1], dtype=int64)
```

```
model.predict([[15.71,55.45,122.25,951.0,1.0000]])
```

```
array([0], dtype=int64)
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 5, Issue 1, January 2025

Decision Tree:

Decision Tree model give 90.21% accurate result and if we train the model it will give 100% accurate result. Now to predict whether the tumor is malignant or benign, we give some input features.

Over all the predictive parameters which we have at first and fourth is predicted as 0 i.e. cancerous and middle two are i.e. not cancerous .

VI. CONCLUSION

In this project we successfully apply Softmax Regression, Logistic Regression and Decision tree model for classification of Breast cancer data to predict the tumor is malignant or benign. Among this three model Softmax Regression model gives highest percentage of accuracy. Required measures are given and the model will predict in the form of 0 or 1 .If we train the model then it is 99% accurate. There may be 1% chance that the prediction is not accurate as we are dealing with Machines.

REFERENCES

[1]. Ashok, P., Jackermeier, M., Jagtap, P., KÅetínský, J., Weininger, M., & Zamani, M. (2020, April 22). DtControl: Decision tree learning algorithms for controller representation. HSCC 2020 - Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control ,Part of CPS-IoT Week.

[2]. Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20–28.

[3]. Deng, Y., Li, Z., & Song, Z. (2023). *Attention Scheme Inspired Softmax Regression*.

[4]. Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., Jemal, A., & Siegel, R. L. (2022). Breast Cancer Statistics, 2022. CA: A Cancer Journal for Clinicians, 72(6), 524–541.

[5]. Hong, R., & Xu, B. (2022). Breast cancer: an up-to-date review and future perspectives. In Cancer Communications (Vol. 42, Issue 10, pp. 913–936). John Wiley and Sons Inc.

[6]. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

[7]. LeCessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc Ser C (Appl Stat). 1992;41(1):191–201.

[8]. Mahesh, B. (2020). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR), 9(1), 381–386.

[9]. Zaidi, A., & al Luhayb, A. S. M. (2023). Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression. Mathematical Problems in Engineering, 2023, 1–11.

[10]. Zhu, J. J., Yang, M., & Ren, Z. J. (2023). Machine Learning in Environmental Research: Common Pitfalls and Best Practices. In Environmental Science and Technology (Vol. 57, Issue 46, pp. 17671–17689). American Chemical Society