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 4, Issue 1, July 2024

# Unlocking the Power of Supervised Machine Learning: Applications and Insights

**Supriya G. Naikwadi[1], Rutuja V. Khairnar[2], Dr. Sharmila S. More[3]**
Research Scholar/Student, FYMSC IMCA, Department of IMCA[1]
Research Scholar/Student, FYMSC IMCA, Department of IMCA[2]
Assistant Professor, Department of Science and Computer Science[3]
MIT ACSC Alandi (D), Pune, Maharashtra, India
supriyagopinathnaikwadi@mitacsc.edu.in, rutujavilaskhairnar@mitacsc.edu.in
ssmore@mitacsc.edu.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. 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. 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.

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. 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.

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 4, Issue 1, July 2024

Logistic regression is a type of supervised learning algorithm. It is commonly used for solving classification issues in machine learning [2]. Logistic regression is a statistical model used for binary classification, where the outcome or 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 $h_{\Theta}(x)=ð(x^{T}\Theta)$.

A decision tree is a popular machine learning algorithm used for both classification and regression tasks. 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. 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.

## II. 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.

## III. METHODOLOGY

Supervised learning is a machine learning technique that uses labeled data to predict output based on patterns learned from training data. It has been around since the early days of artificial intelligence research, with the perceptron being one of its most influential applications. Supervised learning is crucial in various fields due to its ability to make accurate predictions and classify data. Common applications include classification, regression, natural language processing (NLP), computer vision, bioinformatics, and finance. Examples include classifying emails, predicting house prices, predicting disease symptoms, and estimating product demand. Supervised learning has also been used in natural language processing, computer vision, bioinformatics, and finance for tasks such as credit scoring, fraud detection, and risk management.

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 4, Issue 1, July 2024

Supervised machine learning is a subfield of artificial intelligence (AI) where algorithms learn from labeled data to make predictions or decisions. In this paradigm, the algorithm is trained on a dataset where each example is paired with the correct output. Through this training process, the algorithm learns the relationship between the input features and the corresponding output labels, enabling it to generalize and make predictions on unseen data.

**Historical Overview:**

Supervised learning has roots dating back to the 1950s and 1960s when pioneers in the field of AI, such as Arthur Samuel, began experimenting with algorithms that could learn from data. One of the earliest applications was in the development of the perceptron, a simple algorithm capable of learning binary classifiers. Over the decades, supervised learning has evolved significantly, driven by advancements in computing power, data availability, and algorithmic innovation. Notable milestones include the development of decision trees, neural networks, support vector machines, and ensemble methods, among others.

**Importance and Applications:**

Supervised learning is of paramount importance in various domains due to its versatility and effectiveness in solving a wide range of problems. Some key reasons for its importance include:

- Predictive Analytics: Supervised learning enables businesses to forecast trends, make predictions, and optimize decision-making processes. Applications include sales forecasting, demand prediction, risk assessment, and customer churn prediction.
- Classification and Recognition: Supervised learning algorithms are widely used for tasks such as image classification, object recognition, speech recognition, and sentiment analysis. These applications are pervasive in industries such as healthcare, finance, marketing, and entertainment.
- Personalization and Recommendation: Supervised learning powers recommendation systems that provide personalized suggestions to users based on their preferences and behavior. Examples include movie recommendations on streaming platforms, product recommendations on e-commerce websites, and personalized content delivery on social media.
- Natural Language Processing (NLP): In NLP, supervised learning techniques are used for tasks such as text classification, named entity recognition, sentiment analysis, and machine translation. These applications have implications in areas such as customer service, content moderation, and language understanding.
- Fraud Detection and Anomaly Detection: Supervised learning algorithms play a crucial role in detecting fraudulent activities and anomalies in various domains, including finance, cybersecurity, and healthcare. By learning from past examples of fraudulent behavior, these algorithms can identify suspicious patterns and flag potentially fraudulent transactions or activities.

**Basic Concepts and Terminology:**

- Features: Features, also known as input variables or predictors, are the measurable characteristics or attributes of the data that are used to make predictions.
- Target Variable: Also referred to as the output variable or response variable, the target variable is the variable that we are trying to predict or model in supervised learning.
- Training Data: Training data is the labeled dataset used to train the supervised learning algorithm. It consists of input-output pairs, where the inputs are the features and the outputs are the corresponding target variables.
- Model: A model is a mathematical representation of the relationship between the input features and the target variable. The goal of supervised learning is to train a model that can accurately predict the target variable for new, unseen data.
- Prediction: Prediction is the process of using a trained model to make forecasts or decisions about the target variable for new instances of data.

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 4, Issue 1, July 2024

**Types of Supervised Learning Algorithms:**

- Classification: Classification algorithms are used when the target variable is categorical or qualitative. The goal is to assign each instance of data to one of a finite number of classes or categories. Examples of classification algorithms include logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks.
- Regression: Regression algorithms are used when the target variable is continuous or quantitative. The goal is to predict a numerical value based on the input features. Examples of regression algorithms include linear regression, polynomial regression, decision trees, support vector regression (SVR), and neural networks.
- Instance-Based Learning: Instance-based learning, also known as lazy learning, involves storing the entire training dataset and making predictions based on the similarity between new instances and existing instances in the training data. K-nearest neighbors (KNN) is a popular instance-based learning algorithm.
- Ensemble Learning: Ensemble learning involves combining multiple individual models to improve prediction accuracy and robustness. Common ensemble learning techniques include bagging (e.g., random forests), boosting (e.g., AdaBoost, gradient boosting), and stacking.
- Deep Learning: Deep learning is a subset of machine learning that uses neural networks with multiple layers to learn complex patterns in data. Deep learning algorithms have achieved remarkable success in tasks such as image recognition, natural language processing, and speech recognition.

**Classification Methods**

**Logistic Regression:**

- Logistic regression is a linear classification algorithm used for binary classification tasks, where the target variable has two possible outcomes (e.g., yes/no, true/false).
- The logistic regression is a classification algorithm rather than a regression algorithm. It models the probability that a given instance belongs to a particular class using the logistic function (also known as the sigmoid function).
- The logistic function maps any real-valued input to a value between 0 and 1, representing the probability of the instance belonging to the positive class.
- During training, logistic regression estimates the parameters of the model by maximizing the likelihood of the observed data.
- Logistic regression is relatively simple, interpretable, and efficient, making it a popular choice for binary classification tasks in various domains.

**Decision Trees:**

- Decision trees are versatile classification algorithms that recursively partition the feature space into regions, with each region corresponding to a specific class label.
- The decision tree structure consists of nodes, branches, and leaves. Nodes represent features, branches represent decision rules based on feature values, and leaves represent class labels.
- Decision trees are constructed iteratively by selecting the best feature to split the data at each node, typically based on criteria such as Gini impurity or information gain.
- Decision trees can handle both categorical and numerical features, and they are capable of capturing complex decision boundaries.
- The decision trees are prone to over-fitting, especially when the tree depth is not properly controlled. Techniques such as pruning, limiting tree depth, and using ensemble methods like random forests can help mitigate over-fitting.

**Softmax Regression (Multinomial Logistic Regression):**

- Softmax regression is an extension of logistic regression that is used for multi-class classification tasks, where the target variable has more than two possible outcomes.

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 4, Issue 1, July 2024

- Softmax regression models the probabilities of each class using the softmax function, which generalizes the logistic function to multiple classes.
- The softmax function computes the probability distribution over all classes, ensuring that the predicted probabilities sum to one.
- During training, softmax regression learns the parameters (weights and biases) of the model by minimizing the cross-entropy loss between the predicted probabilities and the true class labels.
- Softmax regression is widely used in scenarios where instances can belong to multiple classes simultaneously, such as image classification, natural language processing, and sentiment analysis.

**Evaluation Metrics for Supervised Learning**

**Classification Report:**

- A classification report provides a comprehensive summary of various evaluation metrics, including precision, recall, F1-score, and support (the number of instances for each class).
- It allows for a quick comparison of the performance of different classes in a multi-class classification problem.

**Accuracy:**

- Accuracy is one of the most straightforward evaluation metrics for classification tasks. It measures the ratio of correctly predicted instances to the total number of instances in the dataset.
- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- However, accuracy may not be the most appropriate metric when dealing with imbalanced datasets, where one class dominates the others.

**Precision:**

- Precision measures the proportion of true positive predictions among all positive predictions made by the classifier.
- Precision = TP / (TP + FP)
- Precision focuses on the accuracy of positive predictions, making it useful when the cost of false positives is high.

**Recall (Sensitivity):**

- Recall measures the proportion of true positive predictions among all actual positive instances in the dataset.
- Recall = TP / (TP + FN)
- Recall is particularly important when the cost of false negatives is high, as it reflects the classifier's ability to detect all positive instances.

**F1-score:**

- The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, making it a useful metric for imbalanced datasets.
- F1-score = 2 * (Precision * Recall) / (Precision + Recall)
- The F1-score ranges from 0 to 1, with higher values indicating better performance.

**Confusion Matrix:**

- A confusion matrix is a table that visualizes the performance of a classification model by summarizing the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.
- It provides insights into the types of errors made by the classifier and can help identify areas for improvement.
- From the confusion matrix, various metrics such as accuracy, precision, recall, and F1-score can be calculated.

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[3].

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 4, Issue 1, July 2024

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. Mangasarian, Computer Sciences Dept., University of Wisconsin[3]. 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[3].

```
import pandas as pd
cancer=pd.read_csv("D:\\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.

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 4, Issue 1, July 2024

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

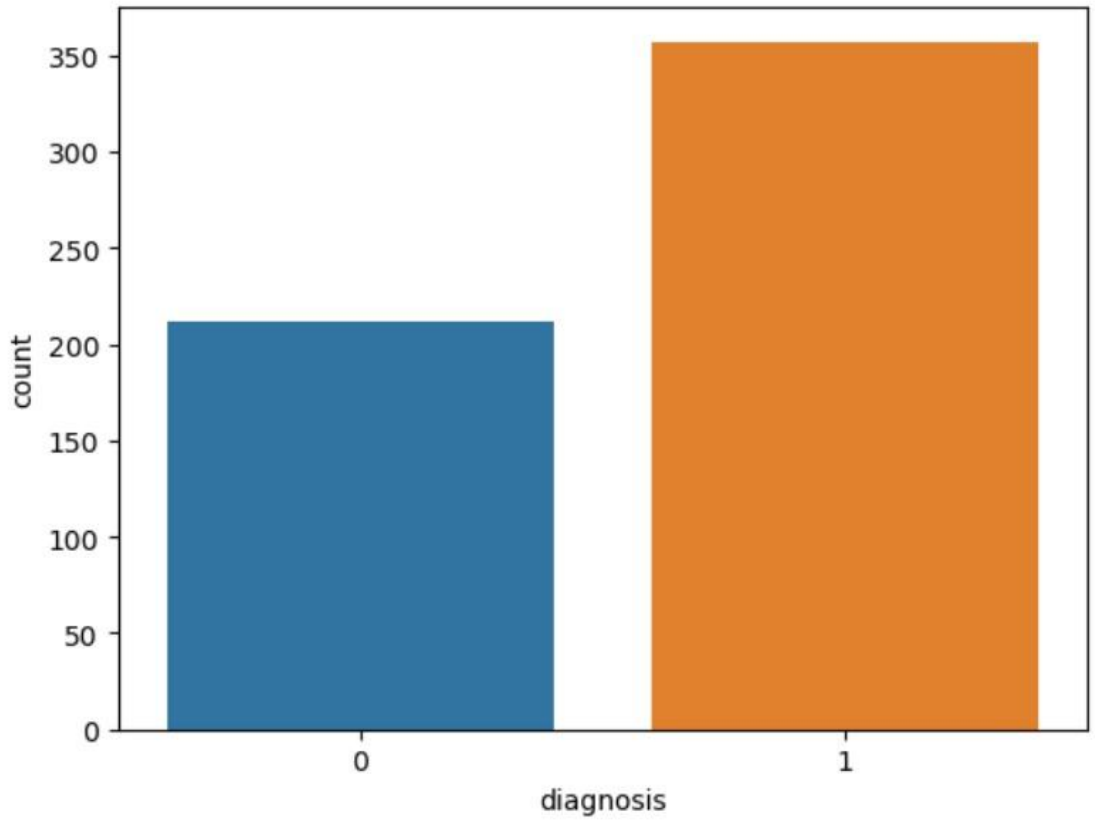
```
1    357
0    212
Name: diagnosis, dtype: int64
```

In our data set we have 357 malignant and 212 benign cases which are predicted using various parameters.
Using seaborn library we draw the plot of malignant and benign cases.

```
import seaborn as sns
```

```
sns.countplot(x="diagnosis",data=cancer)
```

```
<AxesSubplot:xlabel='diagnosis', ylabel='count'>
```




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.
The describe() function gives overall description of data set.

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 4, Issue 1, July 2024

```
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 |

We count the statistical summary such as mean, standard deviation, minimum, maximum, quartiles of our data respective to each parameter.

## IV. RESULT

<table>
<tr><th>Method</th><th>Types</th><th>Description</th></tr>
<tr><td>Fitting of model</td><td>a)Softmax Regression</td><td>Softmax Regression is a generalization of Logistic Regression to the case where we want to handle multiple classes[1]. 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.<br><pre>from sklearn.linear_model import LogisticRegression
model=LogisticRegression(multi_class="multinomial",solver="lbfgs",C=10)
model.fit(x,y.ravel())</pre><pre>LogisticRegression(C=10, multi_class='multinomial')</pre></td></tr>
<tr><td></td><td>b) Logistic Regression</td><td>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. We fit Logistic Regression which is approximately similar to softmax regression i.e. softmax is advancement of logistic regression.<br><pre>from sklearn.linear_model import LogisticRegression
model1=LogisticRegression()
model1.fit(x,y)</pre><pre>C:\ProgramData\Anaconda3\lib\site-packages\sklearn\utils\validation.py:993: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model_logistic.py:814: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(</pre><pre>LogisticRegression()</pre></td></tr>
<tr><td></td><td>c) Decision Tree</td><td>Decision Tree is a supervised machine learning algorithm use for classification and prediction but, mostly used for classification.<br>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</td></tr>
</table>

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 4, Issue 1, July 2024

| | | |
|---|---|---|
| | | library and at last we fit the model.<br>`target=cancer['diagnosis']`<br>`data=cancer.drop(['diagnosis'],axis='columns')`<br>`from sklearn.model_selection import train_test_split`<br>`x_train , x_test , y_train , y_test=train_test_split(data,target,stratify=target,random_state=0)`<br>`from sklearn.tree import DecisionTreeClassifier`<br>`tree=DecisionTreeClassifier(random_state=0)`<br>`tree.fit(x_train,y_train)`<br>`DecisionTreeClassifier(random_state=0)` |
| Score | a)Softmax Regression | Softmax Regression model give 91.74% accurate result.<br>`model.score(x,y)`<br>`0.9173989455184535` |
| | b) Logistic Regression | Logistic Regression model give 91.21% accurate result.<br>`model1.score(x,y)`<br>`0.9121265377855887` |
| | c) Decision Tree | Decision Tree model give 90.21% accurate result and if we train the model it will give 100% accurate result.<br>`tree.score(x_test,y_test)`<br>`0.9020979020979021`<br>`tree.score(x_train,y_train)`<br>`1.0` |
| Prediction(The 1 shows benign means non cancerous and 0 denotes malignant means cancerous | a)Softmax Regression | `model.predict([[19.99,115.38,152.80,1111.01,1.11840]])`<br>`array([0], dtype=int64)`<br>`model.predict([[7.59,14.6,45.36,195.2,0.0452]])`<br>`array([1], dtype=int64)`<br>`model.predict([[8.23,25.0,13.56,225.7,2.99]])`<br>`array([1], dtype=int64)`<br>`model.predict([[15.71,55.45,122.25,951.0,1.0000]])`<br>`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 4, Issue 1, July 2024

| | | |
|---|---|---|
| | b) Logistic Regression | `model1.predict([[19.99,115.38,152.80,1111.01,1.11840]])`<br>`array([0], dtype=int64)`<br>`model1.predict([[7.59,14.6,45.36,195.2,0.0452]])`<br>`array([1], dtype=int64)`<br>`model1.predict([[8.23,25.0,13.56,225.7,2.99]])`<br>`array([1], dtype=int64)`<br>`model1.predict([[15.71,55.45,122.25,951.0,1.0000]])`<br>`array([0], dtype=int64)` |
| | c) Decision Tree | `model1.predict([[19.99,115.38,152.80,1111.01,1.11840]])`<br>`array([0], dtype=int64)`<br>`model1.predict([[7.59,14.6,45.36,195.2,0.0452]])`<br>`array([1], dtype=int64)`<br>`model1.predict([[8.23,25.0,13.56,225.7,2.99]])`<br>`array([1], dtype=int64)`<br>`model1.predict([[15.71,55.45,122.25,951.0,1.0000]])`<br>`array([0], dtype=int64)` |
| Confusion Matrix | a)Softmax Regression | `print(metrics.confusion_matrix(expected,predected))`<br>`[[183 29]`<br>`[ 18 339]]` |
| | b) Logistic Regression | `print(metrics.confusion_matrix(expected,predected))`<br>`[[182 30]`<br>`[ 20 337]]` |

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 4, Issue 1, July 2024

| | | |
|---|---|---|
| | c) Decision Tree | `print(metrics.confusion_matrix(expected,predected))`<br>`[[206 6]`<br>`[ 8 349]]` |
| Classification Report | a)Softmax Regression | `from sklearn import metrics`<br>`print(metrics.classification_report(expected,predected))`<br>precision recall f1-score support<br>0 0.91 0.86 0.89 212<br>1 0.92 0.95 0.94 357<br>accuracy 0.92 569<br>macro avg 0.92 0.91 0.91 569<br>weighted avg 0.92 0.92 0.92 569 |
| | b) Logistic Regression | `from sklearn import metrics`<br>`print(metrics.classification_report(expected,predected))`<br>precision recall f1-score support<br>0 0.90 0.86 0.88 212<br>1 0.92 0.94 0.93 357<br>accuracy 0.91 569<br>macro avg 0.91 0.90 0.91 569<br>weighted avg 0.91 0.91 0.91 569 |
| | c) Decision Tree | `from sklearn import metrics`<br>`print(metrics.classification_report(expected,predected))`<br>precision recall f1-score support<br>0 0.96 0.97 0.97 212<br>1 0.98 0.98 0.98 357<br>accuracy 0.98 569<br>macro avg 0.97 0.97 0.97 569<br>weighted avg 0.98 0.98 0.98 569 |
| | | |

## V. 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

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 4, Issue 1, July 2024

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].** Dr.Kalyanrao Takale, Machine Learning II, Golden Series By Nirali, 2021
**[2].** LeCessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc Ser C (Appl Stat). 1992;41(1):191–201.
**[3].** https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.
**[4].** Fujiyoshi H, Hirakawa T, Yamashita T. Deep learning-based image recognition for autonomous driving. IATSS Res. 2019;43(4):244–52.
**[5].** Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inform Theory. 1975;21(1):32–40.
**[6].** Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge: MIT Press; 2016.
**[7].** Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems. 2014: 2672–2680.
**[8].** Guerrero-Ibáñez J, Zeadally S, Contreras-Castillo J. Sensor technologies for intelligent transportation systems. Sensors. 2018;18(4):1212.
**[9].** Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011.
**[10].** Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: ACM Sigmod Record, ACM. 2000;29: 1–12.
**[11].** Harmon SA, Sanford TH, Sheng X, Turkbey EB, Roth H, Ziyue X, Yang D, Myronenko A, Anderson V, Amalou A, et al. Artifcial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. Nat Commun. 2020;11(1):1–7.
**[12].** He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904–16.
**[13].** He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770–778. 46. Hinton GE. A practical guide to training restricted boltzmann machines. In: Neural networks: Tricks of the trade. Springer. 2012; 599-619