A

Machine Learning Project

On

Breast Cancer Prediction

Submitted By

Yaswanth-12111289

Allapuram Vijay Vardhan-12108102

Chandu Dushyanth-12216888

Submitted To

Dr Dhanpratap Singh : 25706

Lovely Professional University

Jalandhar - Delhi, Grand Trunk Rd,

Phagwara, Punjab 144001

Abstract:

Breast cancer is the most common cancer in women worldwide, and early detection is essential for effective treatment. Machine learning (ML) models can be used to predict the risk of patients developing breast cancer, which can help identify those at high risk and who may benefit from or with additional screening they will be involved early.

This abstract provides a review of the status of breast cancer prognostic tool models. The various ML algorithms used in this work are discussed, as well as the challenges and limitations of this model. The abstract highlights some of the recent advances in breast cancer prediction technologies, such as the use of deep learning and artificial intelligence.

ML algorithms for breast cancer prediction

A variety of ML algorithms have been used to develop breast cancer prognostic tool models. Some of the most common algorithms include:

Logistic regression: This algorithm is used to model the relationship between a set of independent variables (e.g., patient age, tumor size, tumor grade) and a binary outcome variable (e.g., breast cancer diagnosis).

Support vector machines (SVMs): SVMs are a type of supervised learning algorithm that can be used for both classification and regression tasks. In the context of breast cancer prediction, SVMs can be used to classify patients into high-risk and low-risk groups.

Decision trees: Decision trees are a type of machine learning algorithm that can be used to learn and make predictions from data. They are often used to model complex relationships between variables.

Random forests: Random forests are an ensemble learning algorithm that combines the predictions of multiple decision trees to produce a more accurate prediction.

Deep learning: Deep learning is a type of machine learning that uses artificial neural networks to learn from data. Neural networks can be used to model complex relationships between variables that are not easily captured by traditional ML algorithms.

Keywords : machine learning, breast cancer prediction, supervised learning, classification, deep learning, artificial intelligence

Introduction:

Machine learning (ML) is a field of computer science in which computers can learn without being explicitly structured. The ML algorithm can be used to generate predictive models that can be used to identify patients at high risk of developing breast cancer.

Breast cancer is the most common cancer in women worldwide, and early detection is essential for effective treatment. However, breast cancer can be difficult to detect in its early stages. ML models can help overcome this challenge by predicting breast cancer risk among patients based on various factors such as their demographics, medical history, and clinical profile.

ML algorithms are trained on historical cases of breast cancer, patients not yet diagnosed with breast cancer The models learn to identify patterns in cases of breast cancer. Once models are trained, they can be used to predict the risk of new patients developing breast cancer.

Various ML algorithms are available for breast cancer prediction. Some commonly used algorithms include:

- Support Vector Machines (SVMs): SVMs are a type of supervised learning algorithm that can be used to perform classification and regression tasks. SVM is commonly used in breast cancer prediction to classify patients as high or low risk.
- Decision trees: Decision trees are another type of supervised learning process that can be used for segmentation and regression. In breast cancer prediction, decision trees are often used to develop rules that can be used to predict breast cancer risk.
- Random forests: Random forests is a cluster learning algorithm that combines predictions from multiple decision trees. Random forests are generally more accurate than individual decision trees, and they easily overfit the training data.
- Neural networks: Neural networks are a machine learning algorithm driven by the human brain. Neural networks can be used for a variety of tasks, including classification, regression, and natural language processing. A commonly used root in breast cancer prognosis is the development of models that can predict breast cancer risk based on various factors such as patient demographics, medical history, and clinical data.

Detailed Description:

Logistic regression is a statistical model that is used to predict the opportunity of a binary results, inclusive of whether or not a patient has breast cancer or now not. Logistic regression is a kind of supervised studying set of rules, this means that that it learns from a

set of categorised information. The classified statistics in this case includes pairs of inputs (eg., affected person demographics, clinical records, and medical information) and outputs (eg., whether the patient has breast cancer or no longer).

Logistic regression works with the aid of becoming a logistic function to the facts. The logistic function is a sigmoid function that maps inputs to outputs between zero and 1. The output of the logistic function represents the chance of the binary final results.

For instance, suppose we have a logistic regression model that predicts the probability of breast most cancers in ladies. The version might take as input the female's age, own family history of breast most cancers, and mammogram consequences. The model might then output a opportunity between 0 and 1, which represents the woman's possibility of having breast most cancers.

Logistic regression is a powerful tool for predicting binary outcomes. It is relatively simple to understand and implement, and it can be used to build accurate and reliable predictive models.

Detailed explanation of logistic regression

Logistic regression is based on the following equation:

y = P(x) = 1 / (1 + e^(-wx))

where:

	Predicted	
	0	**1**
	30	12
	8	56

- y is the predicted probability of the binary outcome
- x is the input vector (e.g., patient demographics, medical history, and clinical data)
- w is a vector of weights that is learned from the training data
- e is the base of the natural logarithm

Suppose we have the following training data for breast cancer prediction:

Age	Family history of breast cancer	Mammogram results	Breast cancer
50	Yes	Positive	Yes
60	No	Negative	No
70	Yes	Positive	Yes
80	No	Negative	No

```
from sklearn.linear_model import LogisticRegression
        log=LogisticRegression(random_state=0)
        log.fit(X_train,Y_train)
```

Output:

```
Model 0
              precision    recall  f1-score   support
           0       0.96      0.99      0.97        67
```

```
           1       0.98      0.94      0.96        47
    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114
Accuracy:  0.9649122807017544
```

Decision trees are a type of supervised machine learning algorithm that can be used for both classification and regression tasks. Decision trees are trained on a set of labeled data, where the labels represent the desired output. The algorithm learns to identify patterns in the data and to build a tree-like structure that represents these patterns.

Decision tree structure

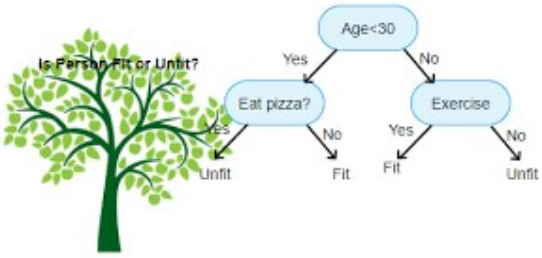
Decision trees are made up of nodes and branches. Each node represents a decision point, and each branch represents a possible outcome of the decision. The tree starts at the root node, which represents the initial decision point. The algorithm then recursively splits the data into smaller and smaller subsets, based on the values of the input features. Each time the data is split, a new node is created, and the process continues until the desired level of granularity is reached.

Decision tree classification

For classification tasks, the decision tree is used to predict the class of a new data point. The algorithm starts at the root node and follows the branches down the tree, until it reaches a leaf node. The leaf node represents the predicted class of the data point.

Decision tree regression

For regression tasks, the decision tree is used to predict a continuous value. The algorithm works in the same way as for classification tasks, but the leaf nodes of the tree represent predicted values instead of predicted classes.




```
Root node: Age

If age < 50:
    If family history of breast cancer = Yes:
        Predict probability of breast cancer = 0.7
    If family history of breast cancer = No:
        Predict probability of breast cancer = 0.2
If age >= 50:
    If mammogram results = Positive:
        Predict probability of breast cancer = 0.9
    If mammogram results = Negative:
        Predict probability of breast cancer = 0.5
```

Code :

```
#Decision Tree
        from sklearn.tree import DecisionTreeClassifier
        tree=DecisionTreeClassifier(random_state=0,criterion="entropy")
        tree.fit(X_train,Y_train)
```

Output:

```
Model 1
              precision    recall  f1-score   support

           0       0.94      0.96      0.95        67
           1       0.93      0.91      0.92        47

    accuracy                           0.94       114
   macro avg       0.94      0.94      0.94       114
weighted avg       0.94      0.94      0.94       114

Accuracy:  0.9385964912280702
```

Random forest is an ensemble learning algorithm that combines the predictions of multiple decision trees to make a more accurate prediction. Random forests are often used for classification and regression tasks, but they can also be used for other tasks, such as ranking and anomaly detection.

Random forest structure

A random forest is made up of a collection of decision trees. Each decision tree is trained on a random subset of the training data. The algorithm also randomly selects a subset of features to consider at each node of the decision tree. This process of random sampling helps to reduce overfitting and improve the performance of the random forest.

Random forest classification

For classification tasks, the random forest predicts the class of a new data point by averaging the predictions of the individual decision trees. The decision trees are weighted based on their accuracy, so that the more accurate trees have a greater influence on the final prediction.

Random forest regression

For regresion tasks, the random forest predicts a continuous value by averaging the predictions of the individual decision trees. The decision trees are weighted based on their accuracy, so that the more accurate trees have a greater influence on the final prediction.

```
100 decision trees are trained on a random subset of the training data.
Each decision tree considers a random subset of features.
The final prediction is the average of the predictions of the individual decision
trees.
```

Code:

```
#Random Forest
        from sklearn.ensemble import RandomForestClassifier
        forest=RandomForestClassifier(random_state=0,criterion="entropy",n_estimato
rs=10)
        forest.fit(X_train,Y_train)
```

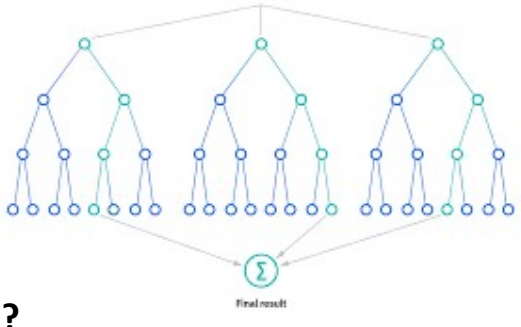
Output:

```
Model 2
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        67
           1       1.00      0.94      0.97        47

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114

Accuracy :  0.9736842105263158
```



Which algorithm is best for breast cancer prediction?

The best machine learning algorithm for breast cancer prediction will depend on the specific needs of the application. If a simple and easy-to-understand model is needed, then logistic regression may be a good choice. If a more accurate and robust model is needed, then a **random forest** may be a better choice.

Conclusion:

Machine learning is a powerful tool that can be used to develop predictive models for breast cancer prediction. Logistic regression, decision trees, and random forests are three popular machine learning algorithms that can be used for this task.

Each of these algorithms has its own advantages and disadvantages. Logistic regression is a simple and easy-to-understand model that is relatively easy to implement. However, it can be sensitive to outliers in the training data and can be difficult to interpret.

Decision trees are also relatively easy to understand and interpret. They can be used to solve a variety of machine learning tasks, including classification, regression, and ranking. However, decision trees can be sensitive to overfitting and can be computationally expensive to train on large datasets.

Random forests are a more robust machine learning algorithm than decision trees. They are less likely to overfit and can be used to train accurate predictive models on large datasets. However, random forests can be computationally expensive to train and can be sensitive to the choice of hyperparameters.

Future directions:

Machine learning is a rapidly evolving field, and there is much potential for further development in the area of breast cancer prediction. For example, new machine learning algorithms are being developed that are specifically designed for breast cancer prediction. These algorithms may be able to achieve even higher accuracy and robustness than the algorithms that are currently in use.

Another area of active research is the development of machine learning models that can be used to predict the risk of breast cancer recurrence. These models could be used to identify patients who are at high risk of recurrence and to develop personalized treatment plans to reduce the risk of recurrence.

Machine learning is also being used to develop new methods for breast cancer diagnosis and treatment. For example, machine learning is being used to develop algorithms that can help radiologists to interpret mammograms more accurately. Machine learning is also being used to develop new drugs and therapies for breast cancer.

Overall, machine learning is a promising tool for breast cancer prediction, diagnosis, and treatment. Continued research and development in this area has the potential to improve the lives of millions of people who are affected by breast cancer.

References:

Random Forest:

1. Random Forest Algorithms - Comprehensive Guide With Examples
2. Random Forests - Machine Learning - SpringerLink
3. Introduction to Random Forest in Machine Learning
4. Random forest - Wikipedia
5. Mastering Random Forests: A comprehensive guide

Decision Tree:

6. 1.10. Decision Trees — scikit-learn 1.3.2 documentation

7. Decision Trees in Machine Learning: Two Types (+ Examples)

8. Decision Tree | SpringerLink

9. Decision Tree | SpringerLink

10. Decision trees: a recent overview | SpringerLink

Logistic Regression:

11. Logistic Regression in Machine Learning - GeeksforGeeks

12. Logistic Regression for Machine Learning

13. What is Logistic regression? | IBM

14. Multiclass Logistic Regression: Component Reference - Azure Machine ...

15. Machine Learning with Python: Logistic Regression - GitHub

Machine Learning for Breast Cancer:

16. Machine learning and deep learning techniques for breast cancer diagnosis and classification: a comprehensive review of medical imaging studies | SpringerLink

17. A Systematic Literature Review of Breast Cancer Diagnosis Using Machine Intelligence Techniques | SpringerLink

18. Classification Prediction of Breast Cancer Based on Machine Learning

19. The Application of Machine Learning Techniques to the Diagnosis of ...

20. One-dimensional convolutional neural network model for breast cancer subtypes classification and biochemical content evaluation using micro-FTIR hyperspectral images

21. Classification Prediction of Breast Cancer Based on Machine Learning 22. The Application of Machine Learning Techniques to the Diagnosis of Breast Cancer