**PROJECT REPORT**

**AIM**

The main aim of this project is to Develop a predictive model for breast cancer diagnosis using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to provide accurate and timely assessments of malignancy. This model aims to provide timely and precise assessments of malignancy, thereby contributing to improved clinical decision-making and patient outcomes.

**Objectives**

1. Data Preprocessing and exploratory data analysis on the dataset. Handling of missing values and looking for outliers so that to check which algorithm will be better suited for this dataset. Also investigate the correlation matrix and perform dimensionality reduction using PCA (Principal Component Analysis) if necessary.
2. Build a model using neural networks algorithm. If there are many inconsistencies in the dataset its better to resolve all those inconsistencies or choose an algorithm that is robust against all those inconsistencies. Model performance will be assessed using established metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) to ensure optimal diagnostic capabilities.
3. Deploy the model in a cloud platform like Stream lit. After the model has reached a good accuracy and satisfies all other metrics like F1 score and recall. The model will b deployed in stream lit platform where a user interface will be created to enter all values and based on that the output will be shown that if it is malignant or benign. The model will only be deployed in the cloud platform if approved.

**Literature Review**

Breast cancer is cancer that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, milk rejection, fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin. In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin. Breast cancer most commonly presents as a lump that feels different from the rest of the breast tissue. More than 80% of cases are discovered when a person detects such a lump with the fingertips. The earliest breast cancers, however, are detected by a mammogram. Lumps found in lymph nodes located in the armpits may also indicate breast cancer.

This dataset consists of a total of 569 samples. There are a total of 32 features that characterize samples, the first of which is the ID of the sample, the second is its class, and the remaining 30 are features that contain various information about the cells. The class label of our samples can be malignant (M) or benign (B). Benign tumours can be removed surgically and do not reappear. Cancerous tumours are those that have undergone a malignant transformation. The Dataset was obtained by research students by imaging a needle-tip-wide breast mass with biosip by Dr. William Wolberg, an employee of Wisconsin Hospital, and digitizing these images by William Nick Street, one of the researchers of the University of Wisconsin Computer Sciences Department, in November 1995.

Cancer is a significant public health issue globally, marked by an elevated incidence and mortality rate.According to the GLOBOCAN 2020 database, approximately 19.3 million new cases and 10 million deaths have been reported annually. The introduction of machine learning and artificial intelligence positively supports cancer prevention and management.Artificial intelligence is commonly defined as a set of computer-coded programs or algorithms that use data analysis and pre-programmed instructions to make predictions and decisions about various aspects of a disease. Machine learning is a specialized field within AI that refers to a group of algorithms designed to automatically learn and improve from experience. In other words, machine learning is an AI subset that focuses on developing algorithms capable of learning from data and refining their performance over time. Deep Learning is a subfield of “Machine Learning” that employs neural network-based models to imitate the human brain’s capacity to analyse huge amounts of complicated data in areas such as language processing, drug discovery, and image recognition. All of these computer algorithms use data such as investigations performed, scans conducted, patients’ medical histories, and other information to forecast or diagnose a cancerous condition.

**Methodology**

The methodology used in this assignment is by using python programming language to build the model using algorithms like neural networks which is a deep learning algorithm. The aim is to build a model based on the dataset which has 32 features and one of them is class which says Malignant or Benign. This class is used as dependent variable and other 30 features are used as independent variables to predict this class. Here the 30 features are 30 variables related to breast cancer findings. Here the class label is a categorical variable so here label encoding is used. The 30 variables are like fractal dimension, concavity, and compactness etc. before building the model its important to preprocess data. All the variables have outliers in it. The method used to look for outliers are robust univariate statistics like Hampel 84 method used to see outliers based on median. Some of the variables are highly correlated and PCA was used to reduce the dimensionality. Here anyways the neural network algorithm is robust against all these problems in the dataset. The neural networks activation functions used are RELU and Sigmoid, sgd optimizer was used and binary cross entropy, here sgd means stochastic gradient descent. 100 epochs were taken. Before feeding the data into model it was scaled using RobustScaler. The validation is also done by taking 50 percent for test and training was done by taking 30 percent for test. So the PCA code is optional. The evaluation metrics used are F1 score, Precision, Accuracy and Recall. The threshold is given as 0.7. the metrics obtained after training the model are.

Accuracy: 0.9418604651162791

F1 Score: 0.9122807017543859

Recall: 0.8387096774193549

Precision: 1.0

These metrics are good with good accuracy of 94 percent which means the model is performing well to identify who is in malignant and benign stage. All other metrics are also perfect.

**Conclusion**

Finally, it is safe to say that the model is performing well with correctly classified cases as we can observe the accuracy, F1 score and Precision which is perfect, and the model can classify new cases based on given inputs with 94 percent accuracy. The training accuracy is also good here which is of 97 percent, the difference between training and test is not so high. However, interpretability can be challenging, making it crucial to understand the model's decision-making process. Data quality and bias in training data may lead to inaccurate predictions. Also neural networks has problems like overfitting sometimes it trains too well on training data that it doesn’t generalize well the new unseen data. In future many such considerations is to be taken and worked on before working on any new project

**References**

1. X. Chen, Y. Ye, G. Williams, and X. Xu, “A survey of open source data mining systems,” *Lecture Notes in Computer Science*, vol. 4819, pp. 3–14, 2007.

View at: [Publisher Site](https://doi.org/10.1007/978-3-540-77018-3_2) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20survey%20of%20open%20source%20data%20mining%20systems&author=X.%20Chen&author=Y.%20Ye&author=G.%20Williams&author=X.%20Xu&publication_year=2007)

1. F. Kayaalp and M. Başarslan, “Open source data mining programs: a case study on R,” *Duzce University Journal of Science and Technolgy*, vol. 6, pp. 455–468, 2018.

View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Open%20source%20data%20mining%20programs%3A%20a%20case%20study%20on%20R&author=F.%20Kayaalp&author=M.%20Ba%C5%9Farslan&publication_year=2018)

1. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, “Data mining: practical machine learning tools and techniques,” *Acm Sigmod Record*, vol. 31, pp. 76-77, 2002.

View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Data%20mining%3A%20practical%20machine%20learning%20tools%20and%20techniques&author=I.%20H.%20Witten&author=E.%20Frank&author=M.%20A.%20Hall&author=C.%20J.%20Pal&publication_year=2002)

1. S. Eltalhi and H. Kutrani, “Breast cancer diagnosis and prediction using machine learning and data mining techniques: a review,” *IOSR Journal of Dental and Medical Sciences*, vol. 18, no. 4, pp. 85–94, 2019.

View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Breast%20cancer%20diagnosis%20and%20prediction%20using%20machine%20learning%20and%20data%20mining%20techniques%3A%20a%20review&author=S.%20Eltalhi&author=H.%20Kutrani&publication_year=2019)

.