**Breast Cancer Risk Prediction System**

|  |  |  |
| --- | --- | --- |
| **1** | **INTRODUCTION**  Overview  Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. One of the best ways to fight cancer is early detection, when it is still confined and can be fully excised surgically or treated pharmacologically. Cancer screening programs, that is, the practice of testing for the presence of cancer in people who have no symptoms, has been medicine’s tool of choice for the earliest detection.  Computer-aided diagnosis systems showed the potential for improving diagnostic accuracy. But early detection and prevention can significantly reduce the chances of death. It is important to detect breast cancer as early as possible.  Here we are developing a machine learning model where in the model gets trained by considering the parameters such as: Radius ,Texture, Perimeter, Area, Smoothness, Concavity, Concaveness, Compactness here all these parameters are taken in mean, se and overall values are been taken. And the model is been trained using Auto AI service in IBM Watson cloud and that can be deployed in an application such as web or mobile applications.  (i)Downloading dataset  (ii) Data Preprocessing  (iii) Model Building  (iv) Application Building | |
|  |
|  |
|  | * 1. Purpose   Since the labels in the data are discrete, the predication falls into two categories, (i.e. Malignant or benign). In machine learning this is a classification problem. Thus, the goal is to classify whether the breast cancer is benign or malignant and predict the recurrence and non-recurrence of malignant cases after a certain period.  To achieve this, I have used machine learning classification methods to fit a function that can predict the discrete class of new input.  Through this repository we can:  ● Apply the fundamental concepts of machine learning from an available dataset  ● Evaluate and interpret my results and justify my interpretation based on observed data set  ● Create notebooks that serve as computational records and document my thought process.  The analysis is divided into four sections, saved in juypter notebooks in this repository  1. Identifying the problem and Data Sources  2. Exploratory Data Analysis  3. Pre-Processing the Data  4. Build model to predict whether breast cell tissue is malignant or Benign | |
| **2** | **LITERATURE SURVEY** | |
|  | 2.1 Existing problem  Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women. Breast Cancer occurs because of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound, and biopsy are commonly used to diagnose breast cancer performed. | |
|  | 2.2 Proposed solution  I will be building a model in Watson Studio and deploying the model in IBM Watson Machine Learning. To interact with the model, I will be using Node-Red and scoring Endpoint.  Solution Requirements:  Develop a model that is capable of detecting the Breast Cancer in early stages. The Machine learning model is trained and deployed on IBM Watson Studio and an endpoint is created. The web application is built using IBM Node-Red. | |
| **3** | **THEORITICAL ANALYSIS** | |
|  | 3.1 Block diagram | |
|  | 3.2 Hardware / Software designing  IBM Nodered,IBM Watson Studio,IBM Machine Learning,IBM Cloud Object Storage | |
|  | **4** EXPERIMENTAL INVESTIGATIONS  The dataset in the given problem consist of 569 columns and 6 rows.  The prediction was to make whether according to the data from the previous data, is the new patient likely to have breast cancer or not.  Diagnosis column has the output as 0 and 1, where 1 represents patient with the cance r ,wheras 0 represents no cancer. During analysis of the dataset, and studying the heatmap of the correlation between different features, I came to a conclusion that no part of the dataset should not be ignored/deleted.  There is no categorical data in the dataset. Batch of the data is bisected into 80:20 ratio of training test: test data. | |
|  | **5. RESULT**      **Graphical user interface, application  Description automatically generated**  **Graphical user interface, application  Description automatically generated** | |
|  | **6. ADVANTAGES & DISADVANTAGES**  **ADVANTAGES:**  1.Both the model and application is lightweight.  2. Prediction speed is high.  3. Server side is authenticated.  4. The prediction is helpful in educational purposes.  **DISADVANTAGES:**  **1**. Node-red is not suitable for commercial purposes.  2. Predictions on missing feature can be inaccurate. | |
|  | **7. APPLICATIONS**  1. Through this model we can predict, whether a patient is likely to have cancer or not, without even doing medical tests.  2. Medical test can be modified and optimized.  3. We can analyse which starta of population (in women) are more likely to have it in the future.  4. The application can be run on android by SL4A.  5. It would be very useful and handy tool in healthcare.  6 . It can run on PC server very fast.  7. It bypasses the first level of manual inspection. | |
|  | **8. CONCLUSION**  Because of this ongoing covid-19 pandemic, now people have an urge to have pre knowledge of all medical ailments and advancements. For better working of the model we would be needed actual and large dataset. Since by the level of dataset in the repository ,the results are "good" .In the source code I have 8 types of models ,but on the IBM cloud ,Node-red I have deployed it by using only LGBM classification algorithm. The model was trained on a dataset of 569 patients, the total number of features were 5. Feature scaling was very important in this problem set ,as the classification algorithm used demands uniform distancing between all features.  The model after initial testing was deployed IBM Watson and NODE-RED.  This intermediate level of machine learning is very necessary to understand to leave scope for potential developments. | |
|  | **9. FUTURE SCOPE**  As per WHO breast cancer is a deadly cancer, which develops inside the human body without even showing symptoms. Since in the initial days of the disease, no symptoms are witnessed by the patient, the diseases develops into later stages of the deadly cancer .If by regular examination ,we can deploy the dataset into the model to predict ,to very good accuracy we can find about the cancer ,without any medical tests. That is why, this prediction algorithm has great future ahead, if it keeps on learning from bigger dataset. ` | |
|  | **APPENDIX** | |
|  | | 1. Source code |
|  | | var rm = global.get("rm")  var pm = global.get("pm")  var am = global.get("am")  var cm = global.get("cm")  var cnm = global.get("cnm")  var cpm = global.get("cpm")  var token=msg.payload.access\_token  msg.headers={'Content-Type': 'application/json',"Authorization":"Bearer "+token,"Accept":"application/json"}  msg.payload={"input\_data":[{"fields":["radius\_mean","perimeter\_mean","area\_mean","compactness\_mean","concavity\_mean","concave points\_mean"],"values":[[rm,pm,am,cm,cnm,cpm]]}]}  return msg; |
|  | | global.set("rm",msg.payload.rm)  global.set("pm",msg.payload.pm)  global.set("am",msg.payload.am)  global.set("cm",msg.payload.cm)  global.set("cnm",msg.payload.cnm)  global.set("cpm",msg.payload.cpm)  var apikey="gLimCx8lIZ3DlqYy5XbTTV5XIP3xKjmGQvO8cYJd2AQG";  msg.headers={"content-type":"application/x-www-form-urlencoded"}  msg.payload={"grant\_type":"urn:ibm:params:oauth:grant-type:apikey","apikey":apikey}  return msg; |