CONCEPT MEDICO (ML BASED -DISEASE PREDICTION SYSTEM)

OVERVIEW

1. Project Background and Description

Purpose of the System

Building a system which can predict different types of disease as well as it can also provide Doctor detail for appointment and also generate a lab report. (lab report is generated as basis on the user input and ML model prediction. If doctor suggestion is there then after doctors suggestion lab report is again generated)

Team Member

1. Achal Rajesh Mate
2. Pushkar Ashok Narkhede
3. Ritu Kumari

Guide :

- Prof . Rajeshree Nayak

2. Project Scope

1. To contribute towards the well-being of living society.
2. This Project helps Patients and Doctors to get easy understanding about that the patient is facing the disease or not
3. As it integrated with ML model so patients and doctors can predict the level of high order diseases like cancer

3. Technology Stack

Tools--
1. Jupyter-Lab or Jupyter Notebook
2. Pycharm
3. Xampaa (DB Tool)
4. PhpMyAdmin (DB Handler)

Technology--
1. Html 5
2. Css
3. Javascript for authentication
4. Python Programming (for web development and integration of ML model)
5. flask as a web framework
6. Machine Learning. (Classification algorithm (Supervised ML)
7. Mysql Database.

4. Development Stages with their Duration

Stage 1	Stage Name	Starting Date	Ending Date	Status
1	Data Collection	1 Sept 2021	6 Sept 2021	Completed
2	Data Analysis	7 Sept 2021	14 Sept 2021	Completed
3	Feature Selection	15 Sept 2021	21 Sept 2021	Completed
4	Model Selection and Building	22 Sept 2021	30 Sept 2021	Completed
5	Model Deployment	1 Oct 2021	31 Oct 2021	In Process

5. Implementation plan

For Implementation we have Prepare the 3 Groups of the Disease

Group 1 (Project step 1 Building ML Model on this 3 diseases)

1. Breast cancer with 2 stage prediction
2. Covid 19 infection prediction
3. Heart (Cardio vascular) disease prediction

Group 2 (Building web interface for Group 1 ML model --
Project step 2 Building ML model on this diseases --

4. Chronical kidney disease prediction
5. Liver disease prediction
6. Lung cancer prediction 2 stages (stage 1. Lung cancer or not stage 2. severity of lung cancer)

Group 3 (adding Group 2 ML model to web interface --
Project step 3 Building ML model on this diseases if possible --

7. All types of regular disease
8. Hypothyroid prediction
9. Tumor prediction (i.e. in which part of body tumor is growing)
10. Diabetes prediction

We Have started with Group 1

1. Breast cancer
2. Covid 19 infection prediction
3. Heart (Cardio vascular) disease prediction

Details Description of each disease

1. **Breast Cancer Prediction**

   Assign to = Achal Mate

Dataset Description –

- Number of Instances: 569
- Number of Attributes: 30 numeric, predictive attributes and the class

Attribute Information:
- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.
For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

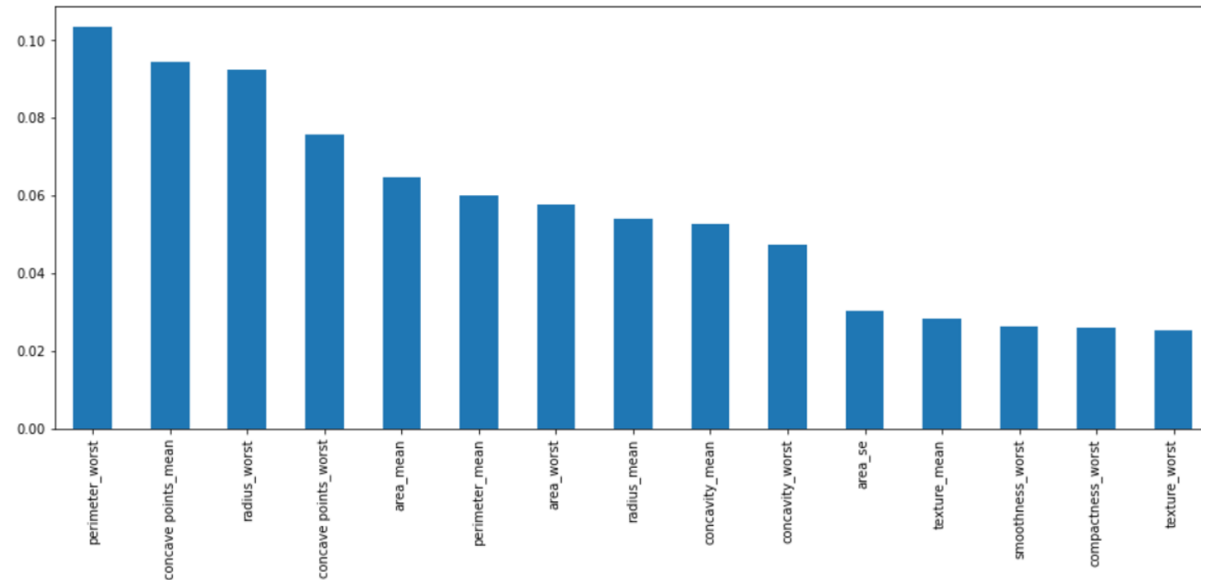
- Missing Values: 569
- Class Distribution: 212 - Malignant, 357 – Benign

Depending on the types of cells in a tumor, it can be:
Benign - The tumor doesn’t contain cancerous cells.
Malignant - The tumor contains cancerous cells.

Observation:

fractal_dimension_mean,texture_se, smoothness_se, symmetry_se, and
fractal_dimension_se are least correlated with the target variable.
All other features have a significant correlation with the target variable.(i.e diagnosis)

Feature Importance -



top 15 features which contributes to target variable

1. Need to verify which feature is significant and which is not from this top 15
2. for this we will be using Hypothesis testing and other feature selection technique

Hypothesis testing -

- Null Hypothesis ==> all the features are not significent
- Alternate Hypothesis ==> all the feature are significent
- alpha values ==>0.05

by Hypothesis testing
if(p_value < alpha) then **reject Null Hypothesis**
else **Fail to reject Null Hypothesis**
or
if(p_value > alpha) then **Fail to reject Null Hypothesis**
else **reject Null Hypothesis**

calculating VIF score we get top 9 feature

1. area_mean
2. texture_mean
3. fractal_dimension_mean
4. fractal_dimension_se
5. texture_se
6. smoothness_se
7. concavity_se
8. symmetry_se
9. area_se

But Our Dataset Contain 10 Independent features which further they are categories into three categories

1. Mean
2. Worst
3. Square(Se)

As Above selected feature does not include any worst categories' feature so to get more significant we drop all worst categories features from the main dataset

Top 15 RFE features

[('radius_mean', True, 1),
('texture_mean', True, 1),
('perimeter_mean', True, 1),
('area_mean', True, 1),
('smoothness_mean', True, 1),
('compactness_mean', True, 1),
('concavity_mean', True, 1),
('concave points_mean', True, 1),
('symmetry_mean', True, 1),
('fractal_dimension_mean', True, 1),
('radius_se', True, 1),
('texture_se', True, 1),
('perimeter_se', True, 1),
('area_se', True, 1),
('smoothness_se', False, 5),
('compactness_se', False, 4),
('concavity_se', False, 2),
('concave points_se', False, 6),
('symmetry_se', False, 3),
('fractal_dimension_se', True, 1)]

Observation

Among this features

- smoothness_mean
- concave points_mean
- radius_se
- area_mean
- perimeter_mean
- fractal_dimension_mean
- perimeter_se
- radius_mean
- ompactness_mean

are the features with higher **p value** than alpha value
So they are drop to get the significant feature
Now, VIF score is also in the range .

Selected Features

1. texture_mean
2. concavity_mean
3. symmetry_mean
4. texture_se
5. area_se
6. fractal_dimension_se

Model Buliding

Model and Training Accuracy

Algorithm	Accuracy
ExtraTreesClassifier	0.982456
Random Forest Classifier	0.973684
XGBClassifier	0.973684
Gradient Boosting Classifier	0.956140
svc rbf Kernal	0.947368
K Neighbors Classifier	0.947368
navieBayes_gaussian	0.938596
Ada Boost Classifier	0.938596
svc linear Kernal	0.929825
logistic_name	0.921053
Decision Tree Classifier	0.912281

After Perfoming Parameter Tuning

Algorithm	Accuracy
ExtraTreesClassifier	0.982456
xgb_classifier_test	0.982456
RandomForestClassifier	0.973684
Logistic Regression	0.964912
SVC	0.964912
KNeighborsClassifier	0.956140
AdaBoostClassifier	0.947368
GaussianNB	0.938596
Gradient Boosting Classifier	0.921053
DecisionTreeClassifier	0.903509

2 . Covid 19 Prediction

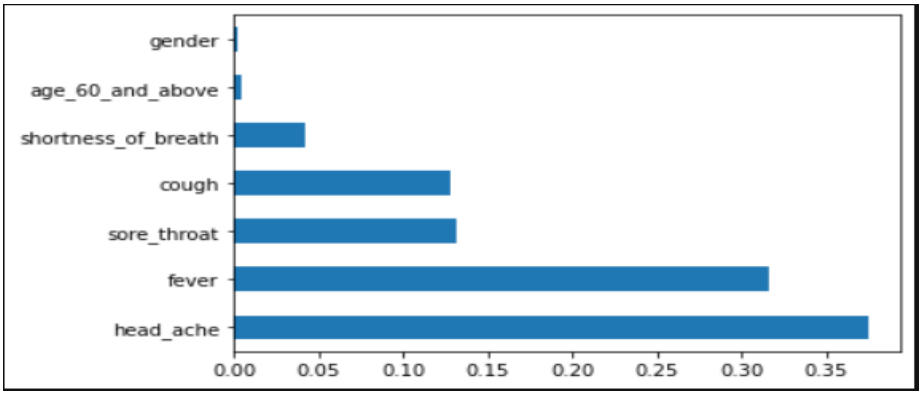
Assign To= Pushkar Narkhede

Dataset Description -

1. test_date ==> Date for covid 19 test
2. cught ==> is the patience having cough or not (binary variable yes=1, no=0)
3. fever ==> is the patience having fever or not (binary variable yes=1, no=0)
4. sore_throat ==> is the patience having sore_throat or not (binary variable yes=1, no=0)
5. shortness_of_breadth ==> is the patience having shortness_of_breadth or not (binary variable yes=1, no=0)
6. head_ache ==> is the patience having head_ache or not (binary variable yes=1, no=0)
7. corona_result ==> is the patience having corona_result or not (categorical variable -ve for covid,+ve for covid, other=> no confirmation with covid or not) *self encoded* **(1= +ve, 0= -ve)**
8. gender ==> is the patience having gender or not (categorical variable male,female) *self encoded* **(1= male, 0=female)**
9. test_indication ==> is the patience having gender or not (categorical variable contact_with_confirm=> covid due to contact of covid 19 person, abord=> covid in other country, other=>no idea about covid infection)

- Total data in dataset : 2742596
- Total null values in dataset : 640530
- Total data remains if we remove all null values = 2102066
- 23.35 % of null data is present in the dataset.

Feature Importance -



P-Values of all Features are 0.00

Interpretation

- After performing hypothesis testing by considering p value **gender** i am on the conclusion that all features we have are significant
- same interpretaion is given by RFE(recursive feature selection) and ExtraTreeClassifier feature contribution score.
- if **age** feature is not contributing more we will be dropping that feature.
- VIF score is also in the range.

Feature selected-

1. cough
2. fever
3. sore_throat
4. shortness_of_breath
5. head_ache
6. age_60_and_above (on hold)

Model Selection

Models and Training Accuracy

1) Decision Tree = 0.945463
2) Random Forest = 0.945463
3) Navie bayes= 0.940577
4) Logistics Regression=0.938743
5) Gradient Boosting=0.945463
6) Xgboost= 0.945463
7) ADABoost=0.934402

After performing hyperparameter tunning

1) Navie bayes = 0.940577
2) Logistics Regression=0.938743
3) Gradient Boosting=0.945463
4) Xgboost=0.945463
5) ADABoost=0.934402

3 Heart Cancer Prediction

Assign to = Ritu Kumari

Dataset Information : The dataset contains 76 features from 303 patients, however, published studies chose only 14 features that are relevant in predicting heart disease. Hence, In this project we will be using the dataset consisting of 303 patients with 14 features set.

Features Explanations:

1) age (Age in years)

2) sex : (1 = male, 0 = female)

3) cp (Chest Pain Type): [0: asymptomatic, 1: atypical angina, 2: non-anginal pain, 3: typical angina]

4) trestbps (Resting Blood Pressure in mm/hg)

5) chol (Serum Cholesterol in mg/dl)

6) fps (Fasting Blood Sugar > 120 mg/dl): [0 = no, 1 = yes]

7) restecg (Resting ECG): [0: showing probable or definite left ventricular hypertrophy by Estes' criteria, 1: normal, 2: having ST-T wave abnormality]

8) thalach (maximum heart rate achieved)

9) exang (Exercise Induced Angina): [1 = yes, 0 = no]

10) oldpeak (ST depression induced by exercise relative to rest)

11) slope (the slope of the peak exercise ST segment): [0: downsloping; 1: flat; 2: upsloping]

12) ca [number of major vessels (0–3)

13) thal : [1 = normal, 2 = fixed defect, 3 = reversible defect]

14) target: [0 = disease, 1 = no disease] ---> In The dataset we have 303 rows with 14 variables

variables types:

1) Binary: sex, fbs, exang, target

2) Categorical: cp, restecg, slope, ca, thal

3) Continuous: age, trestbps, chol, thalac, oldpea