Problem Definition

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. 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.Breast cancer is a cancer in which the cells of breast tissue get altered and undergo uncontrolled division, resulting in a lump or mass in that region. It is generally diagnosed as one of the two types:

- Benign (Non-cancerous)
- Malignant (Cancerous)

Brief on the dataset used: the Wisconsin breast cancer diagnostic data set for predictive analysis

Attribute Information:

1) ID number 2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

1) ID number

2) Diagnosis (M = malignant, B = benign)

3–32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter² / area — 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension (“coastline approximation” — 1)

All columns contain numerical data except the “diagnosis” attribute. This diagnosis attribute contains cancer type, i.e.- M(for malignant) or B(for benign), it is text data.

Python packages

- Numpy v1.19.1
- Matplotlib v3.3.1
- Plotly v4.9.0
- Seaborn v0.10.1
- Sci-kit learn v0.23.2

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data.

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

The plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.

Seaborn is a library mostly used for statistical plotting in Python. It is built on top of Matplotlib and provides beautiful default styles and color palettes to make statistical plots more attractive.

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Importing the necessary libraries

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# keeps the plots in one place. calls image as static pngs
%matplotlib inline
import matplotlib.pyplot as plt # side-stepping mpl backend
import matplotlib.gridspec as gridspec # subplots


#Import models from scikit learn module:
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn import metrics
```

```
import seaborn as sns
import itertools
from itertools import chain
from sklearn.feature_selection import RFE
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import VotingClassifier
from sklearn.model_selection import GridSearchCV, cross_val_score, learning_curve, train_t
from sklearn.metrics import precision_score, recall_score, confusion_matrix, roc_curve, pr
import warnings
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls
import plotly.figure_factory as ff
```

Load the data

```
from google.colab import files
uploaded = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving data.csv to data (1).csv

```
p = 'data.csv'
df = pd.read_csv(p)
print(df.shape)
```

```
(569, 33)
```

```
df.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothr
0	842302	M	17.99	10.38	122.80	1001.0	
1	842517	M	20.57	17.77	132.90	1326.0	
2	84300903	M	19.69	21.25	130.00	1203.0	
3	84348301	M	11.42	20.38	77.58	386.1	
4	84358402	M	20.29	14.34	135.10	1297.0	

5 rows × 33 columns

Data Pre-processing:

↳ *1 cell hidden*

Normalization

[ ] ↳ *8 cells hidden*

Standardization

[ ] ↳ *5 cells hidden*

Discretization

[ ] ↳ *4 cells hidden*

•Data Summarization:

[ ] ↳ *9 cells hidden*

Data Visualization:

↳ *1 cell hidden*

Historgram

[ ] ↳ *3 cells hidden*

Line plot

[ ] ↳ *1 cell hidden*

Scatter plot

[ ] ↳ *1 cell hidden*

nucleus features vs diagnosis

[ ] ↳ *1 cell hidden*

Stack the data

[ ] ↳ *1 cell hidden*

Observations

1. mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in classification of the cancer. Larger values of these parameters tends to show a correlation with malignant tumors.
    2. mean values of texture, smoothness, symmetry or fractual dimension does not show a particular preference of one diagnosis over the other. In any of the histograms there are no noticeable large outliers that warrants further cleanup.

[ ] ↳ *2 cells hidden*

Positive correlated features

[ ] ↳ *1 cell hidden*

Box plot

[ ] ↳ *1 cell hidden*

Data Interpretation

Record all your findings and summary about data

[ ] ↳ *3 cells hidden*

Model Classification

[ ] ↳ *3 cells hidden*

Logistic Regression model

[ ] ↳ *5 cells hidden*

Decision Tree Model

[ ] ↳ *4 cells hidden*

Random Forest

[ ] ↳ *7 cells hidden*

Using on the test data set

[ ] ↳ *2 cells hidden*

Conclusion

↳ *1 cell hidden*

Downloading the final dataset

[ ] ↳ *1 cell hidden*