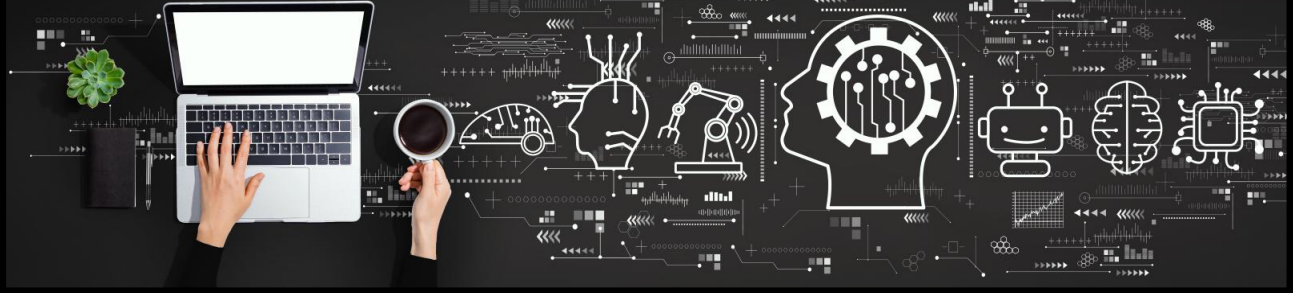
BREAST CANCER MACHINE LEARNING CAPSTONE PROJECT REPORT

BY: UTPAL MISHRA



DATASET

FOR MODELING:

SOURCE

Breast Cancer Wisconsin (Diagnostic) Data Set
Abstract: Diagnostic Wisconsin Breast Cancer Database

DATA SET INFORMATION

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:
ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/

ATTRIBUTE INFORMATION

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 (perimeter^2 / 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)

FOR COUNTRY WISE ANALYSIS:

SOURCE

Link:
https://www.wcrf.org/dietandcancer/cancer-trends/data-cancer-frequency-country

ABOUT

The age-standardized rate for all cancers (including non-melanoma skin cancer) for men and women combined was 197.9 per 100,000 in 2018. The rate was higher for men (218.6 per 100,000) than women (182.6 per 100,000).

Age-standardized rates are used in the tables. This is a summary measure of the rate of disease that a population would have if it had a standard age structure. Standardization is necessary when comparing populations that differ concerning age because age has a powerful influence on the risk of dying from cancer.

GLOBAL CANCER RATE:

The highest cancer rate for men and women together was in Australia, at 468.0 people per 100,000.
The age-standardized rate was at least 320 per 100,000 for 12 countries: Australia, New Zealand, Ireland, Hungary, the US, Belgium, France (metropolitan), Denmark, Norway, the Netherlands, Canada and New Caledonia (France).
The countries in the top 12 come from Oceania, Europe and North America.

CANCER RATE IN MENS:

The highest cancer rate was found in Australia at 579.9 men per 100,000.
The age-standardized rate was at least 360 per 100,000 in 15 countries: Australia, New Zealand, Ireland, Hungary, France (metropolitan), the US, Latvia, Belgium, Norway, Slovenia, Estonia, Slovakia, Denmark, New Caledonia (France) and the Netherlands.
The countries in the top ten come from Europe, Oceania and the Americas.

CANCER RATE IN WOMENS:

The highest cancer rate was found in Australia at 363.0 women per 100,000.
The age-standardized rate was at least 300 per 100,000 in 11 countries (Australia, New Zealand, Hungary, Belgium, Canada, Denmark, Ireland, the US, the Netherlands, Norway and South Korea).
The countries in the top ten come from Europe, Oceania and the Americas.