Livia Halim (lhalim)

Ryan Halim (rhalim14)

Dataset Name: breastcancer

Dataset Description: This dataset contains information from 1,000 examinations of women who received mammography (specialized medical imaging of the breast) within the Breast Cancer Surveillance Consortium (BCSC) between January 2000 and December 2009. The following dataset includes characteristics of US women who have been previously diagnosed with breast cancer, such as age, race/ethnicity, family history of breast cancer, menarche age, age at first birth, breast density, use of hormone replacement therapy, menopausal status, BMI, and history of biopsy.

This dataset instantiates an interesting problem appropriate for machine learning because it is beneficial for people who are fascinated in exploring the breast cancer distribution in US women. Using this dataset, we will be able to determine breast cancer risk factors and explore associations between them. Furthermore, we believe our dataset is not too easy because we have done extensive research that confirms this dataset or similar datasets have not been previously used for a machine learning study, so published models do not exist. No one knows yet what the highest risk factors are for breast cancer. On the other hand, this dataset is not too hard because it is widely accessible due to the fact that mammography is common among women living in the US, and everyday there are many patients that come in to do the procedure.

In order to ensure that the dataset is of high quality, we selected data that comes from a period of ten years. From the original dataset, we also removed mammography data that gives an unknown value for its output –whether or not the women was diagnosed with breast cancer. We believe that having the unknown value in our dataset will cause our machine learning study to be less accurate and less reliable.

The format of the input instance is provided in the table below:

We have decided to divide the data into training and test using the ratio of 80:20. From 1000 data, we randomly choose 800 data to be in the training set and the other 200 to be in the test set. This is because we believe by following the Pareto principle (80/20 rule), there will be less variance in performance statistics. Furthermore, the 80:20 split is a common occurring split in the field of machine learning.

We believe that the dataset is ethical because it follows the ACM Code of Ethics. The dataset does not provide any information that could be used to either directly or indirectly harm others. We do not provide the patients' names or birth dates to respect their privacy. In order to identify each data without the names, we provide a unique data id for each of the data.

Livia Halim (lhalim)

Ryan Halim (rhalim14)

Variable Name	Description	Coding
data_id	Unique ID for each data	1-1000
year	Calendar year of observation	Numerical, 2000-2009
age_group_5_years	Age (years) in 5 year groups	1 = Age 18-29 2 = Age 30-34 3 = Age 35-39 4 = Age 40-44 5 = Age 45-49 6 = Age 50-54 7 = Age 55-59 8 = Age 60-64 9 = Age 65-69 10 = Age 70-74 11 = Age 75-79 12 = Age 80-84 13 = Age >85
race_eth	Race/ethnicity	1 = Non-Hispanic white 2 = Non-Hispanic black 3 = Asian/Pacific Islander 4 = Native American 5 = Hispanic 6 = Other/mixed 9 = Unknown
first_degree_hx	History of breast cancer in a first degree relative	0 = No 1 = Yes 9 = Unknown
age_menarche	Age (years) at menarche	0 = Age >14 1 = Age 12-13 2 = Age <12 9 = Unknown
age_first_birth	Age (years) at first birth	0 = Age < 20 1 = Age 20-24 2 = Age 25-29 3 = Age >30 4 = Nulliparous 9 = Unknown
BIRADS_breast_density	BI-RADS breast density	1 = Almost entirely fat 2 = Scattered fibroglandular densities 3 = Heterogeneously dense 4 = Extremely dense 9 = Unknown or different measurement system
current_hrt	Use of hormone replacement therapy	0 = No 1 = Yes 9 = Unknown
menopaus	Menopausal status	1 = Pre- or peri-menopausal 2 = Post-menopausal 3 = Surgical menopause 9 = Unknown
bmi_group	Body mass index	1 = 10-24.99 2 = 25-29.99 3 = 30-34.99 4 = 35 or more 9 = Unknown
biophx	Previous breast biopsy or aspiration	0 = No 1 = Yes 9 = Unknown
breast_cancer_history	Prior breast cancer diagnosis	0 = No 1 = Yes 9 = Unknown
count	Frequency count of this combination of covariates	Numerical