

**BAN NBB - 130**

**PROJECT**

BAN 130 NBB



March 17, 2021

:

**SENECA COLLEGE, BLACKBOARD ULTRA**



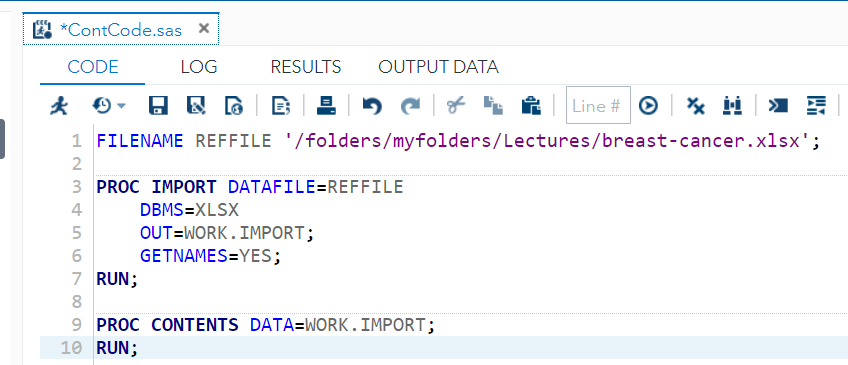
**DATASET – BREAST CANCER**



## Q-1. What is the file size of the dataset you have chosen for your project?

## **A-1.** The size of the dataset (Breast Cancer) we have chosen was 122.27 KB in the csv format when downloaded, when we opened in Excel it was 163KB but after loading into SAS finally it become- 256KB

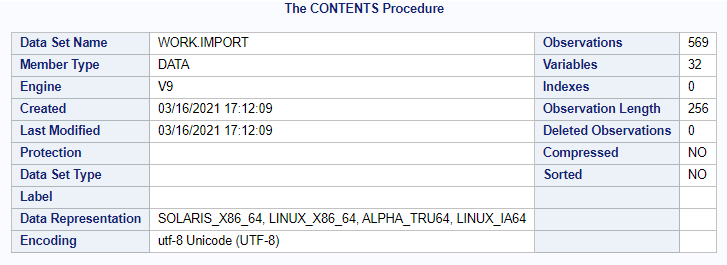
**Code:**

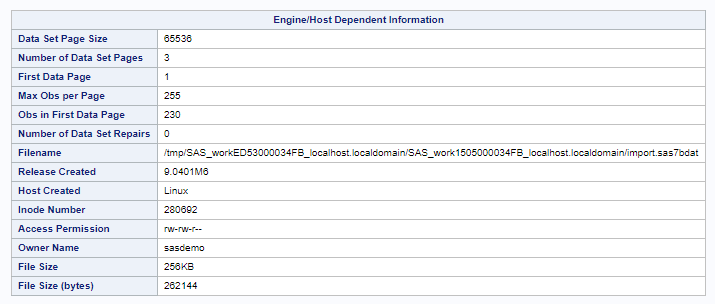


**Output:**



Additional information about the size and dimensions of the dataset are as follows:

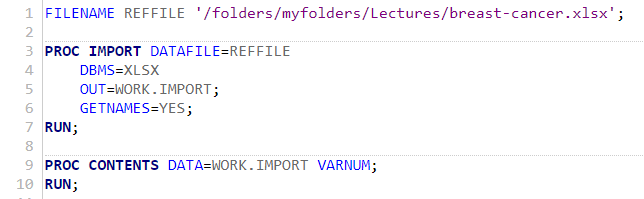




## Q-2. What are the variables (description) used in the dataset?

## **A-2.** There are 32 variables (description) used in the dataset, following is the table showing the name of the variable, it’s data type, Length along with the description of each variable.

**Code:** (breast-cancer file imported and then used PROC CONTENTS for details.)

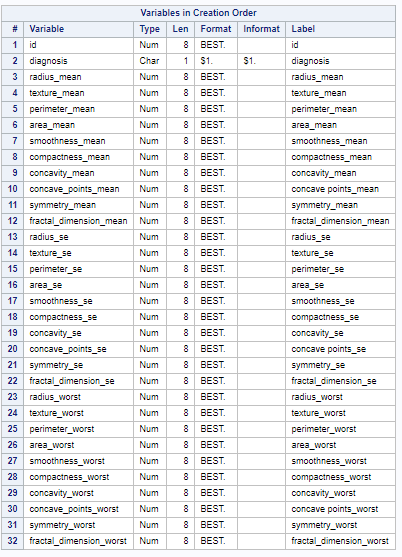


**Output**: (We edited the table from out put removed format, informat and label columns and added Description column.)

(Table is on the next page)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables in Creation Order | | | | |
| # | **Variable** | **Type** | **Len** | **Description** |
| 1 | id | Num | 8 | #ID number |
| 2 | diagnosis | Char | 1 | The diagnosis of breast tissues (M = malignant, B = benign) |
| 3 | radius\_mean | Num | 8 | mean of distances from center to points on the perimeter |
| 4 | texture\_mean | Num | 8 | standard deviation of gray-scale values |
| 5 | perimeter\_mean | Num | 8 | mean size of the core tumor |
| 6 | area\_mean | Num | 8 | mean size of the area of the tumor |
| 7 | smoothness\_mean | Num | 8 | mean of local variation in radius lengths |
| 8 | compactness\_mean | Num | 8 | mean of perimeter^2 / area - 1.0 |
| 9 | concavity\_mean | Num | 8 | mean of severity of concave portions of the contour |
| 10 | concave\_points\_mean | Num | 8 | mean for number of concave portions of the contour |
| 11 | symmetry\_mean | Num | 8 | mean for symmetry |
| 12 | fractal\_dimension\_mean | Num | 8 | mean for "coastline approximation" - 1 |
| 13 | radius\_se | Num | 8 | standard error for the mean of distances from center to points on the perimeter |
| 14 | texture\_se | Num | 8 | standard error for standard deviation of gray-scale values |
| 15 | perimeter\_se | Num | 8 | standard error for the size of the core tumor |
| 16 | area\_se | Num | 8 | standard error for the area of the tumor |
| 17 | smoothness\_se | Num | 8 | standard error for local variation in radius lengths |
| 18 | compactness\_se | Num | 8 | standard error for perimeter^2 / area - 1.0 |
| 19 | concavity\_se | Num | 8 | standard error for severity of concave portions of the contour |
| 20 | concave\_points\_se | Num | 8 | standard error for number of concave portions of the contour |
| 21 | symmetry\_se | Num | 8 | standard error for symmetry |
| 22 | fractal\_dimension\_se | Num | 8 | standard error for "coastline approximation" - 1 |
| 23 | radius\_worst | Num | 8 | "worst" or largest mean value for mean of distances from center to points on the perimeter |
| 24 | texture\_worst | Num | 8 | "worst" or largest mean value for standard deviation of gray-scale values |
| 25 | perimeter\_worst | Num | 8 | "worst" or largest mean value for size of the core tumor |
| 26 | area\_worst | Num | 8 | "worst" or largest mean value for area |
| 27 | smoothness\_worst | Num | 8 | "worst" or largest mean value for local variation in radius lengths |
| 28 | compactness\_worst | Num | 8 | "worst" or largest mean value for perimeter^2 / area - 1.0 |
| 29 | concavity\_worst | Num | 8 | "worst" or largest mean value for severity of concave portions of the contour |
| 30 | concave\_points\_worst | Num | 8 | "worst" or largest mean value for number of concave portions of the contour |
| 31 | symmetry\_worst | Num | 8 | "worst" or largest mean value for symmetry |
| 32 | fractal\_dimension\_worst | Num | 8 | "worst" or largest mean value for "coastline approximation" - 1 |

Additional Information on the Variables from SAS (Original Output):



## Q-3. In the dataset which variable do you think will be used as a predicting variable for any business analytics project?

## **A-3.** In any data set the variable which is useful to predict the solution of underlying problem is considered as predicting variable, For example in property price prediction in future on the basis of various surrounding factors, the main variable to be predicted is the **Price,** The value of which is **dependent** on the other various independent or interdependent variables is our **predicting variable.**

## The Predicting variable for our Dataset – Breast Cancer is the variable **‘Diagnosis”.**

The values of the variable are either M – Malignant or B – Benign.

If the Diagnosis is M – Malignant, it means that the affected cells in the body are cancerous.

If the Diagnosis is B – Benign, then it means that the tumor in the body is not cancerous.

Since it is used to determine whether the person is affected with cancer or not, the variable “**Diagnosis”** is the **predicting variable.**

## Q-4 Write 4 types of questions that you can extract from the dataset that adds business value in the project.

A-4 We can extract answers of following types of questions.

### First Question:

How essential or relevant are the features such as Smoothness and Texture in analyzing the type of tumor as either benign or malignant?

### Second Question:

How easy is it to detect accurately in early stages and what are the complications with time?

### Third Question:

Can tumor size and dimensions be relied upon to detect the malignancy?

### Fourth Question:

Since the inaccurate prediction would lead to delay in treatment and result in further complications, what is the ideal accuracy that needs to be reached for a model to be considered efficient in prediction of breast cancer?

## Q-5 Define a proposal using the chosen dataset.

**A-5** According to the global statistics, Breast Cancer is one of the most common cancers and second leading cause of cancer death among women, posing to be a significant public health problem. This cancer occurs due to the abnormal growth of the cells in breast tissue which is called as a tumor. A tumor can be classified as either malignant (cancerous) or benign (non-cancerous). If the breast cancer is detected at an early stage, it can significantly improve the chances of treatment and survival. Also if accurately diagnosed, it can help patients from undergoing unnecessary treatments or risking their lives.

The current dataset is rich with different variables that are derived based on the lumps detected from fluid samples taken from breast cancer patients. There are primarily ten features from each one of the cells in the sample, then it calculates the mean value, extreme value and standard error of each feature.

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 and field 23 is Worst Radius.

We aim to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in our work. For doing so, data preparation is a very crucial step for our analysis. Data cleaning is used to remove noisy data and missing values. Further categorizing the data, normalizing it, eliminating redundancy and incongruent data to use it for our analysis. The data set chosen has rich data to help in our project and hence it is proposed to be the dataset we would base our project on.