Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(4), 1057-1063 | ISSN:2251-6727

Interaction Among Parameters of Breast Cancer at Wisconsin Diagnostic Centre Using Large Data

Alimi Jamiu[a], Michael Omonkheoa Oyakhire[b], Okoseimiema Sonny Clement[b]

[a]*Resource Treatment Centre 1404 S State Ave, Indianapolis, IN 46203,*
Jamiu.alimi@gmail.com.

[b]*Department of Anatomy, Faculty of basic Medical Sciences, University of Port Harcourt, Port Harcourt, Nigeria.*

(**Received**: 14 April 2024 **Revised**: 01 May 2024 **Accepted**: 18 June 2024)

KEYWORDS
Interaction, malignant, benignant, binary regression and large data

ABSTRACT:
This study applied a binary logistic regression model to determine the effect among some selected variables of breast cancer data collected from Wisconsin (diagnostic data). The data set is made of 569 observations and 25 variables. In this research, we examined the above mentioned data using python programming. The critical examination of the chart of the two diagnosis stages malignant and benign, revealed high levels of benign and low levels of malignant among the patients diagnosed with breast cancer. The covariance matrix between the variables shows a strong positive relationship between the variables with correlation value of +1. The normal distribution curve was determined in two stages with outliers and without outliers. The Coefficient of Determination of the binary logistic regression model (R2) is 0.556. This implies that 55.7% of a high degree of statistical significance emphasizes how well the model performs based on the predictors distinguishing between some selected variable of breast cancer. The binary logistic regression model supports the claim that specific tumour characteristics correlate highly with the tumour's diagnosis. These discoveries have the potential to improve oncology predictive modelling and clinical decision-making, which could improve patient outcomes and diagnostic accuracy.

1.1 Introduction

Breast cancer is a disease that spreads around the world. where abnormal breast cells proliferate and develop into tumours. Tumours have the potential to spread throughout the body and become lethal if ignored. Inside the breast's milk ducts and milk-producing lobules are where breast cancer cells first proliferate. The earliest form is detectable in its early stages and is not life-threatening. Invasive cancers can spread to adjacent lymph nodes or other organs. [9]. Early detection and screening provide the best chance for a decrease in breast cancer mortality, particularly when paired with appropriate therapy. It is critical in these circumstances to be able to identify a tumour's kind swiftly and clearly. This enables medical professionals to assist in the selection and prescription of treatment from the earliest phases of their growth. [1]. The goal of screening for breast cancer is to identify the illness in people who are asymptomatic at a pre-clinical stage to lower the death rate and improve prognostic survival curves while avoiding drastic and costly procedures that could lower the patient's quality of life.[12]

Mammography is the main imaging-based diagnostic used in clinical practice and is regarded as the gold standard for detecting abnormalities of the breast, including lesions that raise suspicion of breast cancer. As a screening test focused on public health, mammography is meant to be affordable, evidence-based, and widely available. Socioeconomic position, ethnicity, and health insurance coverage were found to be significant predictors of the absence of primary screening and follow-up visits in earlier research. [5]. Cultural factors and ignorance among women also play a role in determining the significance, benefits, and likelihood that routine screenings will find breast abnormalities early on, leading to longer life expectancies and survival curves. [10]

In the year 2020, 2.3 million women worldwide had a breast cancer diagnosis in 2022, and 670,000 people died from the disease. Across the world, breast cancer affects women at any age after adolescence, however, its prevalence rises with age. Global estimations show stark differences in the incidence of breast cancer based on human development. For example, in nations with a very high Human Development Index (HDI), 1 in 12 women may receive a breast cancer diagnosis during their lifetime, and 1 in 71 will pass away from the disease. In comparison, 1 in 48 women will die from breast cancer in nations with a low HDI, even though only 1 in 27

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(4), 1057-1063 | ISSN:2251-6727

women will receive a diagnosis of the disease throughout their lifetime. [9]

The strongest risk factor for breast cancer is female sex. Women get breast cancer in about 99% of cases, while men get breast cancer in 0.5–1% of cases. The care of breast cancer in men is based on the same concepts as in women. [10] Ageing, obesity, heavy alcohol consumption, radiation exposure history, family history of breast cancer, reproductive history (including age at first pregnancy and menstruation onset), tobacco use, and postmenopausal hormone therapy are some of the factors that raise the risk of breast cancer. Although the majority of women diagnosed with breast cancer have family history of the disease, having a family history of breast cancer enhances one's chance of developing the disease. A woman's risk does not always decrease if her family history is unknown. The most common hereditary high penetrance gene mutations that significantly raise the risk of breast cancer are those found in the BRCA1, BRCA2, and PALB-2 genes. Women who have mutations in these key genes may want to think about chemoprevention or surgically removed both breasts as risk reduction measures [11].

2.0 MATERIALS AND METHODS

Nevertheless, research has demonstrated that, even in the relatively early stages of the disease's development, the diagnosis of cancer's intrinsic character is not always accurate. This incorrect diagnosis results in insufficient therapy, which raises the disease's mortality rate. Hence, the largest difficulty is to develop instruments for the accurate and timely identification of tumour types. Statistical and mathematical modelling methods, in addition to mammography examinations, can assist us in this procedure. Several models have previously been put out to explain how breast cancer develops. [4]In the research, we are going to study the interaction between the parameters of breast cancer in application to the machine learning process using python programming.

2.1 Linear Regression Models

The regression model studies the relationship between the variable of the interest y and one or more explanatory variables $X^{(j)}$

The general models

$$Y_i=h\left(X_1^{(1)}, X_{2,}^{(2)} X_3^{(3)} \dots\dots\dots X_i^{(m)}: \emptyset_1, \emptyset_2, \dots\dots\dots \emptyset_p\right)+\varepsilon_t$$

Here, h is an appropriate function that depends on the explanatory variable and parameter that we want to summarize with vector $\underline{X}=\left\{X_1^{(1)}, X_{2,}^{(2)} X_3^{(3)} \dots\dots\dots X_i^{(m)}\right\}^T$ and $\underline{\emptyset}=\left\{\emptyset_1, \emptyset_2, \dots\dots\dots \emptyset_p\right\}^T$. The unstructured deviations from the function h are described via the random errors ε_t. The normal distribution is assumed for the distribution of this random error, so $\varepsilon_t \sim$ N $(0,\sigma^2)$ independent [8].

2.3 Binary Logistic Regression Model

In a generalised linear model that fits data that has a binary outcome. if the data has multiple classes the logistic regression generalizes into a multinomial regression model. [3]

$$pr(Y_i = 1/X_i) = \pi_i$$

$$pr(Y_i = 0/X_i) = 1 - \pi_i$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right)= \widehat{\beta_0}+ \widehat{\beta_1 X_i}$$

$$\pi_i=\frac{exp(\widehat{\beta_0}+ \widehat{\beta_1 X_i})}{1+exp(\widehat{\beta_0}+ \widehat{\beta_1 X_i})}$$

2.4 Multicollinearity problem in binary regression

The multicollinearity problem is defined as the association between two or more explanatory variables through a strong linear relationship in which the effect of the dependent variables cannot be separated from that of the explanatory variables**.** The problem of linear multicollinearity is also described through the concept of "orthogonality": when the explanatory variables are orthogonal (not linked to each other), all the eigenvalues are equal to one; if one of these eigenvalues is less than one, especially when it is equal to or near zero, than it is not orthogonal, leading to the problem of linear multicollinearity. Linear multicollinearity is considered to be of two types:[7]

3.6.1 The Prefect Multicollinearity*:* In this type of linear multicollinearity, the specific matrix of information (XX) has a determinant equal to zero (|XX́ | = 0) and thus we cannot find the estimators of the general linear regression model since it is not possible to find the inverse matrix.

3.6.2 Semi-Multicollinearity*:* It is the subject of the research in which the value of a specific determinant of

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(4), 1057-1063 | ISSN:2251-6727

the matrix of information ($X\acute{X}$) is very small and close to zero: $|X\acute{X}| \approx 0$.

Therefore, the estimators of the parameters can be found, but the consequence of this type of

multicollinearity is when estimates are inaccurate, and the estimated differences in parameters are very large.[7]

3.7.2. Variance Inflation Factor

Proposed by Farrar and Glauber the Variance Inflation Factor (VIF) measures the inflation of the parameter estimates and is computed for all explanatory variables in the model.

The VIF formula is as follows:

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

$$j = 0,1,2 \ldots . m$$

$$1 - R_j^2, \text{ j} = 0,1,2, \ldots, \text{m}$$

where R_j^2 is the Coefficient of Determination for the explanatory variable. [7]

The coefficient of determination ranges between 0 and 1 and is calculated in the case of four explanatory variables according to the following formula:

$$R_j^2 = 1 - (1 - r_{12}^2)(1 - r_{12}^2, 3)(1 - r_{14,23}^2)$$

where:

r_{12}^2 Represents the simple correlation coefficient and $r_{12}^2, 3\ and\ r_{14,23}^2$ represents the partial correlation coefficients.

2.5 covariance matrix

Since our observations consist of many observed variables, we employ a random vector of the covariance matrix for the interaction between the parameters in the study. Every element in the combination has an average[6]

$$X = (X_1, X_2 \ldots \ldots \ldots X_n)$$

$$\mu = (E(X_1), E(X_1) \ldots \ldots E(X_n))$$

and a covariance matrix

$$\sigma = \begin{bmatrix} V(X_1) & COV(X_1X_2) & . & . & COV(X_1X_n) \\ COV(X_2X_1) & V(X_2) & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ COV(X_nX_1) & . & . & . & V(X)_n \end{bmatrix}$$

With $\sigma_{ij} = \sigma_{ji}$, $\sigma_{11} = var(X_1) = COV(X_1X_1)$

$$P_{11} = \sum_{i=1}^{n} P_{ii} = 1$$

3. Results

The study used Breast Cancer data from Wisconsin (Diagnostic) to identify the level of spread of cancer among the study sample. The UCI machine learning repository received the data set from Wisconsin Madison Hospital Between 1989 and 1991 and collected data on breast cancer in the UCI archive (Mangasarian, et al 2008). In the Wisconsin dataset, there is a unique ID for each of the 569 rows and 25 columns. Aspirating malignant tissue from patient breasts that featured nuclear features that could be evaluated electronically.

Plot of Malignant (M) and Benign (B) Breast Cancer

The chart shows that 62.74% of the patients have benign breast cancer while 37.26% of patients have malignant breast cancer representing the total 569 patients that were diagnosed in the Wisconsin data. These indicate that breast cancer in the patient spreads fast to other part of the body.

Covariance Matrix Breast Cancer

The covariance among the variables shows a strong relationship between the variables with values ranging from -1. to +1. A covariance coefficient of +1 indicates that the two variables are perfectly related in a positive (linear) manner, a covariance coefficient of -1 indicates that two variables are perfectly related in a negative (linear) manner, while a covariance coefficient of zero indicates that there is no linear relationship.

3.1 The Normal Probability Plots Before Outliers Were Removed.

A non-linear regression model requires that the error term for a normal distribution produce unbiased estimates with the minimum variance. However, to satisfy this assumption the researcher performs statistical normality testing. This is determined by the residuals plot of the normal probability plot. The residuals plots of some variables are not normally distributed.

3.2 The normal probability plots of the variables after outliers were removed.

This is determined by the residual plot of the normal probability distribution. The residuals after outliers were removed are normally distributed with mean, mode and median all lying at the centre of the normality plot. This

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(4), 1057-1063 | ISSN:2251-6727

implied that the variables were normally distributed after the outliers were removed.

3.3 Binary Logistic Regression Models Malignant and Benign Breast Cancer

The Coefficient of Determination (R^2) between the binary variables is strong, with a probability value of 0.000 for all the variables. Since the R square is close to 1.Texture_se: 'texture_se' has a coefficient of -2.7262, a p-value of 0.000. This negative coefficient suggests that a higher 'texture_se' value corresponds to a lower likelihood of the tumor being malignant. 'texture_se' is a significant predictor in the model, as indicated by the statistical significance (p-value < 0.05).Area_se: The p-value for 'area_se' is 0.000, and the coefficient is 0.0968. Higher 'area_se' values appear to enhance the risk of the tumor being malignant, according to this positive coefficient. Strong statistical significance highlights how crucial 'area_se' is to the model's prediction process. Fractal_dimension_se: This variable has a substantial negative correlation with the likelihood of malignancy, with a coefficient of -878.8111 and a p-value of 0.000. A strong inverse association is indicated by the huge size of the coefficient, which suggests that greater values of 'fractal_dimension_se' are linked to a significantly lower likelihood of a malignant diagnosis. Concavity_worst:' concavity_worst' has a p-value of 0.000 and a coefficient of 9.8154. The positive coefficient suggests that there is a substantial correlation between larger 'concavity_worst' values and a higher likelihood of the tumor being malignant. The statistical significance demonstrates that 'concavity_worst' is a significant predictor of the diagnosis.

4.0 Conclusion

This study focused on the interaction among the variables of breast cancer in application to Wisconsin (diagnostic data). The data set is made of 569 observations and 25 variables. To determine the inter-relationship between the study variables. The critical examination of the time chart of the two states malignant and benign were plotted. These revealed a high number of benign and low spread of malignant among the patients diagnosed with breast detained in Figure 1. The 25 variables involved in this study were tested for inter-relationships using the covariance matrix detained in Figure 2. The covariance among the variables shows a strong relationship between the variables with value ranging from -1 to +1. The test for statistical significance among the variables was done by the normal probability plot (with outliers and without outliers) in Figures 2 and 3. The residuals after outliers were removed are normally distributed with mean, mode and median all lying at the centre of the normality plot. The binary logistic regression model was used to recast diagnosis as the independent variable and some selected variable such as texture_ se, area_se, fractal_ dimensional radius and Concavity worst. With a pseudo R-squared of 0.5566, the model accounts for roughly 55.66% of the variation in the tumor Given that the binary logistic regression model accounts for a sizable amount of variability in the outcome variable, this suggests a reasonably good fit for the model. Moreover, the remarkably p-values of 0.000 indicates that the predictive model outperforms better. This high degree of statistical significance emphasizes how well the model performs based on the provided predictors in distinguishing between malignant and benign tumours. This logistic regression model's findings offer important new information on the variables affecting breast cancer tumor diagnosis. Tumors with bigger areas and higher degrees of concavity are more likely to be malignant, according to the positive correlations found between 'area_se' and 'concavity_worst' and malignancy. In contrast, there is a negative correlation between 'texture_se' and 'fractal_dimension_se' with malignancy, meaning that a lower probability of a malignant tumor corresponds to larger values of these predictors. This can be highlights as follows: The model is a valuable tool for predicting cancer outcomes based on the identified predictors, as seen by its reasonably high Pseudo R-squared value, which further validates its usefulness in capturing the variance in tumour diagnosis. The binary logistic regression model supports the claim that specific tumour characteristics correlate highly with the tumor's diagnosis. These discoveries have the potential to improve oncology predictive modelling and clinical decision-making, which could improve patient outcomes and diagnostic accuracy. Subsequent investigations may examine supplementary factors or other modelling methodologies to enhance the prediction capability and practicality of these models in clinical settings.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(4), 1057-1063 | ISSN:2251-6727

References

[1] Azamjah, N., Soltan-Zadeh, Y., & Zayeri, F. (2019). Global Trend of Breast Cancer ortality Rate: *Asian Pac J Cancer Prev, 20*(7),

[2] Ahuja, H.A. (2012). *Modern Macroeconomics*. New Delhi: Schan Publishers.

[3] Altunç, O. & Aydin.F (2014). An estimation of the consumption function under the permanent Income Hypothesis: The Case of D-8 countries. *Journal of Economic Cooperation and Development.* 35(3):29-42

[4] Bicar, A. (2018) Le cancer du sein chez la jeune femme et saprise encharge. PhD Thesis, *University of Limoge, Limoge,*

[5] Banks E, Reeves G, & Beral V, (2004) Influence of personal characteristics of individual women on sensitivity and specificity of mammography in the Million Women Study: cohort study. *BMJ*. 2004; 329(7464):

[6] De Juan JP,. & Seater JJ.(1997) A cross-country test of the permanent income hypothesis. *International Review of Applied Economics.* 11(3): 451-468.

[7] Farrar, Donald E., and Robert R. Glauber. (1967) "Multicollinearity in Regression Analysis: The Problem Revisited." *The Review of Economics and Statistics*, vol. 49, no. 1,

[8] Granger, C., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2, 111-120.

[9] Guide to cancer early diagnosis. World Health Organization. https://www.who.int/cancer/publications/cancer_early_diagnosis/en/. Published 2021.

[10] Khan M, & Chollet A.(2021) Breast Cancer Screening: Common Questions and Answers. *I am a famous physician*. 103(1):33-41.

[11] Tabár L, Chen TH, Yen AM, (2021). Early detection of breast cancer rectifies inequality of breast cancer outcomes. *J Med Screen*. 28(1).

[12] Raherinirina, A.,Randriamandroso, A., Hajalalaina, A.R., Rakotoarivelo, R.A. and Rafamatantantsoa, F. (2021) A Gaussian Multivariate Hidden Markov Model for Breast Tumor Diagnosis. Applied Mathematics , 12, 679-693.

Table

Table 3.1. Binary Logistic Regression Models Malignant and Benign Breast Cancer

Dep. Variable:	diagnosis	No. Observations:	589
Model:	MNLogit	Df Residuals:	565
Method:	MLE	Df Model:	3
Date:	mon, 17 Jun 2024	Pseudo R-squ.:	0.5566
Time:	07:20:03	Log-Likelihood:	-166.60
converged:	true	LL-Null:	-375.72
Covariance Type:	nonrobust	LLR p-value:	2.467e-90

	coef	std err	z	P>|z|	[0.025	0.975]
Texture_se	-2.7262	0.357	-7.646	0.000**	-3.425	-2.0.27
Area_se	0.0968	0.011	8.825	0.000**	0.075	0.119
Fracta_dimension_se	-878.8111	131.8279	-6.669	0.000**	-1137.08	-620.524
Concavity_worst	9.8154	1.233	7.959	0.000**	7.398	12.233

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(4), 1057-1063 | ISSN:2251-6727

Footnote: ** = sig. At 5%

Figures

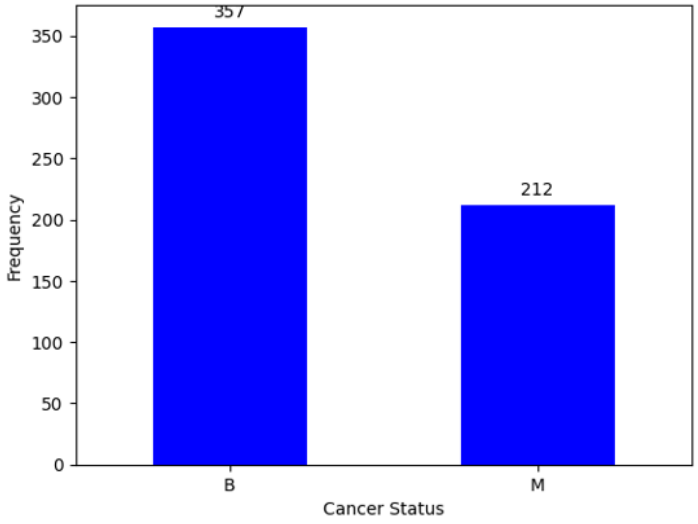

Figure 3.1: Plot of Malignant (M) and Benign (B) Breast Cancer

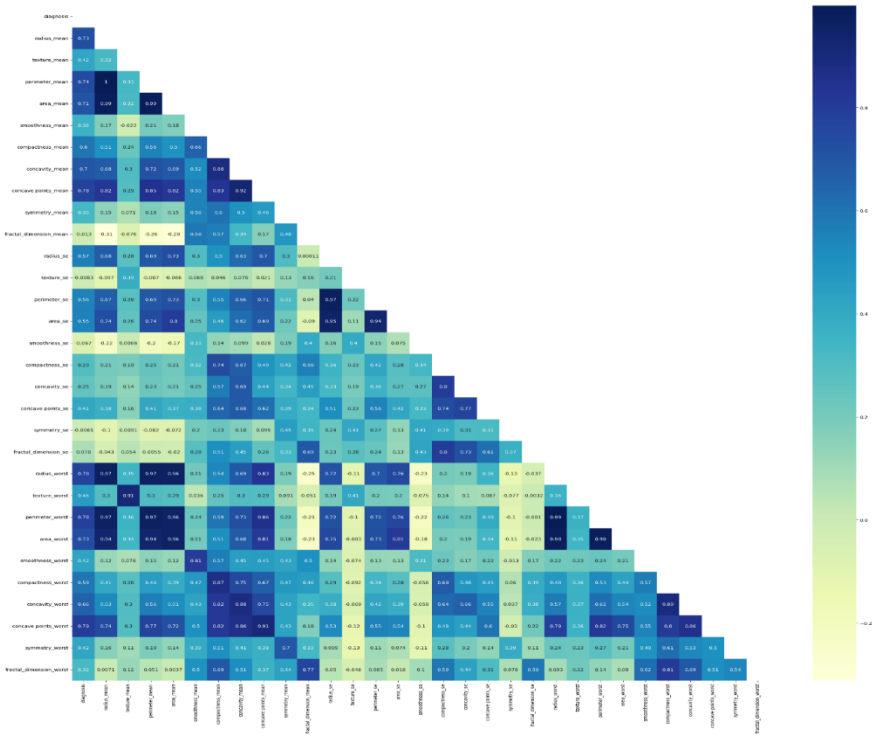

Figure 3.2: Covariance Matrix Breast Cancer

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(4), 1057-1063 | ISSN:2251-6727

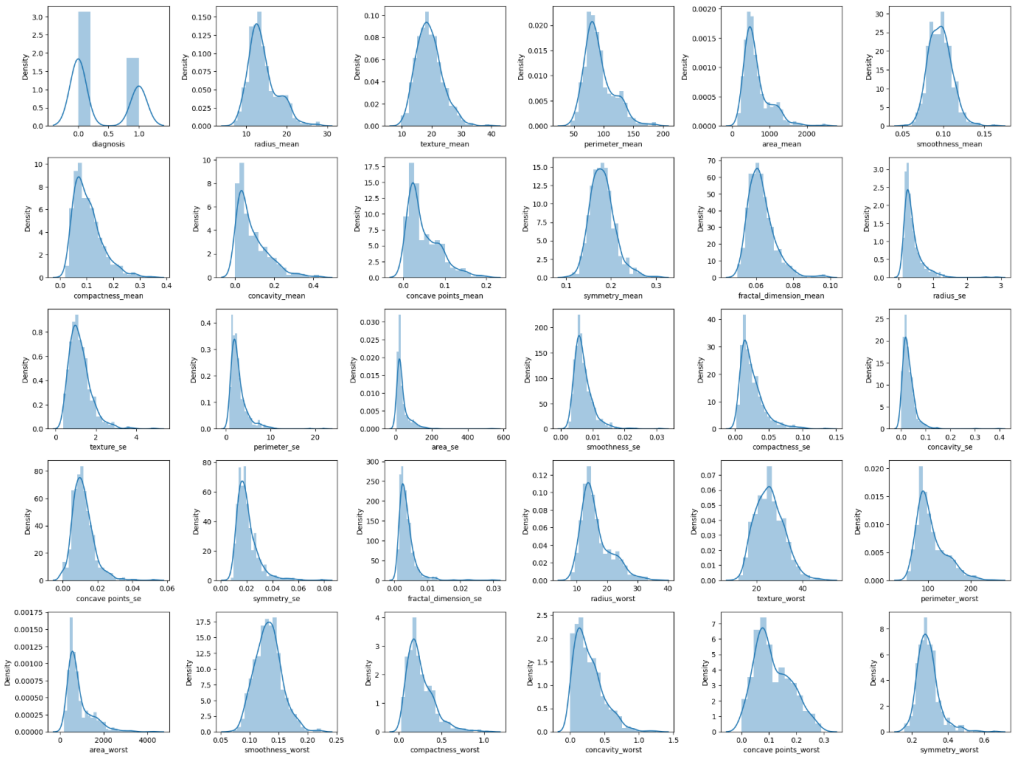

Figure 3.3: The Normal Probability Plots Before Outliers Were Removed.

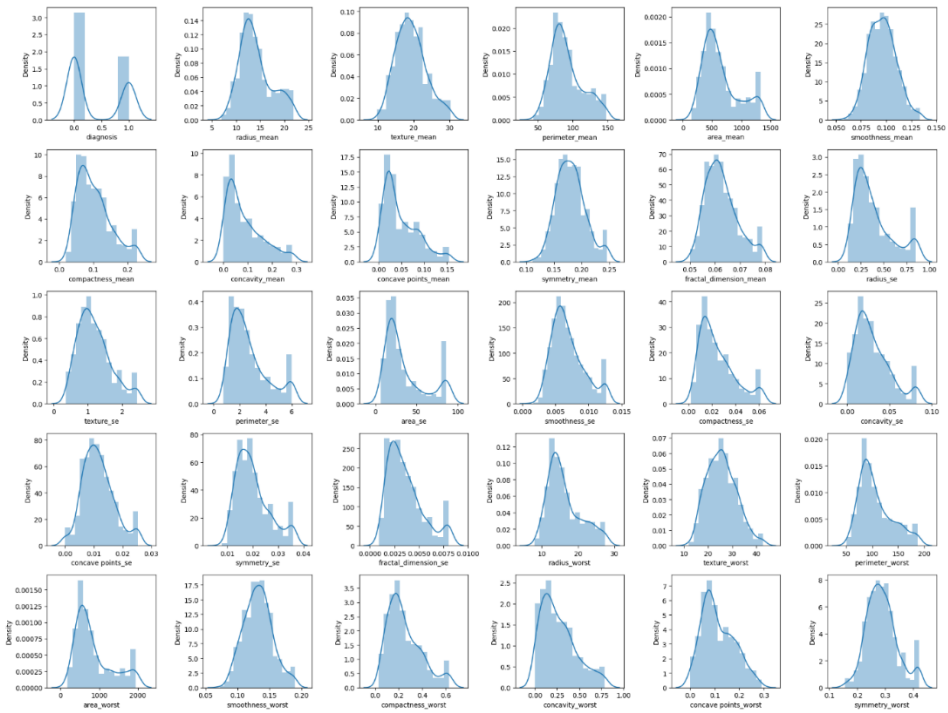

Figure 3.4: The Normal Probability Plots after Outliers Were Removed.