EXECUTIVE SUMMARY

Breast cancer remains a prevalent global concern, affecting millions annually, necessitating ongoing research into its etiology, risk factors, and treatment modalities. Early detection is imperative, as delayed diagnoses often result in more complex treatments and decreased survival rates. Thus, there's a pressing need to enhance early detection methods to mitigate the disease's impact worldwide. Our study aims to contribute to this global effort by leveraging data analysis techniques to improve breast cancer diagnosis.

Following a review of relevant literature, we formulated research questions centered on the predictive utility of various clinical features in diagnosing breast cancer. Subsequently, we explored the Breast Cancer Wisconsin (Diagnostic) dataset, comprising 569 samples from the University of Wisconsin Hospital, to address these inquiries.

Methodologically, our study employed a multifaceted approach, incorporating Principal Component Analysis (PCA), Confirmatory Factor Analysis (CFA), Canonical Correlation Analysis (CCA), and Automatic Variable Selection Regression. These techniques enabled comprehensive exploration of the dataset, unveiling patterns, validating groupings, and elucidating interrelationships among clinical variables. Notably, Ridge Regression facilitated the development of an accurate predictive model, achieving a commendable 97.25% accuracy in diagnosing tumor malignancy.

Our findings underscored the significance of tumor size, shape, and surface characteristics in distinguishing between malignant and benign cases. Variables such as radius, perimeter, area, concavity, and fractal dimension emerged as pivotal predictors of tumor malignancy, underscoring the importance of considering diverse features in diagnostic assessments.

However, the study is not without limitations. Firstly, the dataset's exclusive reliance on samples from Wisconsin raises concerns regarding its generalizability to other regions. Future research should encompass broader geographic representation to enhance the robustness of findings. Secondly, the absence of patient demographic information limits our understanding of how factors like age, sex, and diet influence tumor characteristics. Incorporating these variables in future investigations would provide a more comprehensive perspective on breast cancer development.

Moreover, the utilization of data from 1995 necessitates periodic reassessment to account for advancements in cancer detection and related fields. Finally, our study underscores the potential of analyzing tumor shapes like fractal dimension and smoothness in refining diagnostic accuracy, empowering medical practitioners to tailor treatment plans to individual patient needs.

In summary, our research underscores the critical role of data analysis techniques in advancing breast cancer diagnosis and underscores the need for ongoing interdisciplinary efforts to combat this pervasive disease.

Abstract

The early diagnosis of breast cancer plays a crucial role in improving treatment outcomes and enhancing survival rates. This study employed various statistical techniques to analyze clinical features and identify significant attributes contributing to the diagnosis of breast tumors. The methods utilized include Principal Component Analysis (PCA), Confirmatory Factor Analysis (CFA), Canonical Correlation Analysis (CCA), and regression techniques such as forward, backward, stepwise, and ridge regression. However, CCA revealed a strong correlation between features of size and shape clinical groups which can be significant in predicting breast cancer diagnosis, while PCA effectively distinguished components representing tumor size, shape, and surface characteristics. CFA further extracted six significant factors related to size, shape, texture, variability, and symmetry. Automatic variable selection regression methods identified key predictor variables associated with tumor shape, such as fractal dimension, smoothness, and worst fractal dimension, contributing to an accurate prediction model. Ridge regression achieved an impressive 97.25% accuracy in predicting tumor malignancy, demonstrating its robustness. The findings highlight the importance of tumor shape characteristics in diagnosing breast cancer and provide valuable insights for early detection and effective treatment strategies.

Introduction

Breast cancer has always been a significant health concern which affects millions of individuals worldwide. It is the most common cancer diagnosed among women and a leading cause of cancer-related mortality. Early detection and accurate diagnosis play a crucial role in improving treatment outcomes and enhances the survival rates for patients with breast cancer. However, the existing screening methods have limitations, and there is a need for more reliable and accurate tools for diagnosis. This study focuses on investigating the clinical features associated with breast tumors and their relationship with malignancy.

This research's importance cannot be overlooked, as breast cancer has serious implications for individuals, families, and society. Breast cancer is not just a significant threat to physical health but also has psychological, emotional, and economic impacts on the individuals affected and also their loved ones. With the help of this study, we could potentially help and enhance the early detection strategies and also reduce the morality rates.

Literature Review

Breast cancer stands as the second most common cancer globally, posing a significant health threat, particularly affecting one in eight females, with higher risks after the age of 40. The studies show that different techniques were applied in the past to predict the tumor to reduce risks of mortalities. (Tsehay Admassu Assegi, 2021). Various studies have focused on selecting features that distinguish between benign and malignant tumors, aiming to create an effective system for recognizing breast cancer. (Sohaila Rehman, 2013). The major problem in breast cancer prediction with machine learning is the imbalance between the benign and malignant observations in breast cancer dataset (Tsehay Admassu Assegi, 2021)

Cancer disease cases are increasing rapidly, and machine-learning algorithms are required for decision support to reduce the epidemic cases by predicting breast cancer as early as possible. The research work applied different machine learning algorithms to develop a predictive model for the classification of breast cancer. Some of the previous research works on breast cancer classification. (Tsehay Admassu Assegi, 2021). In addition, several studies had been conducted to analyze breast cancer survivability through data mining methods, majority of studies considered patients that were alive after 5 years since the first diagnosis as having survived the cancer. Prediction accuracies are generally lacking although some factors are known to greatly affect the prediction accuracy. (Nagesh Shuklaa, 2018)

Previous literature shows, breast utility instrument was developed using CFA model The factors considered included comprehensive coverage of relevant items, prioritization of patient-rated important dimensions, removal of overlapping dimensions, exclusion of low-consistency and low-importance dimensions, prioritization of patient experience in sex-related dimensions, and re-specification of models based on clinical relevance. (Teresa C. O. Tsui, 2022). Another study suggests breast cancer classification using machine learning methods, including Support Vector Machines, Logistic Regression, K-Nearest Neighbour, Decision Trees, and Naive Bayes. This study highlights the identification of crucial features to enhance the understanding of predictive models applied in breast cancer diagnosis. (Farahnaz Sadoughi, 2016).

This study addresses a critical question: which traits are most predictive of breast cancer presence? To enhance breast cancer diagnosis and maybe save lives. Looking into various traits like size, texture, smoothness, concavity, fractal dimension etc, we have tried to determine whether a detailed study of these traits can help in early screening and establishing a better treatment plan for the patient. Using a wide range of statistical tools, we hope to clarify how various approaches rank these important characteristics in addition to identifying them. Through more specific testing procedures, earlier detection, and better treatment decisions for patients, this comparison analysis can improve diagnostic practices.

Methods

This study is solely based on secondary data where we have used the previous dataset of Wisconsin. The sample size of this dataset is 569 observations, which is good to perform a thorough study with no missing values. The variables of this dataset are shown in table 1.

We started by employing Canonical Correlation Analysis (CCA) to explore associations between tumor shape variables and tumor size variables (radius and area). Given the significance of the shape variables in predicting the malignancy, we aimed to uncover any high correlations between size and shape features.

Principal Component Analysis (PCA) was then applied to understand the underlying structure and relationships between size and shape variables post-CCA. This helped us in identifying distinct factors representing cell nuclei size, its shape and surface characteristics for dimensionality reduction and variable exclusion. This step was crucial in simplifying the dataset while preserving key information.

Following this a Confirmatory Factor Analysis (CFA) was utilized to extract significant features and identify groupings of variables explaining the most variance in the dataset. This helped us in determining the predictive power of the shape variables in determining tumor malignancy. This process involved identifying clusters of variables that explained most of the dataset’s variance.

Upon identifying relevant features, we pruned the dataset by eliminating highly correlated variables. Subsequently, we employed ridge regression, leveraging its ability to handle multicollinearity, to predict tumor malignancies. This regression technique allowed us to utilize the remaining eight variables, which included metrics such as mean symmetry, mean fractal dimension, and various standard errors.

We used automatic variable selection methods, such as forward, backward, and stepwise selection, to optimize our predictive models. By methodically identifying the most informative variables, these techniques improved the predicted accuracy and model refinement.

Our thorough investigation attempted to clarify the complex connection between tumor features and malignancy prediction, which we hope will advance knowledge of breast cancer prognosis.

Results and Discussion

Canonical Correlation Analysis (CCA):

The insights gained from CCA can inform subsequent steps in predictive modeling. Features like smoothness, concavity, concave points, fractal dimension, compactness, and symmetry are responsible for expressing the shape of the nucleus, therefore taken into consideration to find out their shared variance and contribution to capturing meaningful patterns or differences between groups.

The canonical correlation for the first variate is found to be 0.99. Since the first function explains most of the information, it means that clinical_group1 and clinical_group2 share 99.1% of information, indicating a significant amount of shared variance. Variables tend to be highly correlated with each other, suggesting a linear combination of features in clinical_group1 and a linear combination of features in clinical_group2.

For CV 1, the test results indicate a very high correlation (rho^2 = 0.99943) and a highly significant Chi-Squared statistic (Chisq = 4334.08) with 16 degrees of freedom. The p-value (Pr(>X)) is extremely low (less than 2.2e-16), denoted by '***,' suggesting a strong rejection of the null hypothesis. This implies that there is a significant relationship between the first canonical variate (CV 1) and the set of variables included. Similarly, for CV 2, the results show a correlation (rho^2 = 0.21047) and a Chi-Squared statistic (Chisq = 132.93) with 7 degrees of freedom. Again, the p-value is extremely low, indicating a highly significant relationship between the second canonical variate (CV 2) and the set of variables.

In summary, the statistically significant results of Bartlett's Chi-Squared Test suggest that both CV 1 and CV 2 are strongly associated. The high correlation values and low p-values provide evidence of the meaningfulness of these relationships, emphasizing the importance of these canonical variates in capturing patterns and relationships within the breast cancer dataset. However, we have considered variate 1 because it explains most of the information for our study.

Table 2 For Clinical Group 1, the CV1 has coefficients of 0.2675, 0.0001646 suggesting a positive association with the radius and area included in this group and their contribution to CV1. For Clinical Group 2, the coefficients for CV1 include several values: 9.7903e-05, 0.0434, 3.3983, -1.4522, -0.1777, -0.0508, 0.0695, and 1.4252. These coefficients represent the weights assigned to different variables in the group. The subsequent coefficients (0.0434, 3.3983, -1.4522, -0.1777, -0.0508, 0.0695, 1.4252) correspond to the variables contributing to CV1 in Clinical Group 2. Furthermore, the loadings of variables in both sets (clinical_group1 and clinical_group2). Both Radius and Area are highly important variables with values 0.99 and 0.98, respectively. In group 2, Perimeter (0.99) and Concave Points (0.82) are the most important variables in the first function of the canonical correlation. This indicates that these variables contribute significantly to the canonical variates. Hence, these variables might be important features for predicting tumor malignancy. Results reveal that these features align in characterizing the shape and size of nuclei for breast cancer diagnosis.

Figure 1 shows the Helio plot of the first canonical variate. As we can visualize, the radius and area of the first clinical group are highly correlated. However, in clinical group 2, perimeter is highly correlated, followed by concave points, concavity, and smoothness. Furthermore, all variables are positively correlated except for fractal dimension, which seems to be negatively correlated. This means that these variables can be considered in doing further analysis of breast cancer tumors.

Principal Component Analysis (PCA):

Principal Component Analysis (PCA) was conducted here to help us identify the distinct factors representing the cell nuclei size, shape, and surface characteristics. The PCA revealed that the first two principal components (PC1 and PC2) explain a significant portion (around 63%) of the total variance in the data.

The first principal component (PC1), accounts for approximately 44.27% of the total variance, primarily represents the tumor size and morphology. Variables such as radius_mean, perimeter_mean, area_mean, and radius-related features loaded heavily on this component. These variables are associated with the overall size and shape of the tumor.

The second principal component (PC2), explaining approximately 18.97% of the variance, represents the texture versus smoothness characteristics of the tumor. Variables like texture_mean, compactness_mean, concavity_mean, concave.points_mean, and fractal_dimension_mean loaded heavily on this component. These features capture the surface texture, irregularities, and fractal properties of the tumor.

The PCA results tell us that the two major underlying components that explain a substantial portion of the variance in the dataset are tumor size/morphology (PC1) and texture/smoothness (PC2). These findings also align with the clinical understanding that the size, shape, and surface characteristics of tumors are crucial factors in distinguishing between benign and malignant cases.

Figure 2, Scree plot helps in identifying the appropriate number of principal components to retain based on the "elbow".

In the scree plot, the first principal component exhibits the highest eigenvalue of around 13.3, accounting for approximately 44.3% of the total variance in the dataset. The second principal component has an eigenvalue of around 5.7, explaining an additional 19% of the variance. Together, these two principal components capture a substantial portion (approx. 63.3%) of the total variance present in the data.

This analysis suggests that the two main sources of variation in the cancer data are tumor size and morphology (PC1) and texture versus smoothness characteristics (PC2).

Confirmatory Factor Analysis (CFA):

Confirmatory Factor Analysis (CFA) was conducted to verify the factors and fit their variables based on Principal Component Analysis (PCA) and check their statistical significance.

Five factors were deemed as sufficient to proceed with further analysis, as mentioned before in the PCA results.

Due to the variance size in variables, the dataset of the variables used was scaled, reducing the variance. Previously when conducting CFA on the dataset before scaling, Standard Error and P-values were deemed as N/A and null, therefore, scaling was necessary to proceed.

Figure 3 shows that statistical significance of all variables included in the factors determined by PCA analysis. Factor 1 (F1) and Factor 2(F2) are both strong factors to use for future analysis.

Figure 4 shows that there is strong relationship between Factor 1 (F1) and Factor 2 (F2), meaning these two factors are associated which each other when determining the result.

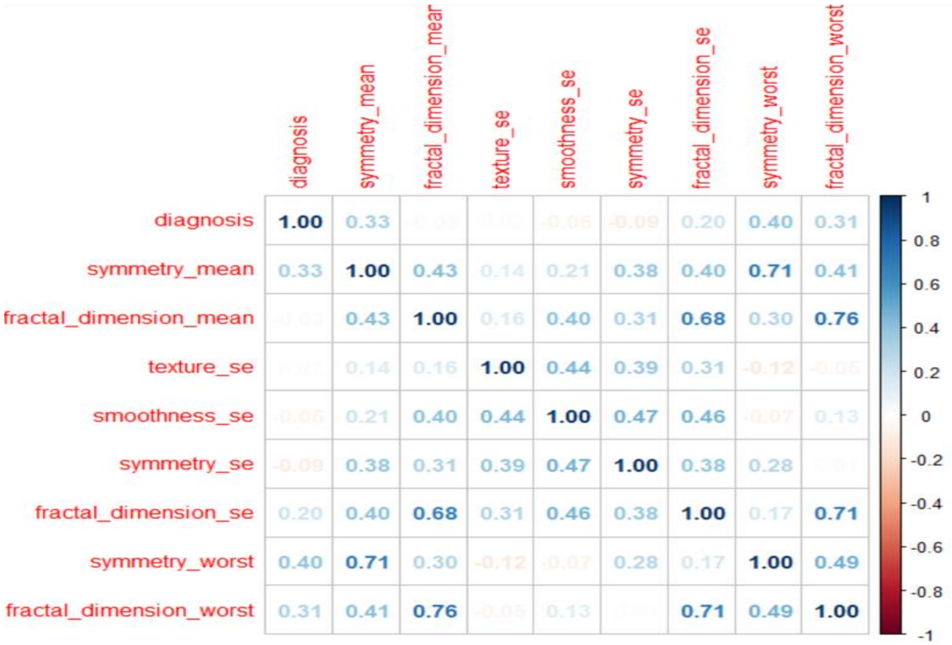
Figure 5 suggests a positive variance and high variance in Factor 1 (F1) and Factor (F2), which relates to a high amount of variability. Since there is a high variance, it shows that these factors can be relied on to explain the data well.

Based on the CFA that was conducted, it can be confirmed that the variables in the factors are valid and have statistical significance. PCA analysis provided the factors determining and CFA validated that these factors work. Factor 1 (F1) and Factor (F2) are strongly associated with each other. The variables radius_mean, perimeter_mean, and area_mean have a strong influence on Factor 1.

Factor 1 includes variables that relate to size and Factor 2 includes variables that are related to texture, this connects back to the Canonical Correlation Analysis. These two factors have a high influence on determining malignancy in tumors.

Automatic Regression Methods:

After removing variables which were experiencing multicollinearity in the dataset, we were left with 8 independent variables with correlations less than 0.8, which did not have VIF values of greater than 10. A correlation matrix of the left-over variables in the dataset is included below:



Three automatic variable selection regression methods were utilized in order to investigate whether the same groups of variables revealed to accurately predict tumor malignancy during literature search, group of variables revealed by the PCA to explain the most variance in the data, and the groups of variables in the most important features found in the CFA were comparable or the same. What was found was that all three automatic variable selection procedures resulted in the same exact final model, with the same variables. In all three instances, all variables left in the final model were variables which measured or described the tumor cells' shape. A summary of the same final model constructed by all three variable selection techniques is provided below.

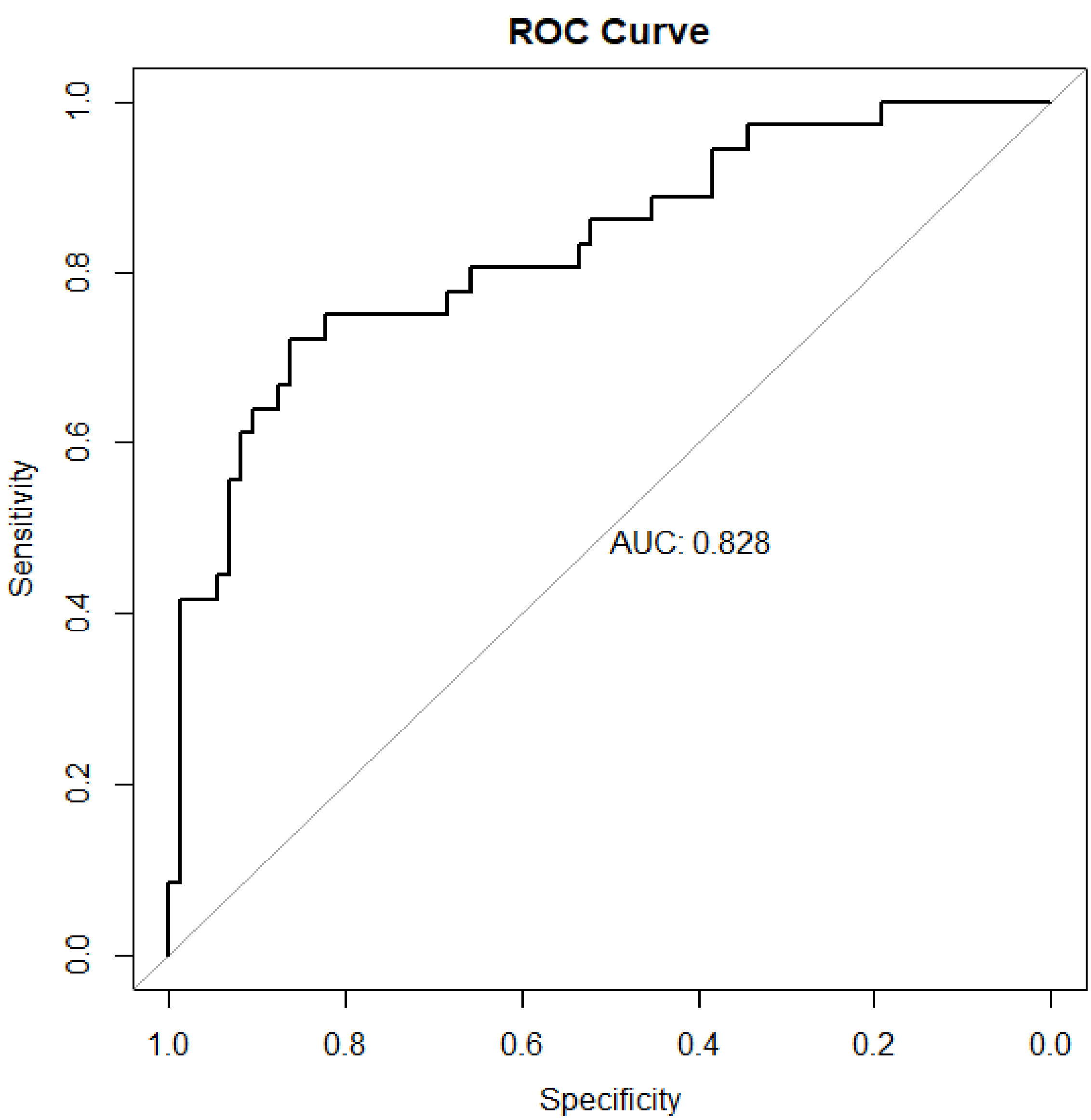
```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.08203 -0.28621 -0.07648  0.31071  1.04813 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.89715    0.20530   4.370 1.54e-05 ***
symmetry_mean             4.57584    1.02571   4.461 1.03e-05 ***
fractal_dimension_mean  -57.44447    5.05536 -11.363  < 2e-16 ***
texture_se                0.10530    0.03824   2.754 0.006132 ** 
smoothness_se            27.62153    7.63054   3.620 0.000328 ***
symmetry_se              -6.27591    3.33983  -1.879 0.060873 .  
symmetry_worst            1.32369    0.55676   2.377 0.017846 *  
fractal_dimension_worst  20.04645    2.02733   9.888  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3733 on 452 degrees of freedom
Multiple R-squared:  0.4203,	Adjusted R-squared:  0.4113 
F-statistic: 46.81 on 7 and 452 DF,  p-value: < 2.2e-16
```

The model is shown to have an adjusted R-squared value of 0.4113, meaning that 41.13% of the variability of the dependent variable is explained by the independent variables. The coefficients reveal that three variables are very significant in the prediction of tumor malignancy in the regression model, variables: mean fractal dimension, smoothness standard error, and worst fractal dimension. Mean fractal dimension had an estimated coefficient of –57.44, showing that the higher the average fractal dimension was for a participant the lower their chance of having a malignant tumor was. Smoothness standard error and worst fractal dimension measures both had positive coefficients respectively of 27.62, 20.05; showing that the higher these variables were for a participant the higher the chance of their tumor being malignant.

To better evaluate the final model produced by the three automatic variable selection methods an ROC curve and confusion matrix were made, included below.



The construction of the ROC Curve suggests that the model produced by these three automatic variable selection methods is good, with an AUC (Area Under Curve) of 0.828, as it can efficiently discriminate between tumors which are malignant verse non-malignant. To further evaluate the predictive accuracy of the model an confusion matrix was made. In order to evaluate the model's performance, since our dependent variable of tumor malignancy is binary, the threshold for converting probabilities of tumor malignancy produced by our regression methods was set to 0.5. If the model predicted that there was greater than a 50.00% chance that a participant had a malignant tumor then the prediction was that they did have a malignant, if the model predicted that there was 50.00% or less chance of a participant having a malignant tumor then the prediction was that the participant had a benign tumor. The confusion matrix is included in the appendices with final model predictability.

The final model produced by the Backward, Forward, and Stepwise variable selection methods predicts tumor malignancy correctly at an overall accuracy of 80.73%. The model correctly identifies true positives or malignant tumors in participants who truly have malignant tumors at a rate of 86.30%. The model correctly identifies true negatives or benign tumors in participants who truly have benign tumors at a rate of 69.44%. While this model was overall a success, we realized that since so many variables were highly correlated and had to be removed all three automatic variable selection methods produced the same result. While an overall accuracy of 80.73% is pretty good, that is still almost 2 out of 10 participants with malignant tumors not being identified as having malignant tumors. Especially in the case of breast cancer research we wanted a more powerful and accurate model, we opted for utilizing ridge regression as it handles multicollinearity well and may be able to construct a more accurate model.

A Ridge Regression model was constructed to utilize as many of the variables in the dataset as possible with the aim of constructing a more accurate model. The coefficients of the final ridge regression model are included below:

```
> coef(best_ridge)
31 x 1 sparse Matrix of class "dgCMatrix"
                                 s0
(Intercept)            -1.172036e+00
radius_mean             1.166786e-02
texture_mean            6.397523e-03
perimeter_mean          1.496133e-03
area_mean              -4.730306e-05
smoothness_mean        -1.960128e-01
compactness_mean       -9.141653e-01
concavity_mean          5.593603e-01
concave.points_mean     1.425290e+00
symmetry_mean          -2.073660e-02
fractal_dimension_mean -6.936930e+00
radius_se               2.626726e-01
texture_se             -4.682743e-03
perimeter_se            8.510589e-03
area_se                -1.379218e-03
smoothness_se           9.218512e+00
compactness_se         -1.885420e+00
concavity_se           -1.681033e+00
concave.points_se       5.228865e+00
symmetry_se             1.151079e+00
```

```
fractal_dimension_se      1.337906e+00
radius_worst              1.506985e-02
texture_worst             7.665641e-03
perimeter_worst           1.472066e-03
area_worst               -3.824774e-05
smoothness_worst          1.573943e+00
compactness_worst         7.515755e-03
concavity_worst           2.807532e-01
concave.points_worst      1.020963e+00
symmetry_worst            8.282728e-01
fractal_dimension_worst   3.053022e+00
```

The coefficients of the final model show that three variables have significantly higher coefficients than other variables: mean fractal dimension, smoothness standard error, and concave points standard error. The mean fractal dimension has an estimated coefficient of –6.94, smoothness standard error has an estimated coefficient of 9.22, and concave points standard error has an estimated coefficient of 5.23.

These coefficients show that when the mean fractal dimension of a participant is higher the chance of them having a malignant tumor is lower. When the smoothness standard error or concave points standard error of a participant is higher the chance of them having a malignant tumor is higher. To again further evaluate the predictive accuracy of the model and confusion matrix was made. To further evaluate the model's performance, since our dependent variable of tumor malignancy is binary, the threshold for converting probabilities of tumor malignancy produced by our regression methods was set to 0.5. If the ridge regression model predicted that there was greater than a 50.00% chance that a participant had a malignant tumor then the prediction was that they did have a malignant, if the ridge regression model predicted that there was 50.00% or less chance of a participant having a malignant tumor then the prediction was that the participant had a benign tumor. The confusion matrix along with final model predictability is included in the appendices.

The confusion matrix reveals that the final model produced by the ridge regression method predicts tumor malignancy correctly at an overall accuracy of 97.25%. The model correctly identifies true positives or malignant tumors in participants who truly have malignant tumors at a rate of 100.00%. The model correctly identifies true negatives or benign tumors in participants who truly have benign tumors at a rate of 91.67%. While this model was an astounding success and is our final predictive model for breast cancer tumor malignancy as in the case of breast cancer research, we want a powerful and accurate model, which can correctly predict malignant tumors, this model predicts malignant tumors of participants who truly have malignant tumors at a rate of 100%.

Limitations and Future Work

This research provides some limitations based on the background of this study. The research was only confined to observations from Wisconsin, meaning that there is a lack of geographical diversity. Due to this limitation, different locations may suggest an effect on the data. In the future, it would be ideal to expand the location and perform similar analyses on different origins. That way it can be confirmed that the findings are not location specific.

Another limitation that can be considered is the lack of sample representation. This sample does not mention any demographic aspects, such as age, sex, diet, etc. Without these aspects, it can be difficult to pinpoint why there are significant differences in observations. In the future, including details about patients' demographics while collecting data can help an analyst have a better understanding of how those factors might affect cell data and give more detailed insights into breast cancer patterns and risk factors.

Finally, a limitation that can be considered is that the data is outdated. The dataset is from 1995, which is almost thirty years ago at the time this study was done. This shows that any conclusions from our analysis could have changed due to possible changes in the last thirty years on tumor identification. For the future, data should be collected every so often to confirm that the analyses work no matter what the time frame or to determine changes in different periods of time.

Conclusion

Our study concludes by highlighting the importance of three distinct aspects of tumor shape—fractal dimension, smoothness, and worst fractal dimension—in differentiating between malignant and benign tumors. These variables' strong predictive power suggests that they could improve patient outcomes by increasing diagnosis accuracy. Since incorrect diagnoses can have a major negative influence on a patient's health and prognosis, accurate tumor identification is essential for directing prompt and appropriate treatment decisions. We can enable medical practitioners to diagnose and treat breast cancer patients with more knowledge if we give priority to these important qualities. In the end, this strategy may have a favorable effect on overall results and survival rates for those with breast cancer, highlighting its significance in the fight against this illness.

REFERENCES

Farahnaz Sadoughi, H. L. (2016). Application of Canonical Correlation Analysis for Detecting Risk Factors Leading to Recurrence of Breast Cancer. doi:10.5812/ircmj.23131

H. K. Chiang, Chui-Mei Tiu, Guo-Shian Hung, Shiao-Chi Wu, T. Y. Chang and Yi-Hong Chou, "Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis," 2001 IEEE Ultrasonics Symposium. Proceedings. An International Symposium (Cat. No.01CH37263), Atlanta, GA, USA, 2001, pp. 1303-1306 vol.2, doi: 10.1109/ULTSYM.2001.991959.

Meshwa Rameshbhai Savalia, J. V. (n.d.). (2016). Classifying Malignant and Benign Tumors of Breast Cancer: A Comparative Investigation Using Machine Learning Techniques. Reliable and Quality E-Healthcare, 12(1).

Sohaila Rehman, S. M. (2013). A Probable Risk Factor of Female Breast Cancer: Study on Benign and Malignant Breast Tissue Samples. Springer Science+Business Media New York. doi:10.1007/s12011-013-9865-7

Toğaçar, M., Ergen, B., & Cömert, Z. (2020). Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders. Medical hypotheses, 135, 109503.

Tsui, T. C. O., Trudeau, M., Mitsakakis, N., Torres, S., Bremner, K. E., Kim, D., Davis, A. M., & Krahn, M. D. (2022). Developing the Breast Utility Instrument, a preference-based instrument to measure health-related quality of life in women with breast cancer: Confirmatory factor analysis of the EORTC QLQ-C30 and BR45 to establish dimensions. PLOS ONE, 17(2), e0262635–e0262635. https://doi.org/10.1371/journal.pone.0262635

APPENDIX

TABLES AND FIGURES

Table 1: Variable Details

Dependent Variable	
Variables	***Description***
Diagnosis	M = malignant, B = benign
Independent Variables	
Variables	***Description***
Radius	mean of distances from center to points on the perimeter
Texture	standard deviation of gray-scale values
Perimeter	The total distance of the cell nucleus boundary.
Area	The area of the cell nucleus
Smoothness	local variation in radius lengths
Compactness	perimeter^2 / area - 1.0
Concavity	severity of concave portions of the contour
Concave points	number of concave portions of the contour)
Symmetry	The longest line from boundary point to boundary point through the center of the nucleus is found.
Fractal dimension	“coastline approximation" - 1

Table 2: Coefficients and Structural Loadings for Function 1

Clinical Group 1		
	Loadings	***Coefficients***
	CV1	***CV1***
radius_mean	0.9999579	0.2675247423
area_mean	0.9887708	0.0001645852
Clinical Group 2		
	CV 1	***CV1***
texture_mean	0.3239416	9.790349e-05
perimeter_mean	0.9981715	0.04337062
smoothness_mean	0.1711212	3.398337
compactness_mean	0.5061759	-1.452199
concavity_mean	0.6779581	-0.1777336
concave.points_mean	0.8233740	-0.05077637
symmetry_mean	0.1480912	0.06949497
fractal_dimension_mean	- 0.3102825	1.425194

Table 3: Principal Components Analysis Summary

Component	**Standard Deviation**	**Proportion of Variance**	**Cumulative Proportion**	**Interpretation**	**Variables Loading Highly**
PC1	3.6444	0.4427	0.4427	Tumor Size/Morphology	radius_mean, perimeter_mean, area_mean, concave.points_mean, radius_worst, perimeter_worst, area_worst
PC2	2.3857	0.1897	0.6324	Texture vs Smoothness	compactness_mean, concavity_mean, fractal_dimension_mean, compactness_se, concavity_se, concave.points_se, fractal_dimension_se, compactness_worst, concavity_worst
PC3	1.6787	0.0939	0.7264	Marginal variance in texture	texture_se
PC4	1.4074	0.0660	0.7924	Marginal variance in smoothness	smoothness_se, symmetry_worst
PC5	1.2840	0.0550	0.8473	Smoothness features	smoothness_mean, smoothness_worst

Table 4: Confusion Matrix for Stepwise Regression and Ridge Regression

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 63 11
         1 10 25

               Accuracy : 0.8073
                 95% CI : (0.7207, 0.8766)
    No Information Rate : 0.6697
    P-Value [Acc > NIR] : 0.001055

                  Kappa : 0.5614

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.8630
            Specificity : 0.6944
         Pos Pred Value : 0.8514
         Neg Pred Value : 0.7143
             Prevalence : 0.6697
         Detection Rate : 0.5780
   Detection Prevalence : 0.6789
      Balanced Accuracy : 0.7787

       'Positive' Class : 0
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 73  3
         1  0 33

               Accuracy : 0.9725
                 95% CI : (0.9217, 0.9943)
    No Information Rate : 0.6697
    P-Value [Acc > NIR] : 2.81e-15

                  Kappa : 0.9364

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 1.0000
            Specificity : 0.9167
         Pos Pred Value : 0.9605
         Neg Pred Value : 1.0000
             Prevalence : 0.6697
         Detection Rate : 0.6697
   Detection Prevalence : 0.6972
      Balanced Accuracy : 0.9583

       'Positive' Class : 0
```

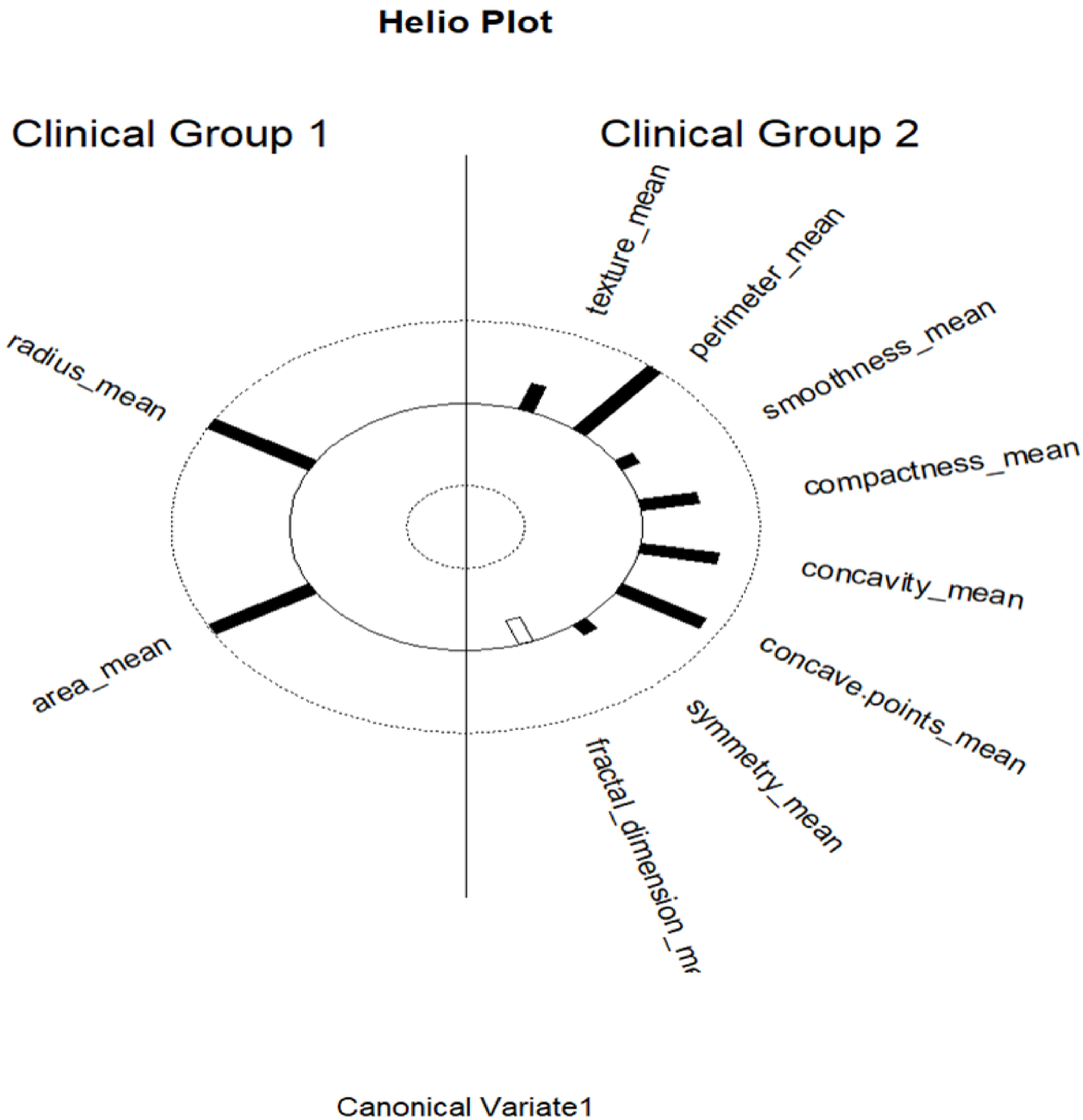


Figure 1: Visualization of the First Canonical Variate

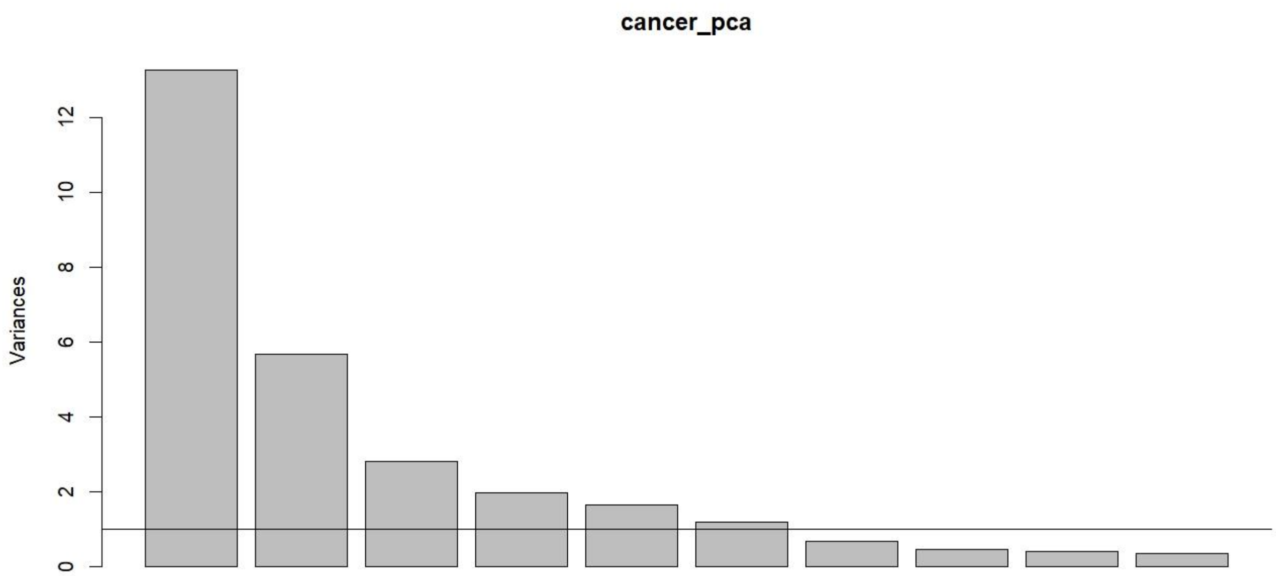


Figure 2: Scree Plot

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F1 =~
    radius_mean       1.000
    perimeter_mean    1.002    0.004  273.469    0.000
    area_mean         0.996    0.017   59.302    0.000
    concav.pnts_mn    0.866    0.027   32.614    0.000
    radius_worst      1.021    0.019   54.822    0.000
    perimeter_wrst    1.020    0.018   55.297    0.000
    area_worst        1.006    0.031   31.979    0.000
  F2 =~
    compactness_mn    1.000
    concavity_mean    0.990    0.040   25.026    0.000
    frctl_dmnsn_mn    0.528    0.049   10.851    0.000
    compactness_se    0.826    0.049   16.938    0.000
    concavity_se      0.744    0.094    7.901    0.000
    concave.pnts_s    0.741    0.060   12.423    0.000
    fractl_dmnsn_s    0.581    0.070    8.297    0.000
    compctnss_wrst    0.924    0.041   22.750    0.000
    concavity_wrst    0.958    0.048   20.032    0.000
  F3 =~
    texture_se        1.000
  F4 =~
    smoothness_se     1.000
    symmetry_worst    0.926    0.277    3.347    0.001
  F5 =~
    smoothness_men    1.000
    smoothnss_wrst    1.076    0.057   19.012    0.000
```

Figure 3: Confirmatory Factor Analysis Summary: Latent Variables

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)
F1 ~~				
F2	0.539	0.055	9.835	0.000
F3	-0.099	0.041	-2.426	0.015
F4	0.004	0.043	0.092	0.927
F5	0.199	0.036	5.469	0.000
F2 ~~				
F3	0.058	0.052	1.124	0.261
F4	0.305	0.043	7.148	0.000
F5	0.489	0.062	7.947	0.000
F3 ~~				
F4	0.148	0.076	1.945	0.052
F5	-0.025	0.052	-0.486	0.627
F4 ~~				
F5	0.383	0.056	6.822	0.000

Figure 4: Confirmatory Factor Analysis Summary: Covariances

Variances:

	Estimate	Std.Err	z-value	P(>|z|)
.radius_mean	0.047	0.009	5.500	0.000
.perimeter_mean	0.043	0.009	5.060	0.000
.area_mean	0.055	0.015	3.640	0.000
.concav.pnts_mn	0.285	0.020	14.236	0.000
.radius_worst	0.007	0.001	4.614	0.000
.perimeter_wrst	0.009	0.002	5.829	0.000
.area_worst	0.036	0.008	4.533	0.000
.compactness_mn	0.110	0.021	5.216	0.000
.concavity_mean	0.128	0.018	7.102	0.000
.frctl_dmnsn_mn	0.751	0.065	11.473	0.000
.compactness_se	0.392	0.057	6.935	0.000
.concavity_se	0.506	0.191	2.654	0.008
.concave.pnts_s	0.511	0.064	8.003	0.000
.fractl_dmnsn_s	0.698	0.170	4.100	0.000
.compctnss_wrst	0.239	0.031	7.820	0.000
.concavity_wrst	0.183	0.023	7.977	0.000
.texture_se	0.000			
.smoothness_se	1.114	0.186	5.994	0.000
.symmetry_worst	1.097	0.132	8.286	0.000
.smoothness_men	0.251	0.036	6.902	0.000
.smoothnss_wrst	0.133	0.035	3.782	0.000
F1	0.951	0.069	13.856	0.000
F2	0.888	0.079	11.197	0.000
F3	0.998	0.113	8.834	0.000
F4	-0.116	0.060	-1.915	0.056
F5	0.747	0.071	10.516	0.000

Figure 5: Confirmatory Factor Analysis Summary: Variances

Code:

CFA

```r
# Natalie Ratowski
# Final project CFA
# 3/3/2024
install.packages("lavaan", dependencies = TRUE)
library(lavaan)

library(Hmisc) #Describe Function
library(psych) #Multiple Functions for Statistics and Multivariate Analysis
library(GGally) #ggpairs Function
library(ggplot2) #ggplot2 Functions
library(vioplot) #Violin Plot Function
library(corrplot) #Plot Correlations
library(REdaS) #Bartlett's Test of Sphericity
library(psych) #PCA/FA functions
library(factoextra) #PCA Visualizations
library("FactoMineR") #PCA functions
library(ade4) #PCA Visualizations

setwd('/Users/DidiRa/Desktop/DSC 324')

dataset <- read.csv(file="data.csv", header=TRUE, sep = ",")
dataset$diagnosis[dataset$diagnosis == "M"] <- 1
dataset$diagnosis[dataset$diagnosis == "B"] <- 0

dim(dataset)
head(dataset)
summary(dataset)
describe(dataset)

dataset2 <- dataset[,3:32]

factors_data <- fa.parallel(dataset2)
factors_data

dataset3 = psych::principal(dataset2, rotate="varimax", nfactors=6, scores=TRUE)
dataset3
print(dataset3$loadings, cutoff=.4, sort=T)
```

```r
dataset_factors <- '
  F1 =~ radius_mean + perimeter_mean + area_mean + concave.points_mean + radius_worst + perimeter_worst + area_worst
  F2 =~ compactness_mean + concavity_mean + fractal_dimension_mean + compactness_se + concavity_se + concave.points_se + fractal_dimension_se + compactness_worst + concavity_worst
   F3 =~ texture_se
  F4 =~ smoothness_se + symmetry_worst
  F5 =~ smoothness_mean + smoothness_worst
'
dataset2_scaled <- as.data.frame(scale(dataset2))

fit <- cfa(dataset_factors, data = dataset2_scaled, estimator = "MLR")
summary(fit)
```

CCA

```
# Fauzia Khan
library(yacca)  # CCA
setwd("~/AdvancedDataAnalyst/Project")
mydata <- read.csv(file = "data (1).csv", header = TRUE,
                   sep = ",")

mydata <- mydata %>% select(-X)
str(mydata)

## Clinical group of combinations
## diagnostic variable : tumor M(Malignant) or B(Benign)
# Research Question: do certain clinical features in clinical_group1
# exhibit significant correlations with specific clinical features in
# clinical_group2

clinical_group1 <-mydata[,c(3, 6)]
clinical_group2 <-mydata[, c(4,5,7:12)]

# run walke's Lamba
ccaWilks = function(set1, set2, cca)
{
  ev = ((1 - cca$cor^2))
  ev

  n = dim(set1)[1]
  p = length(set1)
  q = length(set2)
  k = min(p, q)
  m = n - 3/2 - (p + q)/2
  m

  w = rev(cumprod(rev(ev)))

  # initialize
  d1 = d2 = f = vector("numeric", k)

  for (i in 1:k)
  {
    s = sqrt((p^2 * q^2 - 4)/(p^2 + q^2 - 5))
    si = 1/s
    d1[i] = p * q
    d2[i] = m * s - p * q/2 + 1
    r = (1 - w[i]^si)/w[i]^si
    f[i] = r * d2[i]/d1[i]
    p = p - 1
```

```
    p = p - 1
    q = q - 1
  }

  pv = pf(f, d1, d2, lower.tail = FALSE)
  dmat = cbind(WilksL = w, F = f, df1 = d1, df2 = d2, p = pv)
}


c2 <- cca(clinical_group1, clinical_group2)
summary(c2)


helio.plot(c2, cv=1, x.name="Clinical Group 1",
           y.name="Clinical Group 2")
```

PCA

```
library(GGally) #ggpairs Function
library(ggplot2) #ggplot2 Functions
library(corrplot) #Plot Correlations
library(REdaS) #Bartlett's Test of Sphericity
library(psych) #PCA/FA functions
library(factoextra) #PCA Visualizations
library("FactoMineR") #PCA functions
library(ade4) #PCA Visualizations

getwd()
cancer <- read.csv(file = "data.csv", header = TRUE, sep = ',')
str(cancer)
dim(cancer)
head(cancer)
names(cancer)
###here we are removing the columns with all NA values
cancer_data<- cancer[, !apply(is.na(cancer) | cancer == "NA", 2, all)]
# Check remaining variables
str(cancer_data)
##checking for any NA values
sum(is.na(cancer_data))
##seperating char and num variables for further analysis
cancer_num <- cancer_data[, sapply(cancer_data, is.numeric)]
cancer_char <- cancer_data[, sapply(cancer_data, is.character)]
#Checking for Multicollinearity with Correlations
M<-cor(cancer_num, method="spearman")
M
###There are many strong, significant correlations between the variables:
##radius_mean, perimeter_mean, area_mean are all very highly correlated (>0.99)
##compactness_mean, concavity_mean, concave.points_mean are strongly correlated (>0.8)
##There are also some moderate correlations in the 0.3 - 0.7 range like:
## texture_mean with radius_mean, perimeter_mean, area_mean
##fractal_dimension_mean with radius_mean and area_mean
# PCA_Plot functions
###########################################################################

PCA_Plot = function(pcaData)
{
  library(ggplot2)
```

```r
  theta = seq(0,2*pi,length.out = 100)
  circle = data.frame(x = cos(theta), y = sin(theta))
  p = ggplot(circle,aes(x,y)) + geom_path()

  loadings = data.frame(pcaData$rotation, .names = row.names(pcaData$rotation))
  p + geom_text(data=loadings, mapping=aes(x = PC1, y = PC2, label = .names, colour = .names, fontface="bold")) +
    coord_fixed(ratio=1) + labs(x = "PC1", y = "PC2")
}

PCA_Plot_Secondary = function(pcaData)
{
  library(ggplot2)

  theta = seq(0,2*pi,length.out = 100)
  circle = data.frame(x = cos(theta), y = sin(theta))
  p = ggplot(circle,aes(x,y)) + geom_path()

  loadings = data.frame(pcaData$rotation, .names = row.names(pcaData$rotation))
  p + geom_text(data=loadings, mapping=aes(x = PC3, y = PC4, label = .names, colour = .names, fontface="bold")) +
    coord_fixed(ratio=1) + labs(x = "PC3", y = "PC4")
}

PCA_Plot_Psyc = function(pcaData)
{
  library(ggplot2)

  theta = seq(0,2*pi,length.out = 100)
  circle = data.frame(x = cos(theta), y = sin(theta))
  p = ggplot(circle,aes(x,y)) + geom_path()

  loadings = as.data.frame(unclass(pcaData$loadings))
  s = rep(0, ncol(loadings))
  for (i in 1:ncol(loadings))
  {
    s[i] = 0
    for (j in 1:nrow(loadings))
      s[i] = s[i] + loadings[j, i]^2
    s[i] = sqrt(s[i])
  }

  for (i in 1:ncol(loadings))
    loadings[, i] = loadings[, i] / s[i]

  loadings$.names = row.names(loadings)

  p + geom_text(data=loadings, mapping=aes(x = PC1, y = PC2, label = .names, colour = .names, fontface="bold")) +
    coord_fixed(ratio=1) + labs(x = "PC1", y = "PC2")
}

PCA_Plot_Psyc_Secondary = function(pcaData)
{
  library(ggplot2)

  theta = seq(0,2*pi,length.out = 100)
  circle = data.frame(x = cos(theta), y = sin(theta))
  p = ggplot(circle,aes(x,y)) + geom_path()

  loadings = as.data.frame(unclass(pcaData$loadings))
  s = rep(0, ncol(loadings))
  for (i in 1:ncol(loadings))
  {
    s[i] = 0
    for (j in 1:nrow(loadings))
      s[i] = s[i] + loadings[j, i]^2
    s[i] = sqrt(s[i])
  }

  for (i in 1:ncol(loadings))
    loadings[, i] = loadings[, i] / s[i]

  loadings$.names = row.names(loadings)

  print(loadings)
  p + geom_text(data=loadings, mapping=aes(x = PC3, y = PC4, label = .names, colour = .names, fontface="bold")) +
    coord_fixed(ratio=1) + labs(x = "PC3", y = "PC4")
}
```

```r
####PCA
library(psych)
KMO(cancer_num)
# overall MSA = 0.83
# sufficeint sample size to run the analysis
library(REdaS)
bart_spher(cancer_num)
# p-value < 2.22e-16
# enough variance to split in different components

# run reliability analysis
library(fmsb)
CronbachAlpha(cancer_num)
#####KMO and Bartlett's test shows PCA is suitable.
library(psych)
# Performing PCA
cancer_pca <- prcomp(scale(cancer_num), center = TRUE)
summary(cancer_pca)
# PCA scores
scores <- predict(cancer_pca)
scores
# Scree plot
fviz_eig(cancer_pca)
# PCA loadings
loadings <- cancer_pca$rotation
# Plot loadings
PCA_Plot(cancer_pca)
PCA_Plot_Secondary(cancer_pca)
rawLoadings = cancer_pca$rotation %*% diag(cancer_pca$sdev, nrow(cancer_pca$rotation), nrow(cancer_pca$rotation))
print(rawLoadings)
v = varimax(rawLoadings)
#Options available under varimax function
ls(v)
v
```

Regression Techniques (Stepwise)

```r
library(corrplot)
library(Hmisc)
library(ggplot2)
library(psych)
library(GGally)
library(vioplot)
library(DescTools)
library(leaps)

bcancer <- read.csv("C:/Users/Raamp/Downloads/data.csv")
bcancer$diagnosis[bcancer$diagnosis=="M"] <- 1
bcancer$diagnosis[bcancer$diagnosis=="B"] <- 0
bcancer$diagnosis<- as.numeric(bcancer$diagnosis)
bcancercorr <- corrplot(cor(bcancer, method = "spearman"))

bcancer2<- bcancer[,c(1:15,17:33)]
bcancer2corr <- corrplot(cor(bcancer2, method = "spearman"))

bcancer3<- bcancer2[,c(1:2,4,7:32)]
bcancer3corr <- corrplot(cor(bcancer3, method = "spearman"))

bcancer4<- bcancer3[,c(1:4,8:9,11,13:18,20,23:29)]
bcancer4corr <- corrplot(cor(bcancer4, method = "spearman"),
                         method = "number")

bcancer5<- bcancer4[,c(2,5:8,12:13,19:20)]
bcancer5corr <- corrplot(cor(bcancer5, method = "spearman"),
                         method = "number")


bcancer_w_corr <- bcancer[,c(2:32)]

set.seed(6483)
lregression_indexes <- sample(2, nrow(bcancer5), replace = TRUE, prob = c(0.8,0.2))
lregression_train <- bcancer5[lregression_indexes==1,]
lregression_test <- bcancer5[lregression_indexes==2,]

library(MASS)
stepwise_model_equation <- lm(diagnosis ~. , data = lregression_train)
VIF(stepwise_model_equation)


forward_stepwise_model <- step(stepwise_model_equation,
                          direction = "forward",
                          scope = formula(~ .))
summary(forward_stepwise_model)
```

```r
backward_stepwise_model <- step(stepwise_model_equation,
                               direction = "both",
                               scope = formula(~ .))
summary(backward_stepwise_model)

library(caret)
#forward
forward_testing_predictions <- predict(forward_stepwise_model, newdata = lregression_test, type = "response")

#Forward ROC Eval
library(pROC)
lregression_forward_actual <- data.frame(lregression_test$diagnosis)
evaluation_lregression_forward <- data.frame(actual = lregression_forward_actual,
                                             predicted = forward_testing_predictions)
evaluation_lregression_forward$predicted<- as.numeric(evaluation_lregression_forward$predicted)
lregression_forward_roc <- roc(evaluation_lregression_forward$lregression_test.diagnosis,
                               evaluation_lregression_forward$predicted)
lregression_forward_roc_plot <- plot(lregression_forward_roc ,
                                            main = "ROC Curve",
                                            print.auc = TRUE)
lregression_forward_roc_plot

forward_testing_predictions <- ifelse(forward_testing_predictions > 0.5, 1, 0)
forward_testing_predictions <- factor(forward_testing_predictions)
lregression_test$diagnosis <- factor(lregression_test$diagnosis)
lregression_forward_confusion_matrix <- confusionMatrix(forward_testing_predictions, lregression_test$diagnosis)
lregression_forward_confusion_matrix

#backward
backward_testing_predictions <- predict(backward_stepwise_model,
                                        newdata = lregression_test, type = "response")


#Backward ROC Eval
library(pROC)
lregression_backward_actual <- data.frame(lregression_test$diagnosis)
evaluation_lregression_backward <- data.frame(actual = lregression_backward_actual,
                                             predicted = backward_testing_predictions)
evaluation_lregression_backward$predicted<- as.numeric(evaluation_lregression_backward$predicted)
lregression_backward_roc <- roc(evaluation_lregression_backward$lregression_test.diagnosis,
                               evaluation_lregression_backward$predicted)
lregression_backward_roc_plot <- plot(lregression_backward_roc ,
                                     main = "ROC Curve",
                                     print.auc = TRUE)
lregression_backward_roc_plot

backward_testing_predictions <- ifelse(backward_testing_predictions > 0.5, 1, 0)
backward_testing_predictions <- factor(backward_testing_predictions)
lregression_test$diagnosis <- factor(lregression_test$diagnosis)
lregression_backward_confusion_matrix <- confusionMatrix(backward_testing_predictions, lregression_test$diagnosis)
lregression_backward_confusion_matrix
```

Regression Techniques (Ridge)

```
library(corrplot)
library(Hmisc)
library(ggplot2)
library(psych)
library(GGally)
library(vioplot)
library(DescTools)
library(leaps)
library(glmnet)

bcancer <- read.csv("C:/Users/amanu/OneDrive/Documents/ADA/Homework3/data.csv")
bcancer$diagnosis[bcancer$diagnosis=="M"] <- 1
bcancer$diagnosis[bcancer$diagnosis=="B"] <- 0
bcancer$diagnosis<- as.numeric(bcancer$diagnosis)

bcancer <- bcancer[,c(2:32)]


set.seed(6483)
lregression_indexes <- sample(2, nrow(bcancer), replace = TRUE, prob = c(0.8,0.2))
lregression_train <- bcancer[lregression_indexes==1,]
lregression_test <- bcancer[lregression_indexes==2,]

# Separate predictors and target variable
X_train <- lregression_train[,c(2:31)]  # Exclude the diagnosis column
y_train <- lregression_train$diagnosis

lambda_seq <- 10^seq(2, -2, by = -.1)
# Performing the ridge regression
ridge_model <- glmnet(as.matrix(X_train), y_train, alpha = 0,
                      lambda = lambda_seq)
summary(ridge_model)

ridge_cv <- cv.glmnet(as.matrix(X_train), y_train, alpha = 0)
best_lambda <- ridge_cv$lambda.min
best_lambda

best_fit <- ridge_cv$glmnet.fit
head(best_fit)

best_ridge <- glmnet(X_train, y_train, alpha = 0, lambda = best_lambda)

coef(best_ridge)

X_test <- lregression_test[,c(2:31)]
X_test <- as.matrix(X_test)
y_test <- lregression_test$diagnosis
```

```r
y_test <- lregression_test$diagnosis
y_test <- as.matrix(y_test)

pred <- predict(best_ridge, s = best_lambda, newx = X_test)
length(pred)
library(caret)
testing_predictions <- ifelse(pred > 0.5, 1, 0)
testing_predictions <- factor(testing_predictions)
lregression_test$diagnosis <- factor(lregression_test$diagnosis)


confusion_matrix <- confusionMatrix(testing_predictions, lregression_test$diagnosis)
confusion_matrix
```