

CASE STUDY: Feature Engineering

Feature engineering is the act of choosing, modifying, and converting unprocessed data into features that can be applied in supervised learning. In simple terms, is the process of employing statistical or machine learning techniques to transform raw observations into desired attributes.

To get the best models, the data preparation and analysis processes below can be used.

- **A Checklist for Statistically Exploring Data**
 - Verify the data dimensions, rows, and columns
 - Names of Columns
 - Unique Values per Column and Datatype(s)
- **Data Cleanup**
 - Find any missing information and error (gsub(), regexpr(), ifelse(),...)
 - Able to recognize observations on missing data
 - Graph the Missing Data Representations
 - Choose whether to add missing data or remove it
 - Determine any potential effects during modeling
 - If appropriate for modeling, identify and transform categorical columns and values to numerical representation using dummy variables (as.numeric(),...)
 - Categorical Columns should be subject to feature engineering
 - If necessary for modeling, determine which numerical columns or values should be converted to a categorical representation using a factor (as.factor(),,..)
 - Categorical columns should be checked for distinct values
- **Data Statistical Overview**
 - Verify the data's head and tail to ensure that all necessary data has been loaded
 - Explain the columns of data
 - Locate the number columns and search for information such as mean, median, mode, etc
 - Understand the connection between columns and how they affect one another.
 - Check Chi-Square (displays the relationship between Categorical columns) or RFE/Correlation/GA (displays the relationship between numerical columns)
 - Create additional columns using feature engineering to aid in a better understanding of the data and predictions.

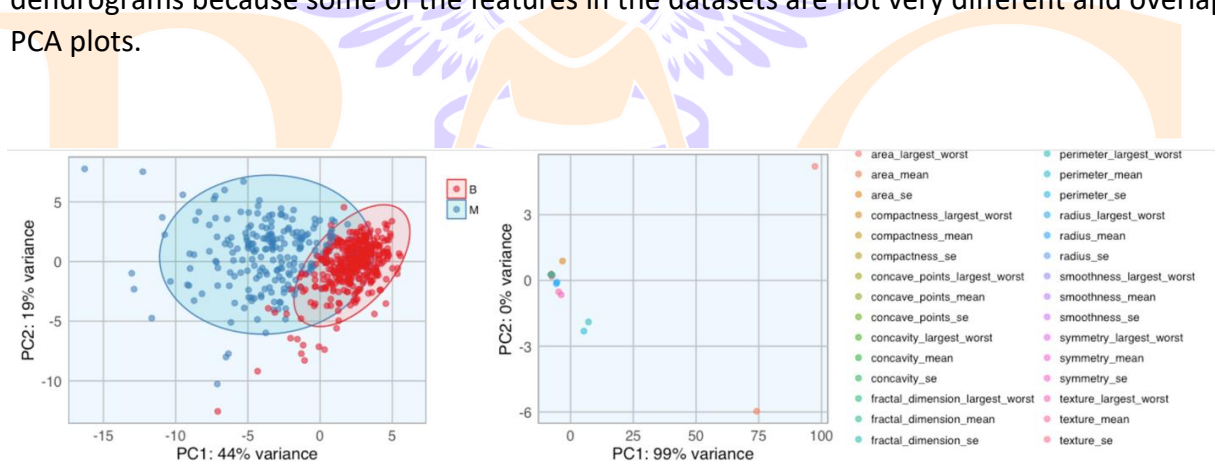
There are numerous techniques to quantify each feature's contribution to the overall model and to limit the number of characteristics that can be used. I will investigate the impact of feature selection via Recursive feature elimination (RFE), correlation and Genetic algorithms (GA) utilizing Random Forest models.

- The Disease State (Diagnostic) Dataset will be the source of the information we utilize to investigate feature selection techniques.

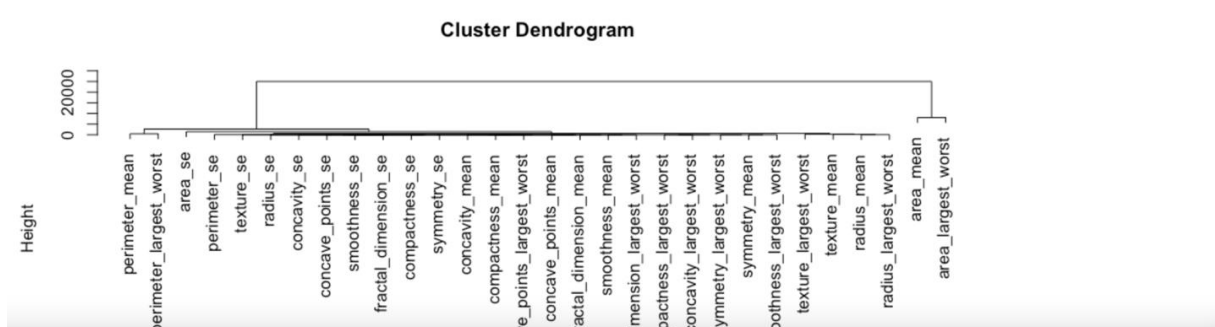
▪ **Analysis**

We are first looking at PCA plots for samples and features to get a sense of the dimensionality and variance of the datasets. The two components that account for most of the variation in the data are displayed in the first two principal components (PCs).

We are developing a function that conducts PCA using the `pcaGoPromoter` package, calculates the ellipses of the data points using the `ellipse` package, and generates the plot using `ggplot2` after specifying our own unique `ggplot2` theme. We are also producing hierarchical clustering dendrograms because some of the features in the datasets are not very different and overlap in the PCA plots.

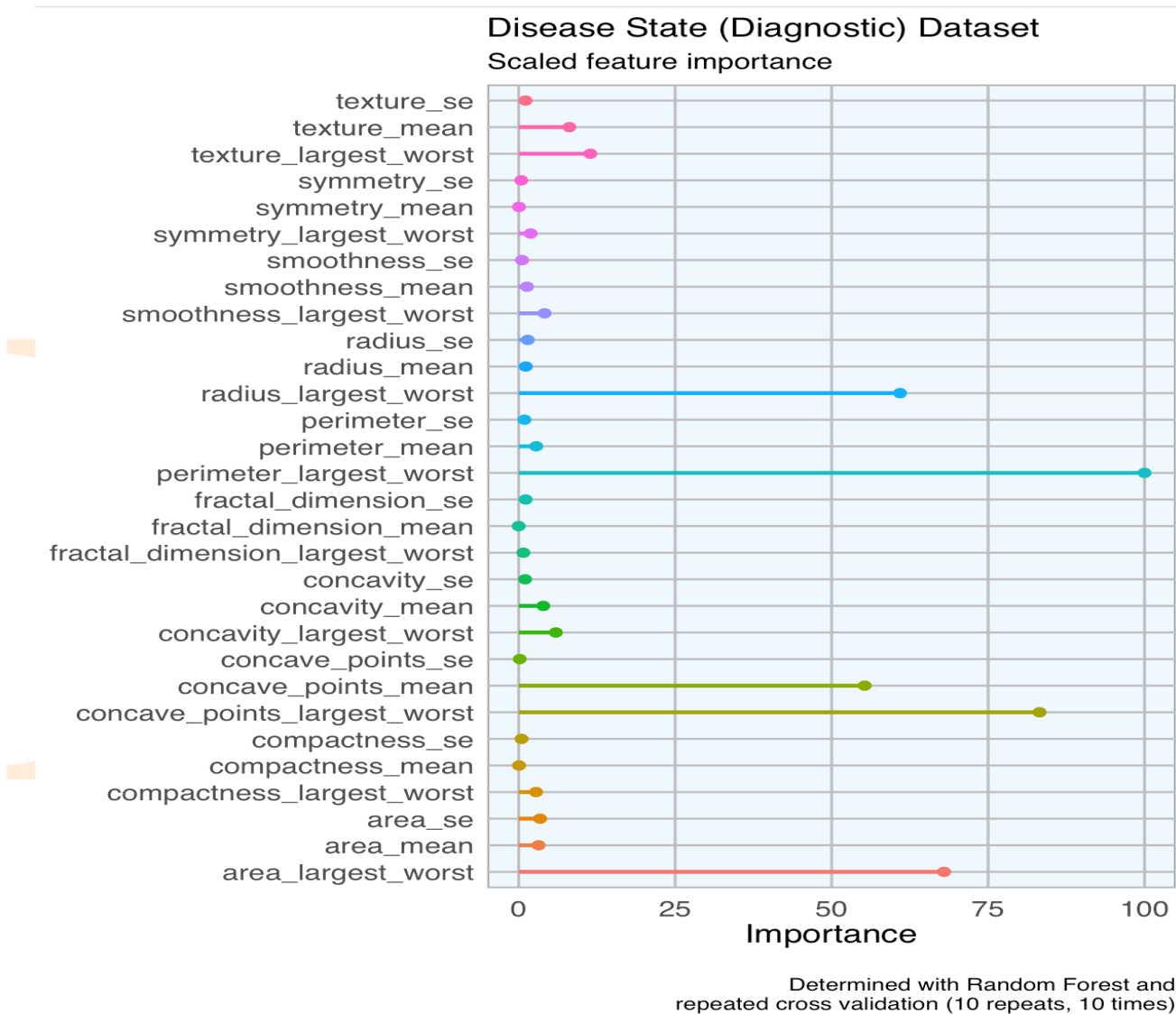


```
h_2 <- hclust(dist(t(bc_data_2[, 3:32])), method = "euclidean", method = "complete")
plot(h_2)
```



- **Feature Importance**

We are using the caret package to run Random Forest models with cross validation to gain an understanding of the relative importance's of the feature. We would need to run feature importance analysis on the training data rather than the entire dataset if we wanted to use feature importance to choose features for modeling.



- **Feature Choice**

Now that we have a general understanding of the data, we will use feature selection technique on the dataset and examine how it affect a Random Forest model's ability to predict outcomes accurately.

- **Generating test and training data**

We must divide the dataset into train and test data before doing anything else with the data. We must only execute the modeling procedure on the training data because performing feature selection on the entire dataset will result in bias in the predictions.

- **Recursive feature elimination (RFE)**

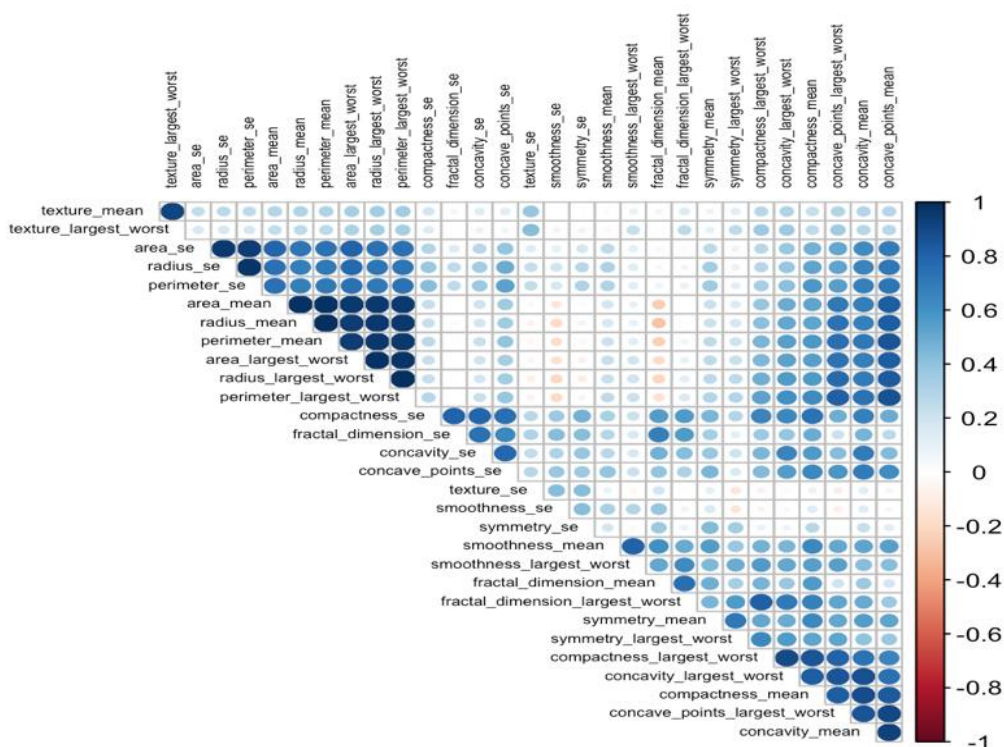
The Random Forest algorithm is used by RFE to evaluate feature combinations and assign an accuracy score to each. Often, the combination with the greatest score is preferred.

```
> # chosen features
> predictors(results_2)
[1] "perimeter_largest_worst"      "area_largest_worst"          "radius_largest_worst"
[4] "concave_points_largest_worst" "concave_points_mean"        "texture_largest_worst"
[7] "area_se"                      "texture_mean"                "concavity_largest_worst"
[10] "concavity_mean"              "radius_se"
>
```

- **Correlation**

We frequently have features that provide redundant information due to their high correlation. We can prevent a prediction bias for the data included in these features by removing highly linked features. This demonstrates the need to remember that just because a feature is useful for predicting an outcome does not mean that it is causative; rather, it may purely be linked with other causal factors when making claims about the biological or medical significance of that feature.

The **corrplot** package calculates and displays correlations between all features. The feature with the lower mean will remain after I have eliminated all features with correlations greater than 0.7.



```

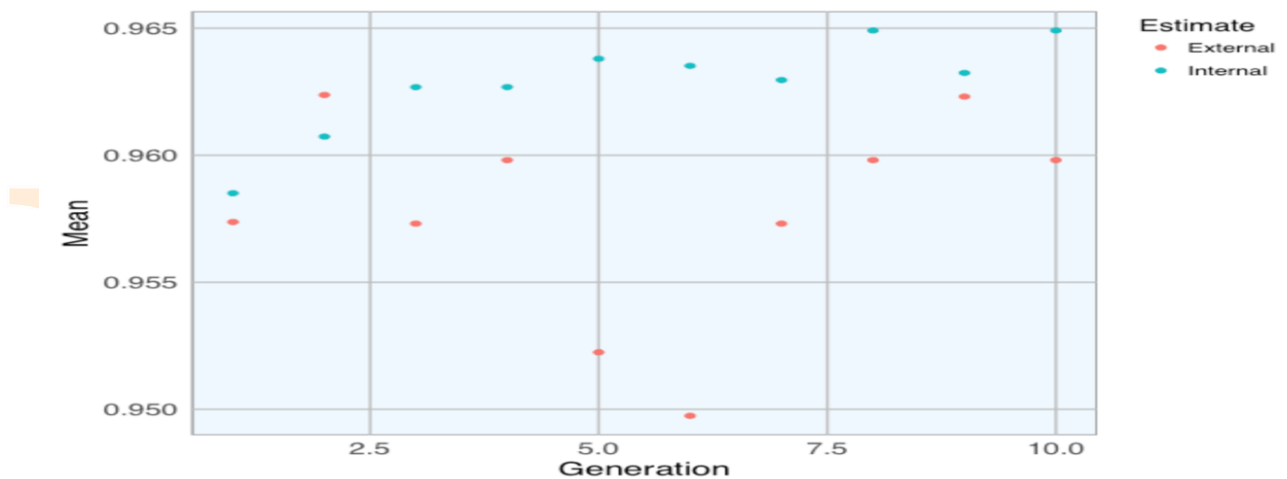
> highlyCor
[1] "concavity_mean"           "concave_points_mean"           "compactness_mean"
[4] "concave_points_largest_worst" "concavity_largest_worst"       "perimeter_largest_worst"
[7] "radius_largest_worst"     "compactness_largest_worst"     "area_largest_worst"
[10] "perimeter_mean"          "perimeter_se"                  "area_mean"
[13] "concave_points_se"       "compactness_se"                "area_se"
[16] "symmetry_mean"           "concavity_se"                  "smoothness_mean"
[19] "fractal_dimension_largest_worst" "texture_largest_worst"
>

```

Out of the 30 features, some seem to be remarkably different while others seem to be closely related. Twenty have removal labels (see output above).

- **Genetic algorithms (GA)**

By simulating selection over time, it seeks to maximize a population of individuals with a specific set of genotypes.

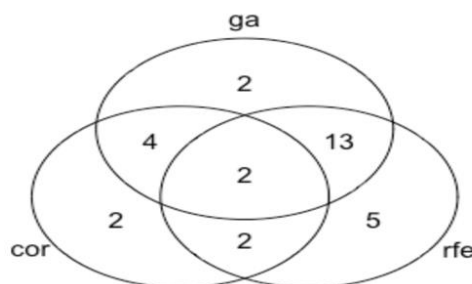


```

[1] "radius_mean"           "texture_mean"
[3] "perimeter_mean"       "area_mean"
[5] "smoothness_mean"      "compactness_mean"
[7] "concavity_mean"       "symmetry_mean"
[9] "fractal_dimension_mean" "texture_se"
[11] "perimeter_se"         "area_se"
[13] "smoothness_se"        "compactness_se"
[15] "concavity_se"         "concave_points_se"
[17] "symmetry_se"          "radius_largest_worst"
[19] "texture_largest_worst" "smoothness_largest_worst"
[21] "compactness_largest_worst" "concave_points_largest_worst"

```

- **Selected features**



All selected just 1 feature, and again, RFE and GA show the greatest overlap, followed by correlation and GA.