**Data cleaning:**

* Remove irrelevant columns from the dataset

**Mutation data:** Deals with change of DNA sequence: deletion, insertion,

**Genomic Data:** A large number of columns appear to represent gene expression levels or other genomic features.

**Gene expression data**: refers to the information that shows which genes are active, and to what extent, in a specific cell or group of cells at a given time. This data is crucial in understanding how genes contribute to the functioning, development, and overall health of an organism. Gene expression is a dynamic process and can change in response to various internal and external factors.

These columns are related to patient information, clinical characteristics, and treatment details.

* **Patient Information:**
  + patient\_id, age\_at\_diagnosis
* **Cancer, Treatment Details, and Clinical Features:**
  + type\_of\_breast\_surgery, cancer\_type, cancer\_type\_detailed, cellularity chemotherapy, hormone\_therapy, radio\_therapy
* **Biological and Clinical Markers:**
  + er\_status\_measured\_by\_ihc, er\_status, pr\_status, her2\_status\_measured\_by\_snp6, her2\_status, 3-gene\_classifier\_subtype
* **Cancer Characteristics:**
  + neoplasm\_histologic\_grade, tumor\_other\_histologic\_subtype, tumor\_size, tumor\_stage primary\_tumor\_laterality, lymph\_nodes\_examined\_positive
* **Prognostic Indices:**
  + nottingham\_prognostic\_index, integrative\_cluster
* **PAM50 plus Claudin-low Subtype:** 
  + A molecular classification of breast cancer
* **Outcome Data:**
  + overall\_survival\_months, overall\_survival, death\_from\_cancer
* **Cohort Information:**
  + cohort, oncotree\_code
* **Other Biomarkers**:
  + er\_status\_measured\_by\_ihc, HER2 status, etc.

**Out of all columns which columns or features can have a significant impact on bringing novel insights?**

**1. Feature Importance Plot**

Given that you've used models like Logistic Regression and potentially others like Random Forest or Gradient Boosting, a plot showing the importance of each feature can be very insightful. This highlights which variables are most influential in predicting the outcome.

**2. Confusion Matrix Heatmap**

A heatmap of the confusion matrix for your model's predictions versus the actual values can visually demonstrate the model's performance in classifying the outcomes

**3. ROC Curve**

The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) provide insights into the model's ability to distinguish between classes. This is especially relevant for binary classification problems.

**4. Model Performance Summary Table**

A table summarizing key performance metrics (e.g., Accuracy, Precision, Recall, F1-Score, AUC) for both the training and testing sets can provide a quick reference to the model's effectiveness.

This can be created using pandas DataFrame and displayed using either print(df) for textual output or df.style for a more stylized HTML representation in Jupyter Notebooks.

**5. Predictions vs. Actual Values Scatter Plot**

As you've already done, scatter plots comparing the actual versus predicted values for both the training and testing data offer a straightforward visualization of the model's accuracy.

These visualizations and summaries can significantly enhance the presentation of your work on GitHub and in publications, providing clear, evidence-based insights into your model's performance and the importance of different predictors in your analysis.