Module_3: *(Template)*

Team Members:

Diya and Alex

Project Title:

ML model for metastatic progression (M0 vs. M1) based on the expression levels of the genes TWIST1, SNAI2, ZEB2, VIM, and CDH1.

Project Goal:

This project seeks to… *(what is the purpose of your project -- i.e., describe the question that you seek to answer by analyzing data.)*

We are planning to analyze whether breast cancer samples exhibit metastatic progression (M0 vs. M1) based on the expression levels of the genes TWIST1, SNAI2, ZEB2, VIM, and CDH1.

Disease Background:

Pick a hallmark to focus on, and figure out what genes you are interested in researching based on that decision. Then fill out the information below.

- Cancer hallmark focus: tissue invasion and metastases
- Overview of hallmark: This is when the primary tumor masses create pioneer cells that go and invade other areas of the body – they form new colonies. Metastases which are distant tumor settlements cause 90 percent of cancer deaths, showing how prominent this hallmark is. At these new sites there is a lot of nutrients and space to grow which allow the cancerous cells to proliferate at a high rate.
- Genes associated with hallmark to be studied (describe the role of each gene, signaling pathway, or gene set you are going to investigate): One protein is E-cadherin which is a cell to cell adhesion molecule that maintains epithelial integrity. When it is lost it changes adhesion which allows cells to detach and “find new homes.” β-catenin links to E-cadherin and is a transcriptional coactivator. When this protein is not attached to E-cadherin it enters the nucleus to promote cell proliferation. Altered integrins promote migration of cells.

Will you be focusing on a single cancer type or looking across cancer types? Depending on your decision, update this section to include relevant information about the disease at the appropriate level of detail. Regardless, each bullet point should be filled in. If you are looking at multiple cancer types, you should investigate differences between the types (e.g. what is the most prevalent cancer type? What type has the highest mortality rate?) and similarities (e.g. what sorts of treatments exist across the board for cancer patients? what is common to all *cancers in terms of biological mechanisms?). Note that this is a smaller list than the initial 11 in Module 1.*

- I think we are going to look at one type of cancer and it to be breast cancer or breast invasive carcinoma.
- Prevalence & incidence: 1 in 8 women in the United States will be diagnosed with breast cancer in their lifetime. In 2020, an estimated 316,950 women and 2,800 men will be diagnosed with invasive breast cancer and an additional 59,080 new cases of non-invasice breast cancer. However the 5 year relative survival rate is 99%. https://www.nationalbreastcancer.org/breast-cancer-facts/
- Risk factors (genetic, lifestyle) & Societal determinants:
  - Getting older: risk for breast cancer increases with age and most cases are diagnosed after the age of 50.
  - Genetic mutations: Inherited changes to genes like BRCA1 and BRCA2 increase risk of breast cancer.
  - Dense breasts: Makes it harder to see tumors and increases risk of breast cancer
  - Personal history: If you have had breast cancer once you are more likely to get it again.
  - Family history: Having a family history of breast cancer will increase your risk of the cancer.
  - Previous radiation therapy: radiation therapy in the past increases the likely hood of breast cancer
  - Exposure to diethylstilbestrol: DES exposure increases likely hood of breast cancer if you took it while pregnanat.
  - Exercise: women who are less active are more likely to develop this cancer
  - Overweight: Older overweight women have a higher risk than healthy weight.
  - Taking hormones: taking oral contraceptives (birth control) increase risk of breast cancer.
  - reproductive hisotry: having pregnancy after 30, not breastfeeding, and never having a full term pregnancy can increase risk.
  - Alchohol: Increased alcohol consumption leads to higher risk. https://www.cdc.gov/breast-cancer/risk-factors/index.html
- Standard of care treatments (& reimbursement)
  - The primary treatment for breast cancer is mainly surgery. There are three main ones that a patient can do. The first one is a lumpectomy which removes the tumor and a small amount of surrounding healthy tissue which is followed by radiation. Next is a mastectomy which removes the entire breast and sometimes nearby lymph nodes. Additionally the last surgery is lymph node surgery which removes the first lymph nodes to which the cancer has spread too.
  - Radiation therapy uses high energy rays to kill the cancer. Normally this is done after surgery to destroy the remaining cancer cells. It can also be used to treat cancer that has metastisized to other parts of the body.

- Chemotherapy uses medicine to kill the cells which is given before surgery to shrink a tumor or after to eliminate the remaining cancerous cells.
- Hormone therapy blocks hormones form attatching to the cancer cells or stops the body from producing estrogen.
- Target therapy is used to attack specifin vulnerabilities within cancerous cells.
- Immuno therapy helps the patients immune system fight the cancer.
- For reinbursemnet most major health insurance planscover standard breast care treatments including chmotherapty, radiation, diagnostic tests, surgery, and medication.
- Additionally the Womens Health and Cancer Rights Act is a law that requires most health plans to cover all stages of breast reconstruction following a mastectomy.
- The cost of treating breatcancer ranges from 60 thousand to 134 thousand depending on the stage. https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/ https://www.themedicareconnection.com/article/will-my-health-insurance-cover-breast-cancer-treatments/

- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology)
  - Structure: Breast is composed of lobes which are subdivided into lobules.
  - Sites of origin for cancer: Most breast cancer begins in ductal epitherlial cells or lobular epithlial cells. Some less comonly start in stromal tissue.
  - Cancer cells often first spread to neighboring lymph nodes before metastasizing to other organs.
  - Breast is hormonally regulated by estrogen and progesterone which control growth and developments and cell division and differentiation. Abnormal hormonal signaling can lead to epitherlial cell proliferation.
  - BRCA1 and BRCA2 mutations impair DNA which can lead to cancer.
  - Additionally HER2 amplification can lead to more cell proliferation. https://www.ncbi.nlm.nih.gov/books/NBK482286/

Data-Set:

Once you decide on the subset of data you want to use (i.e. only 1 cancer type or many; any clinical features needed?; which genes will you look at?) describe the dataset. There are a ton of clinical features, so you don't need to describe them all, only the ones pertinent to your question.

(Describe the data set(s) you will analyze. Cite the source(s) of the data. Describe how the data was collected -- What techniques were used? What units are the data measured in? Etc.) We are planning to analyze the relationship between breast cancer metastasis status, indicated by the AJCC pathologic metastasis stage, and the expression levels of five genes: TWIST1, ZEB2, VIM, CDH1, and SNAI2. Our analysis focuses on breast cancer (BRCA) cases in the TCGA Pan-Cancer Clinical Data Resource, which includes molecular profiles of over 11,000 tumors across 33 cancer types. The dataset contains 1,097 BRCA cases with RNA-sequencing–based gene expression data and clinical metadata, including metastasis status (M0 for no metastasis, M1 for distant metastasis).

The selected genes are key regulators of epithelial-to-mesenchymal transition (EMT), a process that promotes cancer cell migration and metastasis. TWIST1, SNAI2 (SLUG), and ZEB2 are transcription factors that repress epithelial markers and induce mesenchymal traits. VIM, which encodes vimentin, serves as a marker of the mesenchymal phenotype, while CDH1, which encodes E-cadherin, is an epithelial marker often downregulated in metastatic cells. By comparing gene expression between M0 and M1 samples, we aim to determine whether these EMT-related genes can help predict metastatic potential in breast cancer patients.

Sources: https://doi.org/10.1016/j.cell.2018.02.052
https://medlineplus.gov/genetics/gene/snai2/ https://medlineplus.gov/genetics/gene/twist1/
https://medlineplus.gov/genetics/gene/zeb2/ https://medlineplus.gov/genetics/gene/cdh1/
http://www.cancerindex.org/geneweb/VIM.htm

Data Analyis:

Methods

The machine learning technique I am using is: *fill in and describe* I am using K-means clustering from sklearn.cluster, an unsupervised learning method that groups samples based on similarity in gene expression across TWIST1, SNAI2, ZEB2, VIM, and CDH1.

What is this method optimizing? How does the model decide it is "good enough"? The method opitmizes coefficients by minimizing within-cluster sum of squares.

**

Analysis

(Describe how you analyzed the data. This is where you should intersperse your Python code so that anyone reading this can run your code to perform the analysis that you did, generate your figures, etc.)

This code creates a subset of the full dataset with all cancers containing only the relevant samples and genes for our analysis and only for breast cancer. First, it loads the gene expression data (GSE62944_subsample_topVar_log2TPM.csv) and corresponding clinical metadata (GSE62944_metadata.csv). It then filters the metadata to include only BRCA samples with defined metastasis status, either M0 (no metastasis) or M1 (metastasis). Next, it selects the expression values for five genes of interest—TWIST1, SNAI2, ZEB2, VIM, and CDH1—and aligns them with the filtered sample IDs. It than creates a new CSV file with all of this data stored

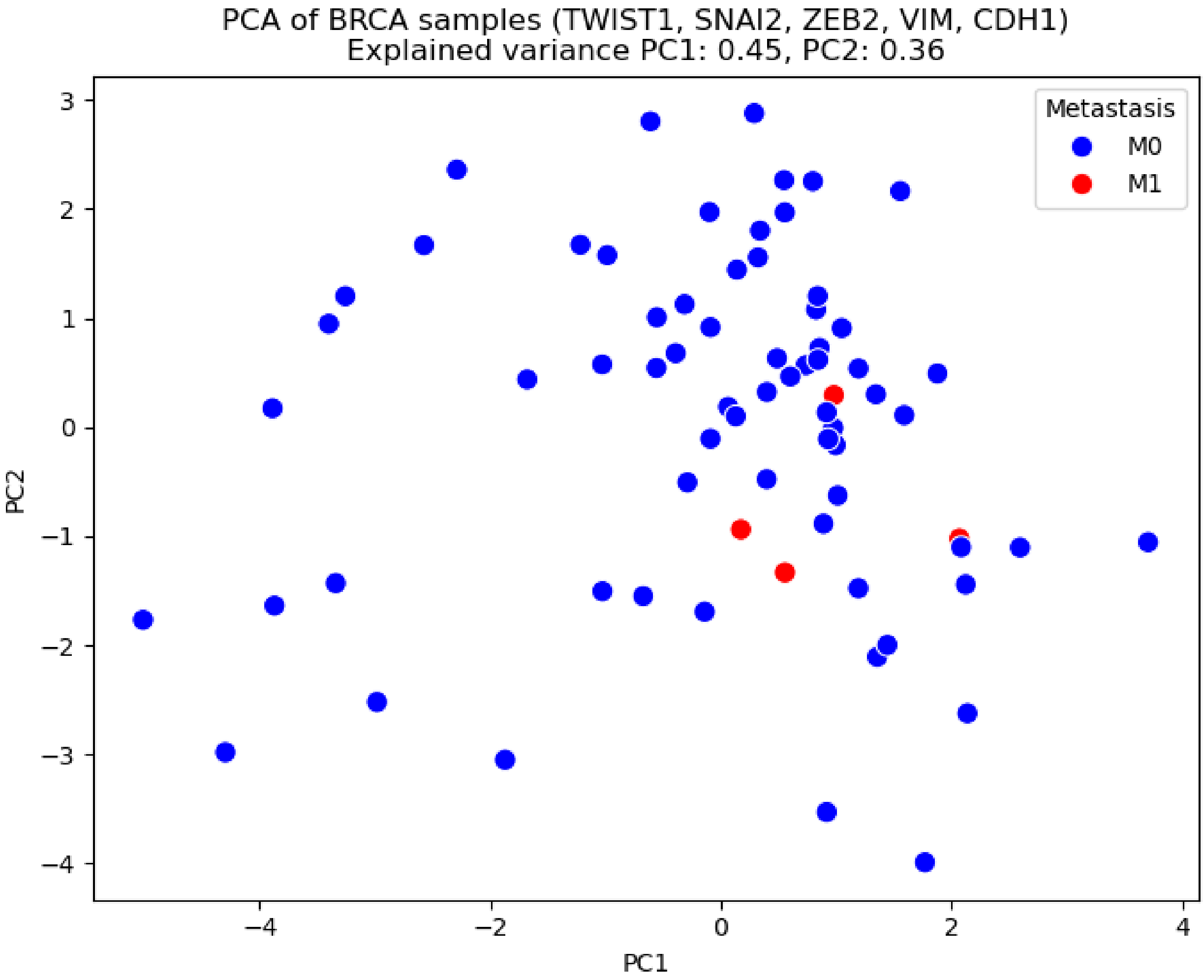
```python
import pandas as pd

# === Step 1: Load data ===
expr_path = "GSE62944_subsample_topVar_log2TPM.csv"
meta_path = "/Users/diyamahendravadi/Downloads/Computational
BME/Module 3/GSE62944_metadata.csv"

expr = pd.read_csv(expr_path, index_col=0)
meta = pd.read_csv(meta_path)
```

```
# === Step 2: Filter metadata for BRCA samples with M0/M1 metastasis
===
meta_subset = meta[
    (meta["cancer_type"] == "BRCA") &
    (meta["ajcc_metastasis_pathologic_pm"].isin(["M0", "M1"]))
]

# === Step 3: Get expression for relevant genes ===
genes_of_interest = ["TWIST1", "SNAI2", "ZEB2", "VIM", "CDH1"]
sample_ids = meta_subset["sample"].values
expr_subset = expr.loc[expr.index.isin(genes_of_interest),
expr.columns.intersection(sample_ids)]

# === Step 4: Transpose and merge ===
expr_subset_T = expr_subset.T.reset_index().rename(columns={"index":
"sample"})
merged = pd.merge(meta_subset, expr_subset_T, on="sample",
how="inner")

# === Step 5: Keep only relevant columns ===
final_subset = merged[["sample", "cancer_type",
"ajcc_metastasis_pathologic_pm"] + genes_of_interest]

# === Step 6: Save minimal CSV ===
final_subset.to_csv("BRCA_selected_genes_minimal.csv", index=False)

print("□ Saved: BRCA_selected_genes_minimal.csv")
print(final_subset.head())
```

```
□ Saved: BRCA_selected_genes_minimal.csv
                         sample cancer_type
ajcc_metastasis_pathologic_pm  \
0  TCGA-E9-A1NI-01A-11R-A14D-07        BRCA
M0
1  TCGA-E2-A1LK-01A-21R-A14D-07        BRCA
M0
2  TCGA-BH-A0B2-01A-11R-A10J-07        BRCA
M0
3  TCGA-E2-A107-01A-11R-A10J-07        BRCA
M0
4  TCGA-LL-A5YN-01A-11R-A28M-07        BRCA
M0

     TWIST1     SNAI2      ZEB2        VIM      CDH1
0  4.071724  5.275409  3.075711  10.373745  7.904312
1  1.919456  7.149143  1.906946  12.121095  8.173847
2  4.236546  5.527577  3.894069  10.672051  4.101489
3  2.416959  3.174286  2.032441   9.279936  4.825750
4  3.111332  3.231575  1.853955  10.033820  6.956183
```

This code performs principal component analysis (PCA) on the subsetted BRCA. First, it loads the minimal CSV file containing only the five genes of interest—TWIST1, SNAI2, ZEB2, VIM, and CDH1—along with the metastasis status for each sample. The gene expression values are extracted into a separate matrix, and PCA is applied to compute the first two principal components, which capture the directions of greatest variance in the data. A scatter plot is generated where each point represents a sample, colored by metastasis status (M0 in blue, M1 in red), allowing visualization of how samples cluster or separate in the reduced two-dimensional space. The explained variance for each principal component is also displayed, providing insight into how much of the original variation is captured by the first two components.

```
import pandas as pd
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns

# === Step 1: Load the subsetted data ===
data = pd.read_csv("BRCA_selected_genes_minimal.csv")

# === Step 2: Extract only the gene expression columns ===
genes = ["TWIST1", "SNAI2", "ZEB2", "VIM", "CDH1"]
X = data[genes]

# === Step 3: Perform PCA ===
pca = PCA(n_components=2)  # First 2 principal components
principal_components = pca.fit_transform(X)
explained_var = pca.explained_variance_ratio_

# === Step 4: Create a DataFrame with PCA results ===
pca_df = pd.DataFrame(data=principal_components, columns=['PC1',
'PC2'])
pca_df = pd.concat([pca_df, data[['ajcc_metastasis_pathologic_pm']]],
axis=1)

# === Step 5: Plot PCA results ===
plt.figure(figsize=(8,6))
sns.scatterplot(
    x='PC1', y='PC2',
    hue='ajcc_metastasis_pathologic_pm',
    data=pca_df,
    palette={'M0':'blue', 'M1':'red'},
    s=80
)
plt.title(f'PCA of BRCA samples (TWIST1, SNAI2, ZEB2, VIM, CDH1)\
nExplained variance PC1: {explained_var[0]:.2f}, PC2:
{explained_var[1]:.2f}')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.legend(title='Metastasis')
plt.show()
```



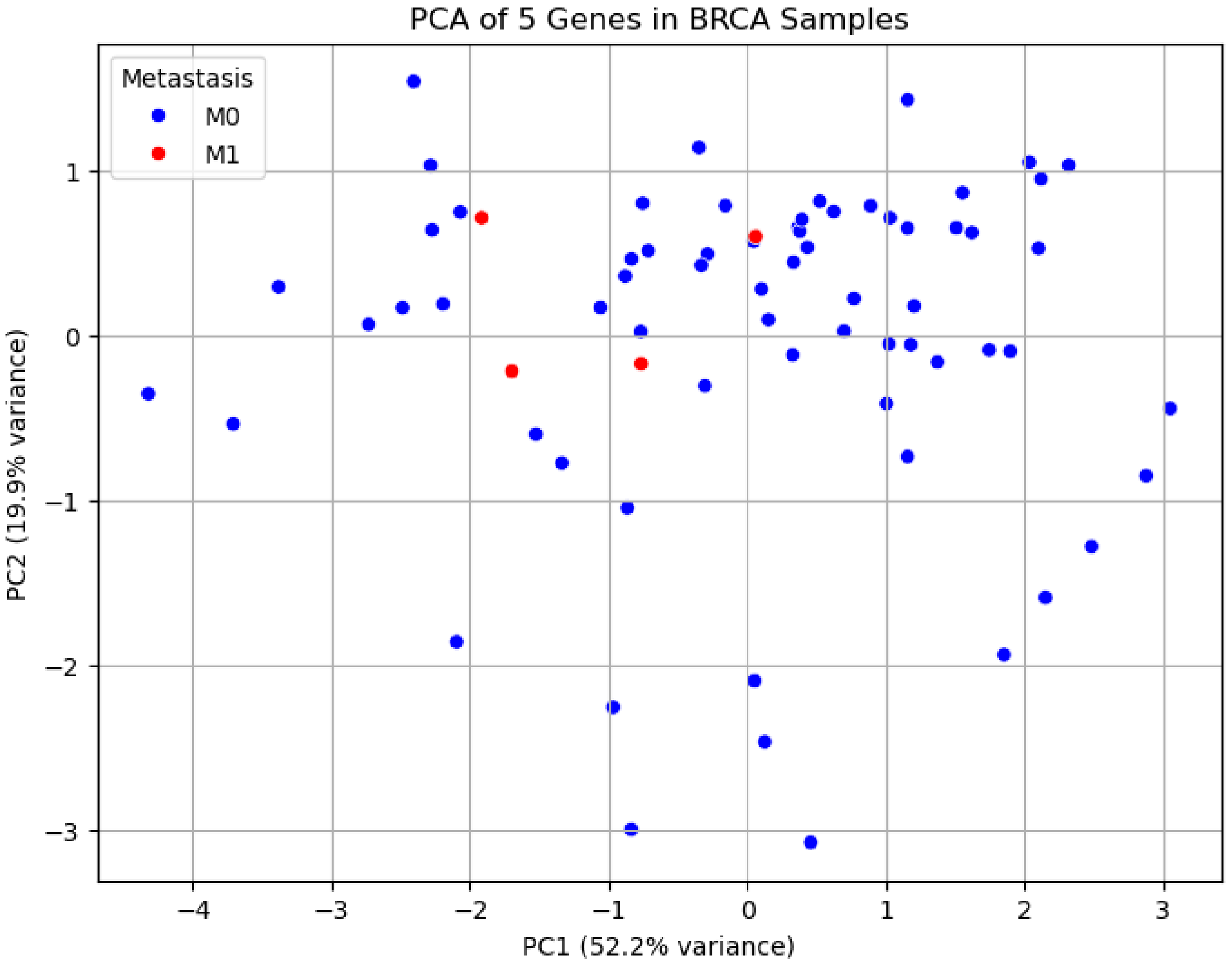
This code performs Principal Component Analysis (PCA) on the expression levels of five genes—TWIST1, SNAI2, ZEB2, VIM, and CDH1—in a subset of breast cancer (BRCA) samples. First, the data is standardized and metastasis labels are encoded numerically (M0=0, M1=1). PCA is then applied to reduce the dimensionality of the data while capturing the maximum variance. The explained variance for each principal component is calculated, and the loadings show how each gene contributes to the components. Finally, a scatter plot of the first two principal components is generated, with points colored by metastasis status, allowing visualization of whether these genes can separate metastatic and non-metastatic samples in the reduced PCA space.

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# === Step 1: Load the already subsetted data ===
df = pd.read_csv("/Users/diyamahendravadi/Downloads/Computational
BME/Module 3/BRCA_selected_genes_minimal.csv")

# Rename metastasis column for clarity
df = df.rename(columns={"ajcc_metastasis_pathologic_pm":
"metastasis"})
df['metastasis_label'] = df['metastasis'].map({'M0':0, 'M1':1})

# === Step 2: Select features ===
genes = ['TWIST1', 'SNAI2', 'ZEB2', 'VIM', 'CDH1']
X = df[genes].values

# Standardize features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# === Step 3: Perform PCA ===
pca = PCA(n_components=5)  # keep all to check explained variance
X_pca = pca.fit_transform(X_scaled)

# === Step 4: Variance explained ===
explained_var = pca.explained_variance_ratio_
cumulative_var = explained_var.cumsum()
print("Explained variance per PC:", explained_var)
print("Cumulative variance:", cumulative_var)

# === Step 5: Loadings ===
loadings = pd.DataFrame(pca.components_.T, index=genes,
columns=[f'PC{i+1}' for i in range(5)])
print("\nPCA Loadings:\n", loadings)

# === Step 6: Scatter plot of first two PCs colored by metastasis ===
pca_df = pd.DataFrame(X_pca[:, :2], columns=['PC1', 'PC2'])
pca_df['metastasis'] = df['metastasis']

plt.figure(figsize=(8,6))
sns.scatterplot(data=pca_df, x='PC1', y='PC2', hue='metastasis',
palette=['blue', 'red'])
plt.title('PCA of 5 Genes in BRCA Samples')
plt.xlabel(f'PC1 ({explained_var[0]*100:.1f}% variance)')
plt.ylabel(f'PC2 ({explained_var[1]*100:.1f}% variance)')
plt.legend(title='Metastasis')
plt.grid(True)
plt.show()
```

```
Explained variance per PC: [0.52152769 0.19938607 0.11202157
0.10463723 0.06242744]
Cumulative variance: [0.52152769 0.72091376 0.83293533 0.93757256 1.
]
```

```
PCA Loadings:
              PC1       PC2       PC3       PC4       PC5
TWIST1  0.467469 -0.006349  0.761360 -0.448564 -0.023549
SNAI2   0.533480  0.155608 -0.167379  0.228632  0.781599
ZEB2    0.480788 -0.005875 -0.622047 -0.538798 -0.302595
VIM     0.507278  0.026006  0.068627  0.672781 -0.533525
CDH1   -0.091564  0.987439  0.025764 -0.059838 -0.111071
```

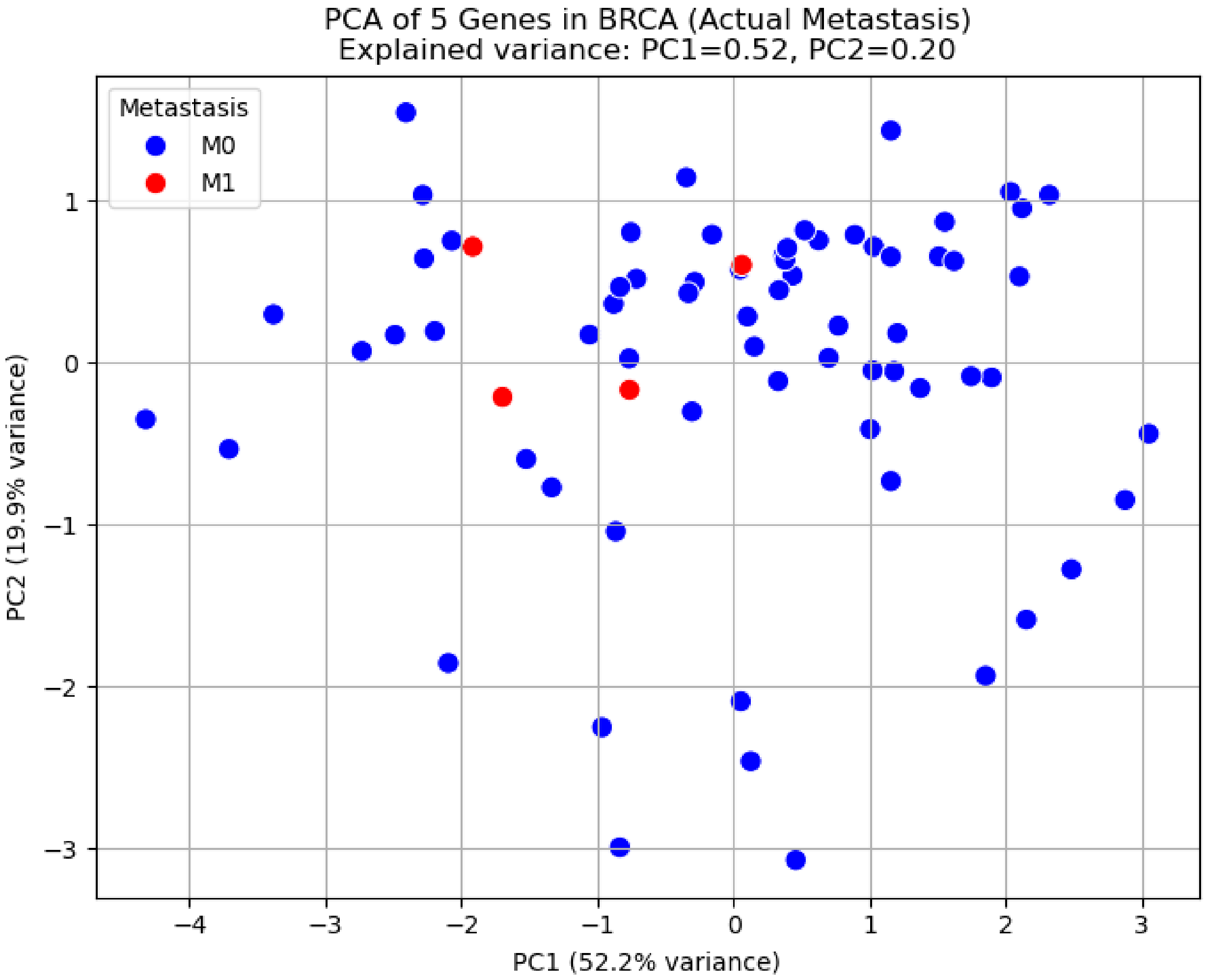


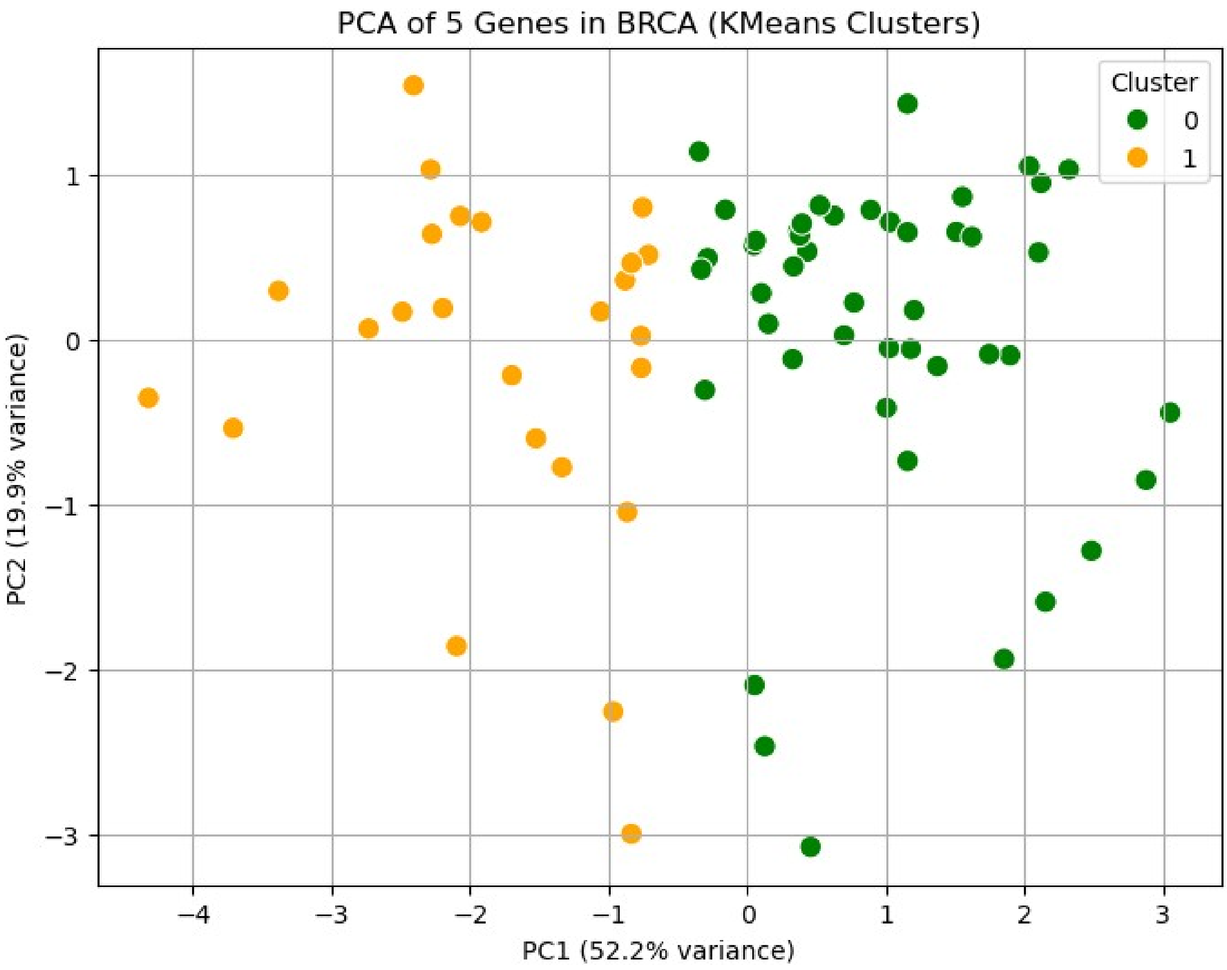
This code applies KMeans clustering to the expression levels of five genes—TWIST1, SNAI2, ZEB2, VIM, and CDH1—in breast cancer (BRCA) samples. First, the gene expression data is standardized. KMeans is then used to partition the samples into two clusters, and the silhouette score quantifies how well the clustering separates the samples, with higher values indicating clearer separation. To visualize the data, PCA reduces the five-dimensional gene expression space to two principal components. Two scatter plots are generated: the first colors the samples by their actual metastasis status (M0 vs. M1), and the second colors them by the KMeans cluster assignments. This allows comparison of the clusters with the true metastasis labels and provides insight into how well the gene expression profiles group metastatic and non-metastatic samples.

```python
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import seaborn as sns

# === Step 1: Load the subsetted data ===
df = pd.read_csv("/Users/diyamahendravadi/Downloads/Computational
BME/Module 3/BRCA_selected_genes_minimal.csv")

# === Step 2: Select features ===
genes = ['TWIST1', 'SNAI2', 'ZEB2', 'VIM', 'CDH1']
X = df[genes].values

# Standardize features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# === Step 3: KMeans clustering using all features ===
kmeans = KMeans(n_clusters=2, random_state=42)
cluster_labels = kmeans.fit_predict(X_scaled)

# Silhouette score
sil_score = silhouette_score(X_scaled, cluster_labels)
print(f"Silhouette Score (using all 5 genes): {sil_score:.3f}")

# === Step 4: PCA for visualization ===
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
explained_var = pca.explained_variance_ratio_

pca_df = pd.DataFrame(X_pca, columns=['PC1', 'PC2'])
pca_df['Metastasis'] = df['ajcc_metastasis_pathologic_pm']
pca_df['Cluster'] = cluster_labels.astype(str)  # for coloring

# === Step 5: Plot PCA colored by actual metastasis ===
plt.figure(figsize=(8,6))
sns.scatterplot(data=pca_df, x='PC1', y='PC2', hue='Metastasis',
palette=['blue','red'], s=80)
plt.title(f'PCA of 5 Genes in BRCA (Actual Metastasis)\nExplained
variance: PC1={explained_var[0]:.2f}, PC2={explained_var[1]:.2f}')
plt.xlabel(f'PC1 ({explained_var[0]*100:.1f}% variance)')
plt.ylabel(f'PC2 ({explained_var[1]*100:.1f}% variance)')
plt.legend(title='Metastasis')
plt.grid(True)
plt.show()

# === Step 6: Plot PCA colored by KMeans clusters ===
plt.figure(figsize=(8,6))
```

```python
sns.scatterplot(data=pca_df, x='PC1', y='PC2', hue='Cluster',
palette=['green','orange'], s=80)
plt.title('PCA of 5 Genes in BRCA (KMeans Clusters)')
plt.xlabel(f'PC1 ({explained_var[0]*100:.1f}% variance)')
plt.ylabel(f'PC2 ({explained_var[1]*100:.1f}% variance)')
plt.legend(title='Cluster')
plt.grid(True)
plt.show()
```

```
Silhouette Score (using all 5 genes): 0.323
```





This is the testing and training for the linear regression model. Since the R^2 we know was very low this makes sense that this shows that it is poor at giving us what we want.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import r2_score, mean_squared_error,
mean_absolute_error

# === Load the already subsetted data ===
df = pd.read_csv("BRCA_selected_genes_minimal.csv")
df = df.rename(columns={"ajcc_metastasis_pathologic_pm":
"metastasis"})
df['metastasis_label'] = df['metastasis'].map({'M0':0, 'M1':1})

# === Select features and target ===
genes = ['TWIST1', 'SNAI2', 'ZEB2', 'VIM', 'CDH1']
X = df[genes].values
y = df['metastasis_label'].values
```

```python
# === Standardize features ===
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# === Split into train/test ===
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.2, random_state=42, stratify=y
)

# === Train linear regression ===
model = LinearRegression()
model.fit(X_train, y_train)

# === Predict on test set ===
y_pred = model.predict(X_test)

# === Evaluate ===
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)

print(f"R²: {r2:.3f}")
print(f"MSE: {mse:.3f}")
print(f"MAE: {mae:.3f}")
```

```
R²: -0.206
MSE: 0.075
MAE: 0.122
```

Verify and validate your analysis:

Pick a SPECIFIC method to determine how well your model is performing and describe how it works here.

- We evaluated the performance of our KMeans clustering model using the silhouette score, which measures how similar each sample is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where higher values indicate better clustering. Our model achieved a silhouette score of 0.323, suggesting a moderate clustering structure: the clusters are somewhat distinct but there is overlap between them. This is consistent with the observation that samples with metastasis (M1) are scattered among non-metastatic (M0) samples, reflecting the complexity of the underlying biological data.
- To assess how well PCA captured the variance in the data, we examined the explained variance ratio for each principal component. The explained variance indicates the proportion of the dataset's variability captured by each PC. For our dataset of five genes (TWIST1, SNAI2, ZEB2, VIM, CDH1), the first two principal components explained approximately 52.2% and 19.9% of the total variance,

respectively, for a cumulative variance of ~72.1%. This suggests that a two-dimensional PCA representation retains most of the information while reducing dimensionality, although some variability remains unaccounted for. The loadings for each gene on the principal components indicate which genes contribute most to each PC, helping interpret the directions of maximum variance.

(Describe how you checked to see that your analysis gave you an answer that you believe (verify). Describe how your determined if your analysis gave you an answer that is supported by other evidence (e.g., a published paper).

- To verify our analysis, we checked that the results were consistent with both the data and established biological mechanisms. The moderate silhouette score and low R^2 were expected, since metastasis is a complex process that likely involves many genes beyond the five genes analyzed. The PCA results highlighted TWIST1, SNAI2, and ZEB2 as major contributors, which aligns with their known roles in epithelial–mesenchymal transition (EMT). To validate these findings, we compared our results to a study examining SNAI2, TWIST1, and CDH1 expression in thyroid carcinoma (https://doi.org/10.1038/modpathol.2012.137). Although the study focused on thyroid rather than breast cancer, it similarly found that increased TWIST1 and SNAI2 expression corresponded with loss of CDH1 in more aggressive, poorly differentiated tumors, supporting the EMT model. This parallel suggests that the gene expression patterns identified in our breast cancer dataset reflect broader EMT-related mechanisms linked to metastasis across epithelial cancers.
- TWIST1 is proven to promote metastasis and invasion in BRCA by blocking Foxa1 expression (Xu, et al.) CDH1 is shown to have heightened expression in BRCA with an unfavorable distant metastasis-free survival outcome (Talib Abdallah, et al.)
- A positive feedback loop between ZEB2 and ACSL4 promote metastasis in BRCA (Lin, et al.)
- Dysregulation of proteins implicated in EMT (CDH1, SNAI2, TWIST1, VIM, and ZEB1) cause mesenchymal properties such as invasiveness and metastasis formation (Savci-Heijink)

Conclusions and Ethical Implications:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

Conclusions

- K-mean clustering: The 5 genes do separate the samples into two distinct clusters from the graph with orange dots indicating metastasis and the green not. The silhouette score of 0.323 does indicate moderate separation

- PCA: When comparing the K-means clusters with the PCA from the clinical outcomes there is significant overlap between the M0 and M1 with the red and blue dots.

Ethical Implications

- Risk of misdiagnosis: Since the model is not very accurate iterating it before people actually use it is important.

- Patient confidentiality: As this data is all personal information from specific patients not giving away any of it is also important.
- Generalizability: There were 100 data points in the subset indicating this might not be generalizable to everyone with breast cancer

Limitations and Future Work:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

Future Work

- Expanding the genes: We only used 5 genes and there are so many more that can have an affect on metastasis that we should also include into our model to make it more effective. These genes can include SLUG, FOX C2, STAT3 and others.
- Improve clinical correlation: As our clusters did not really fit the clinical data we had we shoudl try to do more with our dataset so we can get the sihouettle score higher and that our k-clusters get closer to fitting the clinical M0 and M1.
- Find a data set with more data points in general as there were less than 100 data points for breast cancer which is not great for a proper ML model as you cannot generalize conclusions for over millions of people with 80 data points.
- Find a data set with more data points for M1. In this data set about 4 of the data points had metastasis making it very skewed and harder to get a good ML model.

Limitations

- We limited the genes we used to 5 when there are so many and we could have missed one that had a impact
- Small sample size of M1 - data is imbalanced
- Generalizibility - Only to BRCA and not all types of breast cancer

NOTES FROM YOUR TEAM:

This is where our team is taking notes and recording activity.

Check point 1

- We finished the data set section and the background information for breast cancer. Additionally we came up with our question and figured out what variables we would be using for it.

Check point 2

- We started data analysis. We picked the method we wanted and choose linear regression we than cleaned the data and made a linear regression model. however since we have a small amoutn of data points a cox linear model was said ot be used whihc was also done.

Checkpoint 3

- We did all of the clustering and linear regression for the 5 genes now and found that the model worked better with clustering and unsupervised rather than the supervised.

Final Checkpoitn

- We finished the conclusions, ethical implications and future work.
- We removed all the linear regression as it was supervised and we had categorical data

QUESTIONS FOR YOUR TA:

These are questions we have for our TA.

- No questions at this time