Project Phase II

Pancreatic Cancer Organoid Profiling – Complete Project Documentation

By: Wahed Shaik – 60302087

Title: Predictive Modeling & Advanced Analytics on RNA-Seq Organoid Profiles

---

1. Introduction

Phase II builds directly on the data engineering foundation established in Phase I.
After transforming raw RNA-Seq count matrices into clean, structured **Bronze**, **Silver**, and **Gold** tables, the focus of Phase II is:

- **Applying unsupervised learning** to explore feature-driven structure
- **Building predictive models** for organoid-level biological properties
- **Evaluating model performance** using quantitative accuracy metrics
- **Producing visual analytics** (PCA, clustering, feature distributions)
- **Interpreting biological and data-driven meaning** behind extracted features

This phase completes the full **ETL → Analytics → Modeling** workflow expected in a modern biomedical data project.

---

2. Objectives of Phase II

The specific goals of Phase II were:

1. Build sample-level predictive models

Using engineered features such as **library size**, **% zero genes**, **avg counts**, and others.

2. Perform unsupervised exploratory modeling

Including **Principal Component Analysis (PCA)** and **KMeans clustering**.

3. Quantify model accuracy using industry-standard metrics:

- **R^2**
- **MAE**
- **MSE**
- **RMSE**
- **MAPE**

4. Generate curated Gold-layer modeling outputs

Stored back in Delta Lake for downstream **Power BI dashboards**.

5. Create interpretable visualizations

To illustrate structure, clusters, and model behavior.

---

3. Modeling Pipeline Overview

The modeling pipeline began by loading the **Gold-layer normalized_counts** and **sample_features** tables created in Phase I.

From these, we engineered a final feature matrix with the following sample-level predictors:

- **avg_count**
- **num_zero_genes**
- **num_detected_genes**
- **pct_zero_genes**

The primary prediction target was:

- **library_size** (sequencing depth)

The pipeline included:

Unsupervised Learning

- **Standard scaling**
- **PCA (2 components)**
- **KMeans clustering (k = 3)**

Supervised Learning

- **Random Forest Regression**
- **Train/test split**
- **Accuracy evaluation**

Final results were stored in the Gold layer at:
gold/sample_predictions

This table includes PCA coordinates, cluster labels, predicted values, and performance metrics.

---

4. Unsupervised Modeling

4.1 Principal Component Analysis (PCA)

PCA was applied to the standardized feature matrix.

The first two principal components explained:

- **PC1:** 84.2% of variance
- **PC2:** 15.7% of variance

Together, this represents **99.9% of total variance**, meaning the dataset can be almost entirely represented in two dimensions.

Interpretation

PC1 strongly correlated with:

- **Library size**
- **Number of detected genes**
- **Mean expression**

PC2 captured subtler variation related to **sparsity** and **sequencing noise**.

4.2 KMeans Clustering

Using the PCA-reduced feature space, **KMeans (k = 3)** was applied.

Observed Cluster Patterns

- **Cluster 0:** Samples with unusually high gene detection and atypical sparsity
- **Cluster 1:** Medium-depth sequencing libraries forming a dense middle cluster
- **Cluster 2:** Majority of organoid samples with typical RNA-Seq characteristics

Cluster labels were saved in **gold/sample_predictions** for Power BI visualization.

---

5. Supervised Predictive Modeling

A **Random Forest Regressor** was trained to predict **library_size** using engineered features.

5.1 Model Training Parameters

- **Train/test split:** 80% / 20%
- **n_estimators:** 200
- **random_state:** 42
- **Parallelization:** n_jobs = -1

Random Forests were chosen because they:

- Handle **non-linear** relationships
- Are robust to **outliers**

- Do not require feature scaling
- Perform well on **biological count-derived** features

---

6. Model Performance Metrics

Below are the exact metrics produced from the Databricks modeling run:

MODEL ACCURACY METRICS

R^2 Score: 0.9719

MAE: 87294.95

MSE: 12837985974.22

RMSE: 113304.84

MAPE: 3.31%

Interpretation of Accuracy

- **R^2 = 0.9719**
  The model explains **97.19%** of the variance in library size — extremely strong performance.
- **MAE ≈ 87k**
  On average, predictions differ from true values by only ~87,000 reads (small relative to multi-million reads).
- **RMSE ≈ 113k**
  Confirms the model maintains low error across all samples.
- **MAPE = 3.31%**
  Predictions deviate from true values by just **3%** on average — outstanding for biological data.

Conclusion:
The engineered sample features from Phase I provide a **highly predictive representation** of sequencing depth.

---

7. Gold Layer Modeling Output

The final Gold output table contains:

Per-Sample Data

- **PC1**, **PC2**
- **Cluster labels**

- **Predicted library_size**
- **Actual library_size**

Global Model Metrics

(Repeated per-row for Power BI compatibility)

- **model_r2**
- **model_rmse**

Stored at:
gold/sample_predictions

This dataset powers all Phase II Power BI visuals.

---

8. Power BI Integration (Phase II Visuals)

The following visuals were created in Power BI:

1. PCA Scatter Plot

Shows sample separation in reduced-dimensional space.

2. KMeans Cluster Visualization

Color-coded grouping of organoid samples.

3. Predicted vs Actual Library Size Plot

Shows strong agreement between model predictions and true values.

4. Feature Distribution Visuals

Histograms and scatterplots of **avg_count**, **num_zero_genes**, **pct_zero_genes**, etc.

Together, these plots provide statistical and biological interpretability.

---

9. Challenges Faced in Modeling Phase

1. RNA-Seq Sparsity

Zero-inflation required normalization and careful feature engineering.

2. No Drug Response Labels

Prediction was limited to QC-related targets (library size).

3. High Feature Correlation

Required PCA for proper visualization and interpretation.

4. Large Data Processing Requirements

Unpivoting and normalization required distributed Spark compute.

All challenges were systematically resolved.

---

10. Conclusion

Phase II successfully extends the RNA-Seq ETL pipeline into a **complete analytical and machine learning system**.

The project now includes:

- Clean, high-quality curated datasets
- Engineered gene- and sample-level features
- PCA + clustering for exploratory modeling
- A strong predictive model ($\mathbf{R^2 \approx 0.97}$)
- Gold-layer modeling outputs
- Power BI dashboards integrating results

This completes the workflow from
raw data → curated data → modeling → insights,
and establishes a strong foundation for future enhancements, including:

- Biomarker discovery
- Drug response prediction
- Classification modeling
- Differential expression analysis

The project fully meets—and exceeds—the requirements of **Project Phase II**.

---

11. References

- **AWS Open Data Registry – Organoid Pancreatic Dataset**
  https://registry.opendata.aws/organoid-pancreatic
- **dbGaP Study phs001611.v1.p1**
  https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001611.v1.p1
- **Ensembl GRCh38 Release 109 GTF**
  https://ftp.ensembl.org/pub/release-109/gtf/homo_sapiens/