**Project:- Cancer Type Prediction from Gene Expression**

**Goal:-**

*To build a machine learning model that can predict the type of cancer based on a patient's gene expression data.*

*Think of gene expression data like a snapshot of which genes are "active" or "inactive" in a cell. Different types of cancer have different gene activity patterns. We’ll use those patterns to identify the cancer type.*

**Why is this project useful?**

* *Helps in early diagnosis of cancer types.*
* *Supports personalized medicine (targeted treatments).*
* *Provides hands-on experience with real-world bioinformatics data.*

**Skills:-**

1. *Data Preprocessing*
   * *Clean and prepare large biological datasets.*
   * *Handle missing values, normalize gene expression.*
2. *Dimensionality Reduction (PCA)*
   * *Reduce thousands of genes into a smaller set of useful features.*
   * *Visualize data in 2D or 3D.*
3. *Classification*
   * *Use machine learning models (like Logistic Regression, Random Forest, SVM) to classify the cancer type.*
4. *Model Evaluation*
   * *Use metrics like accuracy, precision, recall, confusion matrix to evaluate performance.*

**Dataset may use:-**

* *TCGA (The Cancer Genome Atlas): A large public dataset of various cancer types and their gene expression.*
* *Kaggle RNA-Seq Data: Easier to access and beginner-friendly datasets with similar data.*

**Libraries may use:-**

* 1. *pandas*
  2. *numpy*
  3. *matplotlib.pyplot*
  4. *sklearn*

**Expected Outcome**

* *A model that can predict cancer type (like lung, breast, or kidney cancer) based on gene expression.*
* *A visualization showing how different cancer types form clusters.*
* *Understanding of how real-world biological data is handled in machine learning.*

**Summary:-**

| **Step** | **What You Do** |
| --- | --- |
| Load Data | *Gene expression values + cancer type* |
| Clean/Prep | *Handle missing values, scale features* |
| Reduce Dimensionality | *Use PCA to reduce to 2 or 3 features* |
| Train Model | *RandomForest/SVM/LogisticRegression* |
| Evaluate | *Accuracy, Confusion Matrix, Classification Report* |
| Visualize | *PCA plot colored by cancer type* |

**Daily Work Plan (02/07 - 13/07)**

**02/07/2025 – (Project Details Day):-**

* *Define the project and finalize resources*

**03/07/2025 – (Dataset & Preprocessing):-**

* *Download and explore the dataset.*
* *Understand the structure: rows = samples, columns = genes + cancer type.*
* *Clean the data (handle missing values, remove duplicates).*
* *Normalize gene expression values using StandardScaler or MinMaxScaler.*

**04/07/2025 – (Feature Engineering & PCA):-**

* *Apply PCA to reduce features (e.g., from 10,000+ genes to 2–100 components).*
* *Visualize PCA components in 2D plot (color-coded by cancer type).*
* *Save the transformed dataset for model input.*

**05/07/2025 – (Model Building):-**

* *Train initial classification models:*
  + *Random Forest*
  + *SVM or Logistic Regression*
* *Use train\_test\_split() and evaluate with accuracy, precision, recall.*
* *Save models using joblib or pickle.*

**06/07/2025 – (Draft 1 Submission):-**

**Deliverables:**

* *Cleaned & processed dataset*
* *PCA visualization*
* *At least one trained model*
* *Evaluation metrics (confusion matrix, accuracy report)*
* *Summary of what’s done so far*

**07/07/2025 – (Model Improvement):-**

* *Try other models (e.g., Gradient Boosting, k-NN).*
* *Tune hyperparameters (e.g., using GridSearchCV).*
* *Compare model performances and choose the best one.*

**08/07/2025 – (Error Analysis & Feature Importance):-**

* *Analyze model errors (misclassifications).*
* *Use feature importance from Random Forest or SHAP to understand key genes.*
* *Write brief interpretation of model behavior.*

**09/07/2025 – (Report Draft):-**

* *Start writing your technical report or markdown notebook:*
  + *Introduction*
  + *Dataset description*
  + *Methods (PCA, models)*
  + *Results & evaluation*
  + *Visualizations*
  + *Conclusions*

**10/07/2025 – (Draft 2 Submission):-**

**Deliverables:**

* *Improved models with tuning*
* *Feature importance/insight*
* *Clean report (Jupyter or PDF)*
* *Graphs & confusion matrix updated*

**11/07/2025 – {Build UI (Start)}:-**

* *Use Streamlit or Tkinter for a basic UI:*
  + *Upload gene expression sample (CSV)*
  + *Predict cancer type using your trained model*
  + *Show output and probability scores*

**12/07/2025 – (UI Finalization):-**

* *Polish UI with:*
  + *Title, instructions*
  + *Upload button + prediction*
  + *Show result in a user-friendly format*
* *Test the UI for a few sample inputs.*

**13/07/2025 – (Final Project Submission):-**

**Deliverables:**

* *Complete project with:*
  + *Final model*
  + *Full report*
  + *Working UI*
* *Folder organized and ready for presentation/submission*