BRACT's
Vishwakarma Institute of Information Technology , Pune-411048
Department of Artificial Intelligence and Data Science

Major Project Review
On
“Ovarian Cancer Detection”

Guided By
Prof. Yashwant Ingle

PRESENTED BY

Sr. NO.	Roll No.	Name	PRN No.
1	471001	Aarya Umbare	22111339
2	472007	Apurva Belsare	22110511
3	472013	Aditi Devade	22110354
4	472025	Pradnya Kedari	22110926

Introduction

Ovarian cancer is a highly aggressive and deadly gynecological disease often detected at advanced stages due to the lack of early symptoms.

Traditional methods like imaging, biopsies, and blood tests are often inadequate for early detection.

Leveraging advancements in bioinformatics and machine learning, this study uses models such as KNN and Random Forest to analyze gene expression data, aiming to improve early diagnosis, identify critical biomarkers, and enable personalized treatment strategies.

Motivation

- Enhancing Early Detection: The project aims to improve early ovarian cancer detection by using data science to identify subtle patterns in diverse datasets, addressing late-stage diagnoses and improving patient outcomes.

- Comprehensive Dataset Exploration: It leverages diverse data types like demographics, genetics, imaging, and clinical records to uncover relationships and patterns for a holistic understanding of ovarian tumors.

- Machine Learning for Diagnosis: Machine learning models are applied to ovarian cancer data to enhance diagnostic accuracy, aiding healthcare professionals in decision-making.

Literature Survey

Sr.	Paper Title	Year	Methodology	Key Findings
1.	Multi-Modal Evolutionary Deep Learning Model for Ovarian Cancer Diagnosis	2021	CNN , LSTM , Ant Lion Optimizer	Hybrid model, Optimized networks, Superior performance
2.	Diagnosing Ovarian Cancer on MRI: A Preliminary Study Comparing Deep Learning and Radiologist Assessments	2022	CNN	DL approaches show promise in improving diagnostic accuracy
3.	New Trends in Ovarian Cancer Diagnosis Using Deep Learning: A Systematic Review	2024	CNN , NN , LSTM	High Mortality Rate, Risk Factors, Imaging Techniques, AI in diagnosis
4.	Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches	2022	LR , DT , Random Forest , LGBM	Significant biomarkers for ovarian tumors include CA125 and HE4

Literature Survey

Sr.	Paper Title	Year	Methodology	Key Findings
5.	Al-Powered classification of Ovarian can□cers Based on Histopathological lmages	2024	CNN , ResNet50	Developed an AI model to classify the major types of EOC, indicating the promise of AI in cancer
6.	Prediction of ovarian cancer using artificial intelligence tools	2023	DT, SVM, AINNM	The study identified HE4, CA125, and NEU as most significant predictive factors for diagnosis, as determined by both Information Gain and Gini Index methods.
7.	Improved Prediction of Ovarian Cancer Using Ensemble Classifier and Shaply Explainable AI	2023	LDA, CART, LR, KNN, SVM	High Accuracy,Ensemble Learning,Explainable AI
8.	Artificial Intelligence in Ultrasound Diagnoses of Ovarian Cancer: A Systematic Review and Meta-Analysis	2024	KNN, SVM	Histopathological Findings as Gold Standard,Promising Role of AI in Clinical Practice

Gap Analysis

- Data Availability and Quality: While diverse datasets exist (e.g., patient demographics, genetics, imaging), there may be gaps in data completeness, quality, or consistency, which can affect model accuracy.
- Model Generalization: The effectiveness of machine learning models on diverse patient populations and in real-world clinical settings may be limited by overfitting or insufficient training data from underrepresented groups.
- Interpretability and Trust in AI: While machine learning can improve diagnostic accuracy, healthcare professionals may face challenges trusting AI-driven insights, especially if the models lack transparency and interpretability in decision-making.

Problem Statement

The objective of this project is to develop predictive models that accurately classify ovarian cancer into subtypes, including 'control', 'endometrioid', 'serous', 'mucinous', 'clear cell', and 'case'. By leveraging a comprehensive analysis of gene composition, including gene expression levels and nomenclature, the goal is to create data-driven tools capable of improving ovarian cancer detection and diagnosis.

Objectives

- Develop reliable and accurate predictive models by integrating diverse genetic information, ensuring generalizability across different patient populations.

- Focus on detecting ovarian cancer at its earliest stages by identifying subtle gene signatures and patterns for more timely and accurate diagnoses.

- Accurately classify ovarian cancer into subtypes like 'control', 'endometrioid', 'serous', 'mucinous', 'clear cell', and 'case' to refine diagnosis and treatment strategies.

Proposed Solution/Methodology

STEP 1 - Data Collection

- Dataset: Used the Ovarian Cancer dataset from Kaggle, containing gene expression data for various ovarian cancer subtypes.
- Data Loading:

1. Loaded multiple CSV files using pandas, each representing different gene types with associated gene names and expression levels for various cancer subgroups.
2. Dynamically named DataFrames for each CSV file to maintain a standardized structure.
3. Renamed the first column in each dataset to gene for consistency.

STEP 2 - Exploratory Data Analysis (EDA)

- Performed a thorough examination of the dataset to gain insights into the gene expression data.
- Explored unique gene group values within each DataFrame to assess the diversity of gene types.
- Identified and corrected data anomalies, such as correcting a typo ('cler cell' to 'clear cell') in the gene subtypes.
- Visualized distributions of gene expressions across different cancer subtypes.

STEP 3 -

- Data Transformation
- Data Splitting
- Feature Scaling
- PCA
- Oversampling for Imbalanced Data

STEP 4- Model Development

- ·K-Nearest Neighbors (KNN): Best accuracy of 78.19% with tuned parameters including n_neighbors, weights, and p.
- ·XGBoost: Achieved 73.04% accuracy by optimizing n_estimators, max_depth, and learning_rate.
- ·Support Vector Machines (SVM): Used the RBF kernel, achieving 30.39% accuracy.
- ·Decision Tree: Achieved 62.75% accuracy with parameters such as criterion, max_depth, and min_samples_split.
- ·Random Forest: Achieved 68.92% accuracy with tuned parameters including n_estimators and max_depth.
- ·Logistic Regression: Used as a baseline model with an accuracy of 55.64%.

Software Requirements

Programming Language:

- Python: For data preprocessing, machine learning, and evaluation.

Libraries and Frameworks:

- Data Handling and Visualization: Pandas, NumPy, Matplotlib, Seaborn.
- Machine Learning: scikit-learn for models like SVM and KNN.

Development Platform:

- Google Colab: For cloud-based coding and GPU support.

Implementation

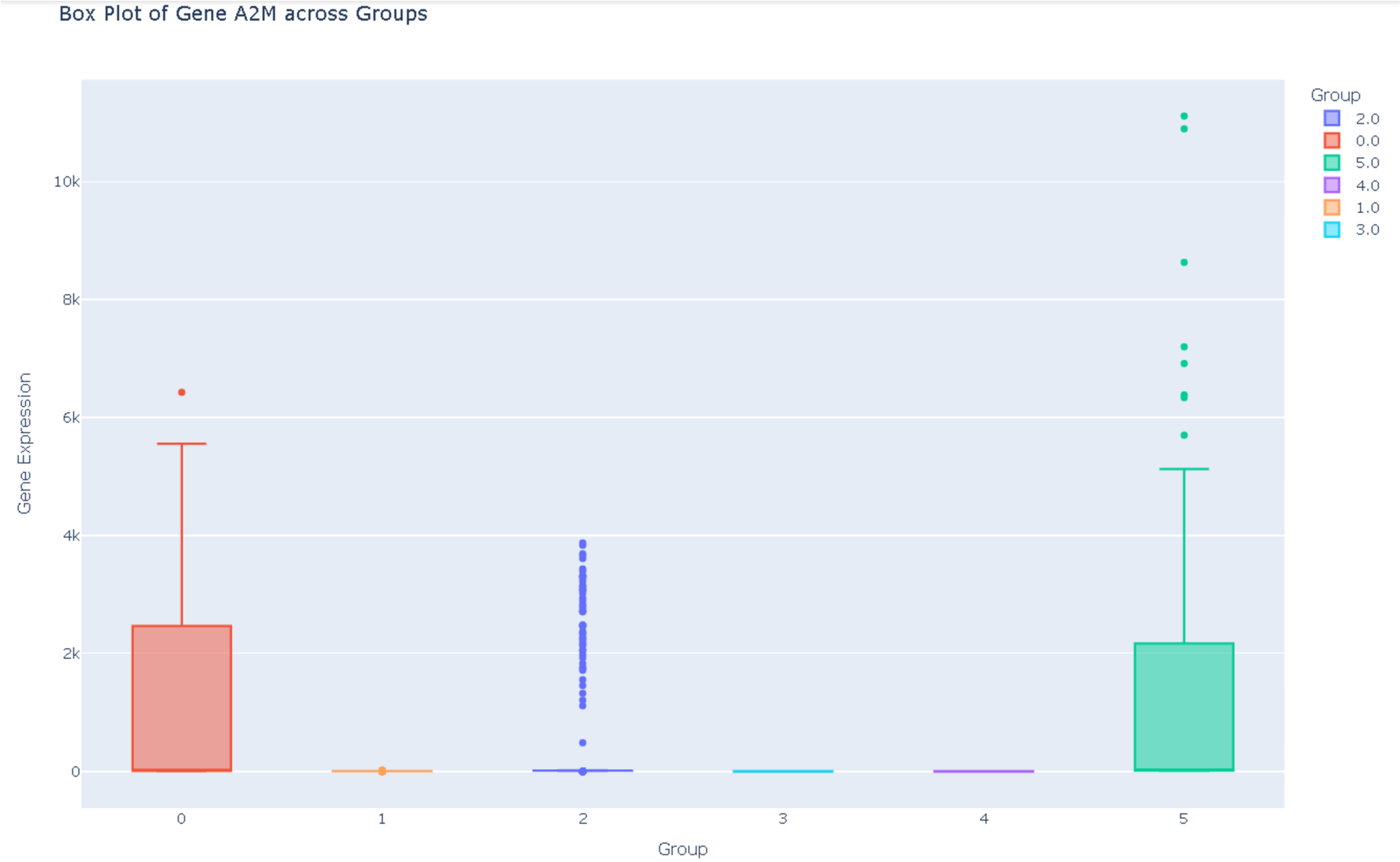
Data Collection

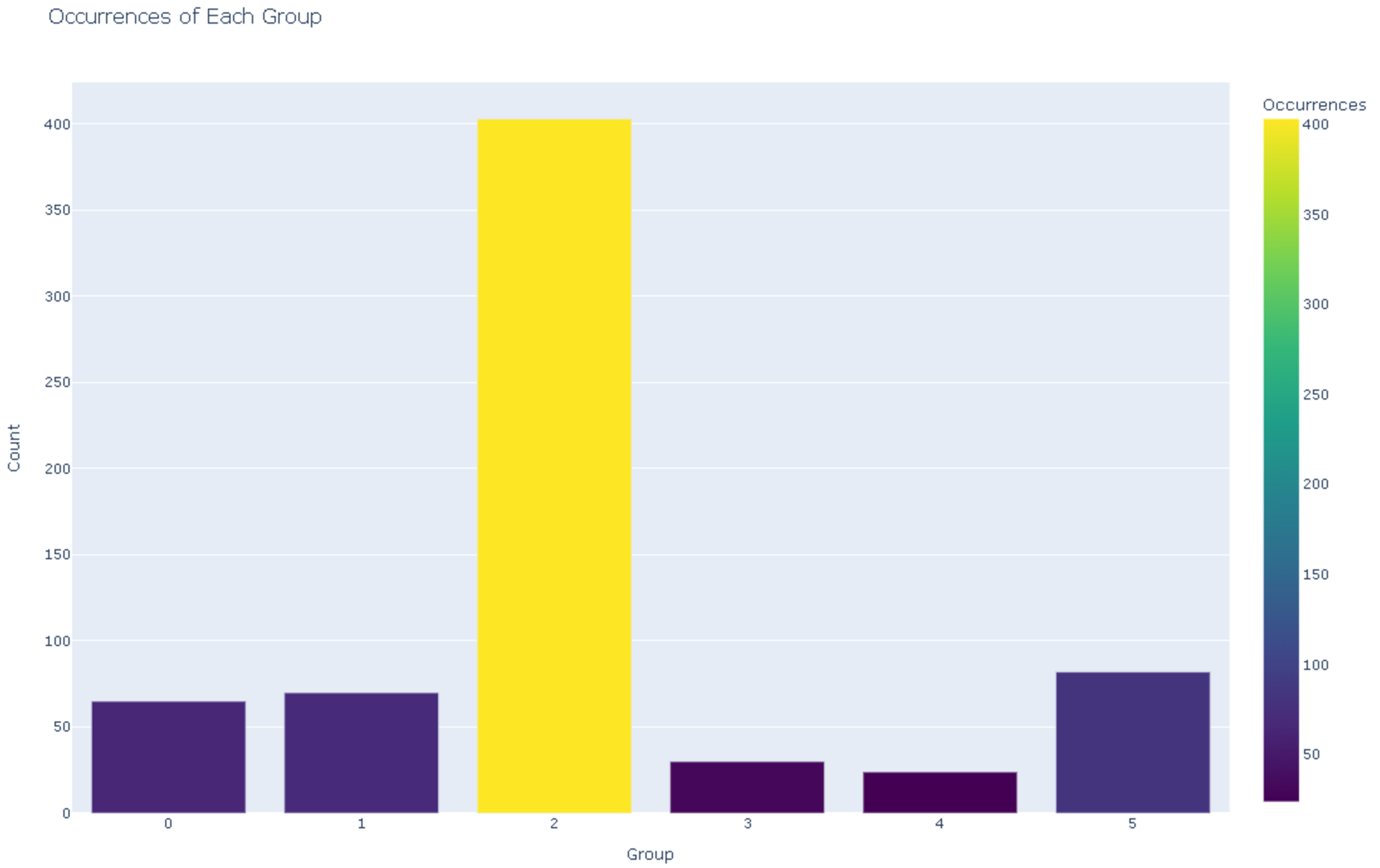
Preprocessing

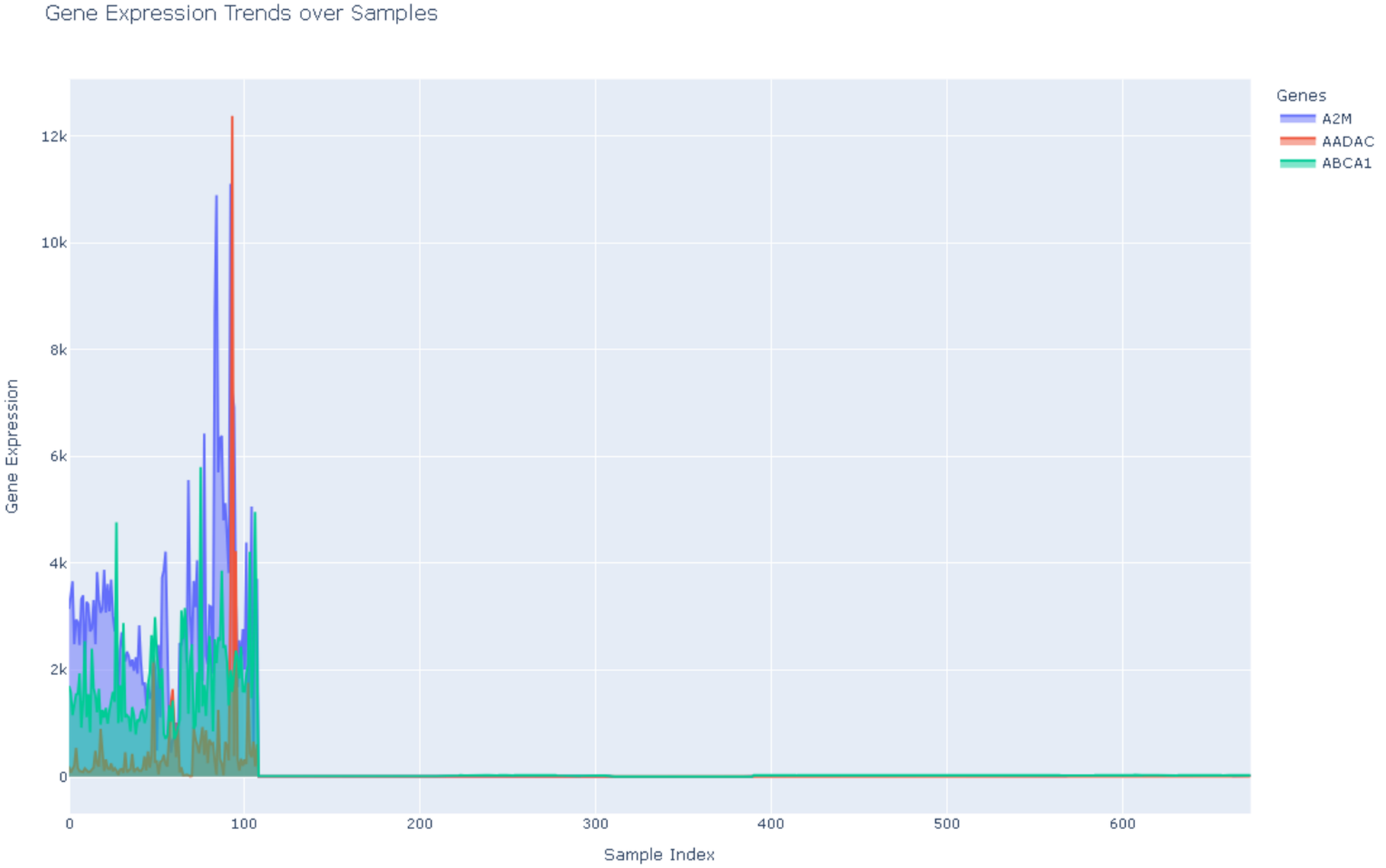
Dimensionality Reduction (PCA)

Model Training

Evaluation


Box Plot of Gene A2M across Groups
Group
2.0
0.0
5.0
4.0
1.0
3.0
Gene Expression
10k
8k
6k
4k
2k
0
0
1
2
3
4
5
Group


Occurrences of Each Group
Count
Group
0
1
2
3
4
5
0
50
100
150
200
250
300
350
400
Occurrences


Gene Expression Trends over Samples
Genes
A2M
AADAC
ABCA1
Gene Expression
0
2k
4k
6k
8k
10k
12k
0
100
200
300
400
500
600
Sample Index

Data Collection and Preprocessing

- Dataset: Gene expression data for ovarian cancer subtypes, including labels such as 'control', 'serous', 'endometrioid', etc.
- Preprocessing Steps:
  - Handling missing values
  - Normalization/standardization of data
  - Data splitting into training, validation, and test sets
  - Oversampling techniques (e.g., SMOTE) to address class imbalance
- Dimensionality Reduction (PCA)
  - Principal Component Analysis (PCA) used to reduce the number of features while preserving important variance in the data.
  - Visualized data using PCA to understand variance explained by the components.

- Model Selection and Training
  - Algorithms Used:
  - K-Nearest Neighbors (KNN)
  - Support Vector Machine (SVM)
  - Logistic Regression
  - XGBoost
  - Decision Trees
  - Training: Applied each algorithm to the preprocessed and PCA-reduced dataset.
- Hyperparameter Tuning
  - GridSearchCV used for hyperparameter tuning (e.g., KNN's n_neighbors, SVM's C and gamma).
  - Optimized models for better accuracy and generalization.
- Model Evaluation

  Metrics Used: Accuracy, F1-score, confusion matrix.

Results and Discussion

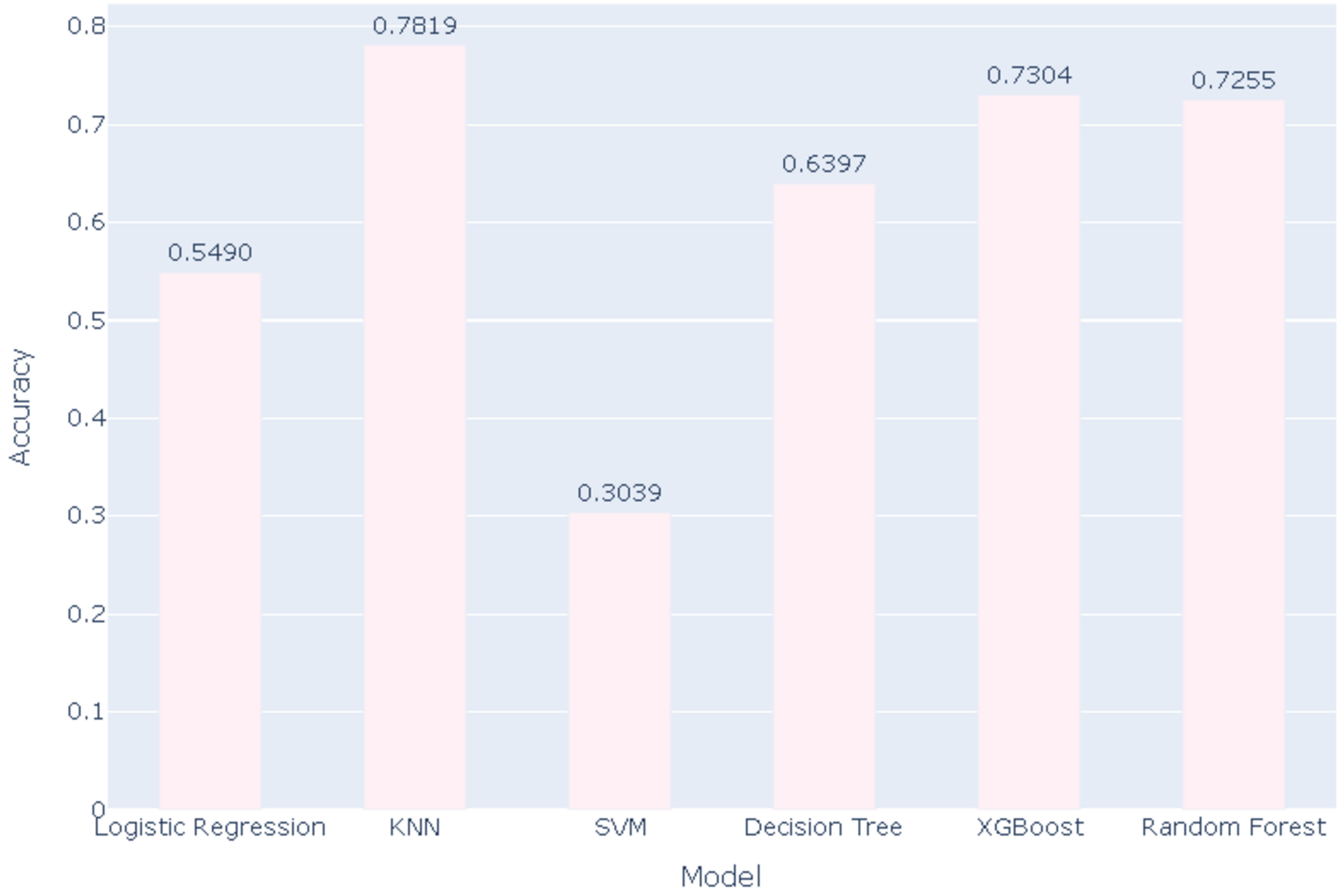
Model Comparison - Accuracy:

- KNN showed superior performance across classes, especially in "case" (Accuracy: 0.7819, F1: 0.99).
- Other models like XGBoost, Logistic Regression, and SVM performed less effectively, especially in minority classes like "mucinous" and "endometrioid."

Optimal Method:

- KNN emerged as the most effective algorithm based on test accuracy, F1-scores, and predictions validated by the confusion matrix.

Model Comparison - Accuracy



Conclusion

This project focused on classifying cancer subtypes through gene expression analysis. Principal Component Analysis (PCA) was used for dimensionality reduction, and machine learning models, including XGBoost, Decision Trees, Logistic Regression, KNN, and SVM, were evaluated.

Model performance was assessed using accuracy metrics, classification reports, and confusion matrices, with visualizations providing additional insights. A final comparison highlighted the strengths of each model, identifying the most effective approach for cancer subtype classification, laying a strong foundation for future applications.

Future Scope

- Integrate genomic and clinical data for improved predictions.
- Explore deep learning models for advanced analysis.
- Develop real-time tools for early cancer diagnosis.
- Enable personalized treatment planning based on subtypes.
- Scale the model to classify other cancer types.
- Collaborate with healthcare for clinical validation and adoption.

References

- ·Multi-Modal Evolutionary Deep Learning Model for Ovarian Cancer Diagnosis
- Rania M. Ghoniem, Abeer D. Algarni, Basel Refky, Ahmed A. Ewees.
- ·Diagnosing Ovarian Cancer on MRI: A Preliminary Study Comparing Deep Learning and Radiologist Assessments
- Tsukasa Saida, Kensaku Mori, Sodai Hoshial, Masafumi Sakai, Aiko Urushibara, Toshitaka Ishiguro, Manabu Minami, Toyomi Satoh, Takahito Nakajima.

Thank You