**Methodology Overview**

This methodology describes the entire process from data preparation to model evaluation and visualization in our machine learning analysis of genes associated with epilepsy (EP). Our approach includes data preprocessing, feature selection, model training, hyperparameter tuning, and visualization of the results to identify significant patterns in gene expression and variant pathogenicity.

**1. Data Collection and Preprocessing**

* **Gene List Collection**: We started by gathering a list of 320 genes associated with epilepsy and identified non-associated genes by excluding known epilepsy genes. The initial gene list was collected using variant screening criteria based on variant rarity, phenotype data, and other annotations.
* **Normalization Tools**:
  + **PyBiomart**: We used PyBiomart to retrieve information for annotation and data enrichment for our gene list. This tool allowed us to gather specific genomic details for each gene effectively.
  + **Data Normalization**: We applied normalization techniques using Python's **pandas** library to ensure the gene data features were on a common scale. This was necessary for improved model training and performance. We included tools such as **StandardScaler** from scikit-learn to standardize features like **CADD scores**, **log fold change** from transcriptome data, and **pathway involvement**.
* **Filtering Criteria for Non-Associated Genes**: The non-associated genes were selected based on the following criteria:
  + **Inclusion Criteria for Selection of Non-Associated Genes**:
    - **Presence in Epilepsy Exome Sequencing Data**: Variants in genes must be observed in more than 25% of the epilepsy exomes analyzed, indicating that these variants are commonly found in individuals with epilepsy but are not necessarily disease-causing.
    - **No Known Epilepsy Association**:
      * **Pathway Analysis**: Genes must not participate in any pathways significantly linked to epilepsy or related neurological disorders.
      * **Phen2Gene Analysis**: Genes must score low or zero, suggesting no association with epilepsy.
      * **MGI Phenotype Analysis**: Genes should not be associated with epilepsy-related phenotypes in the MGI database.
      * **Literature Review**: Genes must not have any documented association with epilepsy based on available literature.
    - **Brain Tissue Expression**: Genes must show expression in brain tissue (e.g., through RNA expression data from Human Protein Atlas or GTEx). This ensures that selected genes are relevant to brain biology.
  + **Exclusion Criteria for Selection of Non-Associated Genes**:
    - **Epilepsy Association**: Genes previously identified as associated with epilepsy through transcriptomic analysis, pathway analysis, Phen2Gene, MGI phenotype analysis, or literature evidence are excluded to ensure these genes do not appear in the non-associated list.
    - **Epilepsy Pathway Involvement**: Genes involved in any epilepsy-related pathways are excluded, as identified through pathway analysis.
    - **Gene Selection from Brain Tissue Expression Data (Post Exome Screening)**: Genes that do not show expression in brain tissue based on RNA expression analysis are excluded. This helps ensure that non-associated genes are still biologically relevant to the tissues involved in epilepsy.

**2. Data Integration and Feature Engineering**

* **Feature Engineering**:
  + We included features such as **CADD scores** (used to predict pathogenicity of variants), **log fold change (logFC)** values from RNA-seq data for brain tissues, **MGI Phenotype annotations**, **Phen2Gene scores**, and **number of pathways involved**.
* **Data Storage and Handling**:
  + **Openpyxl**: We used Openpyxl to handle the data in Excel files for easier data exchange and visualization during intermediate analysis steps.
* **Data Splitting**:
  + **train\_test\_split** (scikit-learn) was used to split the dataset into training and testing sets, ensuring that we had a robust evaluation setup.

**3. Model Building and Training**

* **Model Selection**:
  + We initially selected a **Decision Tree Classifier** for a simple interpretative approach and then extended our work to a more complex model, the **Random Forest Classifier**.
* **Training the Decision Tree**:
  + We trained a decision tree using scikit-learn's **DecisionTreeClassifier**, setting a maximum depth to avoid overfitting and improve interpretability.
  + We extracted the tree's nodes and edges to visualize the decision-making process, showing how features like **Phen2Gene score** and **CADD scores** influenced classification.
* **Random Forest Classifier**:
  + For better accuracy and generalizability, we used **RandomForestClassifier** from scikit-learn. We employed **GridSearchCV** to tune hyperparameters like n\_estimators, max\_depth, and min\_samples\_split to find the best-performing model configuration.
  + **Cross-Validation**: **cross\_val\_score** was used to evaluate the model performance using k-fold cross-validation.

**4. Model Evaluation**

* **Metrics Used**:
  + **ROC-AUC Score**: The model's performance was primarily evaluated using the **ROC-AUC score**, which reached a high value (e.g., 0.9815), indicating good discrimination ability between epilepsy-associated and non-associated genes.
  + **Confusion Matrix**: We generated a **confusion matrix** using **ConfusionMatrixDisplay** to assess true positives, false positives, true negatives, and false negatives in model predictions.
  + **Precision-Recall Curve**: This was used to better understand model performance in light of any data imbalance.
* **Feature Importance**:
  + We determined the importance of each feature using the **feature\_importances\_** attribute from RandomForestClassifier. The results showed **Phen2Gene Score** as the most critical predictor, followed by **CADD Score** and **MGI Phenotype annotations**.

**5. Visualization**

* **Feature Importance Plot**:
  + We plotted feature importance using **matplotlib** and **seaborn** libraries, allowing us to visualize the significance of each feature in predicting epilepsy association.
* **ROC Curve and Precision-Recall Curve**:
  + We plotted the **ROC Curve** and **Precision-Recall Curve** to assess model sensitivity and precision.
* **Decision Tree Visualization**:
  + We visualized the decision tree using **graphviz** and **pydotplus**, creating a comprehensive diagram of decision splits. Later, we also generated a more visually appealing tree using **Plotly**, similar to a network graph.
* **Network Graph in Cytoscape**:
  + Finally, we exported a file that could be imported into **Cytoscape** to create custom visualizations of the relationships between different features and genes, allowing for a more interactive exploration of the data.

**6. Filtered Genes Identification**

* **Weighted Score Calculation**:
  + We assigned a weighted score for each gene based on model predictions and the importance of features, ranging from 0 to 1.
* **High Confidence Gene Score Filtration**:
  + Genes were ranked based on their weighted scores:
    - **Rank 1**: Genes with a weighted score > 0.2 and all other features (such as CADD Score, Phen2Gene Score, and log fold change) showing values > 0.1.
    - **Rank 2 and Beyond**: Genes were further ranked based on descending weighted scores, considering lower thresholds for feature values.
* **Confidence Score**:
  + We filtered genes based on a relaxed **confidence score** threshold, slightly reducing the cutoff due to the observed distribution of scores, thus retaining more genes for downstream validation.
* **Detailed and High-Confidence Genes**:
  + The final filtered list of genes was saved into **complete\_filtered\_genes.csv** for all genes and **high\_confidence\_genes.csv** for those passing the confidence score threshold.

This ranking method helps in categorizing genes based on their association strength and feature values, giving a clear pathway for prioritizing genes for further analysis. Let me know if this addresses your requirement, or if you need further adjustments.

### 7. \*\*Conclusion\*\*

- Our machine learning approach successfully identified significant genes potentially linked to epilepsy by combining transcriptome data, variant pathogenicity scores, and feature importance derived from a random forest model. The decision tree and other visualizations provided insight into feature interactions and the decision-making process, which can be further explored using Cytoscape for network-based analysis.

- The results showed that \*\*Phen2Gene Score\*\*, \*\*CADD Score\*\*, and \*\*MGI Phenotype annotations\*\* played significant roles in distinguishing epilepsy-associated genes, demonstrating the utility of our approach in variant prioritization and gene characterization.