Assignment 3
Cheyanne Durham
Analysis Report
Github Link: https://github.com/cheyannewdurham/CPBS7602

Dimensionality and Clustering

Data Preparation

I wanted to be able to compare the effects of transforming the data (scaling using StandardScaler before moving on with my analysis. For this section of the code, I made sure to retain two data frames of the scaled and unscaled data.

PCA Exploration

To fully understand the effects of scaling and the PCA plots, I looked at how well the genes cluster for both the scaled and unscaled data. Using a linear regression model, I fit a coordinate plate to see how the points were transformed to fit into the PCA components for the scaled and unscaled data. ChatGPT helped me create the mesh and the formulas for the projections. Unsurprisingly, the PCA retained more of the relationships between the scaled data points across multiple components vs the unscaled - understandable given their smaller overall distribution and deviation. The variance prediction was consistent with these qualitative finding, as the PCA unscaled was able to explain more of the variance along PC1 while the variance in the PCA for the scaled was distributed throughout the the various components/dimensions. Given the nature of the assignment, I thought it was worth exploring both kinds of data throughout the computational tasks.

UMAP Exploration

After trying several different UMAP settings, the scaled data gave the best visual cluster presentation. This stage serves as a computational benchmarks to measure the subsequent clustering improvements against. 2 components, 3-7 neighbors, and min distance of 0.3 seemed to preserve the local and global relationships the best as analyzed by just personal interpretation (at this stage).

MDS Exploration

To visualize the data as clusters with likely many overlapping groups, MDS was used along with ConvexHull from the SciPy.spatial package. This gives a clearer idea of the boundaries of each cluster (likely tissue specific) and the pairwise distances between points. From the scaled data, we were are able to see a few distinct points where each tissue type may differ. Unscaled data almost looks like it preserves hierarchical data (broader clusters representing a larger biological group) and shows that the skin and blood may have the most varied gene expression. This further indicates a potential need for data scaling in this data set (given the overlap) when using certain visualization techniques.

PCA, UMAP, and Spectral Clustering

Parameters for the best dimensionality for PCA and UMAP presentation before spectral cluster analysis were explored. The highest silhouette scores came from the rbf kernel, which is very effective when catching non-linear relationships in the data. Other important aspects were use of the scaled data frame, a small number of PCA components (reduce noise), n_neighbor of 10, and minimum distance of 0.1. A 3 component UMAP also had a better silhouette score than a 2 component. Moving forward, the dimensions and clusters with the best silhouette scores were used as the latent spaces. That said, the elongated nature of some of the clusters may indicate a high level of within-cluster variance which could mean poor separation despite the high silhouette score.

Ensemble-based Models to Predict Tissue

Random Forest and Bagging with a Decision Tree were used as our overfitting prone model and weak learning model respectively. They were evaluated using precision, recall, f1-score, support, accuracy, macro average, weighted average, and heat maps by tissue classification to understand accuracy of the generalization with untrained data. Unsurprisingly, the random forest heavily outperformed bagging. Clusters 4 and 5 tend to be our noisiest clusters with the most overlap - this is seen across both learning models. Random forest handled it much better, keeping all F1-scores and recall scores above 0.65 (false positives and negatives respectively) while maintaining high accuracy. The bagging model could not predict with any accuracy or precision and preformed poorly on accuracy and recall for clusters 0, 4, 5, 6, 7. These clusters are likely highly overlapping in the data. One caveat here is that the Random Forest is likely better designed in my code than the bagging model, leading to its better performance overall.

Ensemble-based Models to Predict Age from Blood Samples

Despite applying both scaled and unscaled data, the R^2 scores for Random Forest and Gradient Boosting models on the 2D and 3D latent spaces remained below 0, with an RMSE of approximately 13. These results indicate that there is no observable correlation between blood gene expression and age. Given the robustness of these ensemble-based models, the lack of performance suggests that age-related signals in blood gene expression may be minimal or obscured. To validate these findings, correlation analyses (using both Pearson and Spearman coefficients) were performed. As expected, age perfectly correlates with itself but does not correlate with the UMAP dimensions, confirming the absence of both linear and monotonic relationships between age and the latent space features. GridSearch with cross-validation was used to determine best parameters for each model. The hyper parameters were further tuned to see if there was any significant change with both the scaled and unscaled data. No difference was found in either.

It is possible that the latent spaces, when clustering did not preserve a global relationships related to age. In the future, the initial clustering could include age if it is believed to be an important biological factor.. However, data leakage here may become a concern and inflate the performance of the models erroneously as measure by biological significance. Furthermore, Blood gene expression to age relationships are further obscured by the high amount diverse signals present in this “tissue” type biologically. These confounding signals create noise at such a magnitude that it would be very difficult to regress them out to get relevant clusters in other tissue types while maintaining relationships between blood’s gene expression and age.