Class Project Progress (11/14)

So far, I've built the core foundation of my Bayesian-network–based gene expression analysis pipeline using the GSE19804 lung cancer dataset. I began by downloading and preparing the data, extracting sample-level metadata (tumor vs. normal labels, patient IDs, and stage information), and performing initial quality checks. The dataset ultimately included 120 samples, 54,675 genes, perfectly balanced tumor/normal classes (60/60), and 60 unique patients, which confirms the dataset is suitable for patient-wise cross-validation.

After cleaning and organizing the dataset, I implemented the first part of the feature engineering pipeline. This included fold-wise variance filtering to select the top 20 most variable genes and a quantile-based discretization step to convert continuous gene expression values to discrete bins. These steps ensure that the later Bayesian network models work with stable and non-sparse discrete inputs.

With the cleaned data and feature pipeline ready, I constructed and trained my baseline Bayesian Network model (H1 Naive Bayes). This model assumes that all genes are conditionally independent given the class label Y (tumor vs normal), and it evaluates two possible dependency structures between Stage and Y using the Bayesian Information Criterion (BIC):

- Stage → Y
- Y → Stage

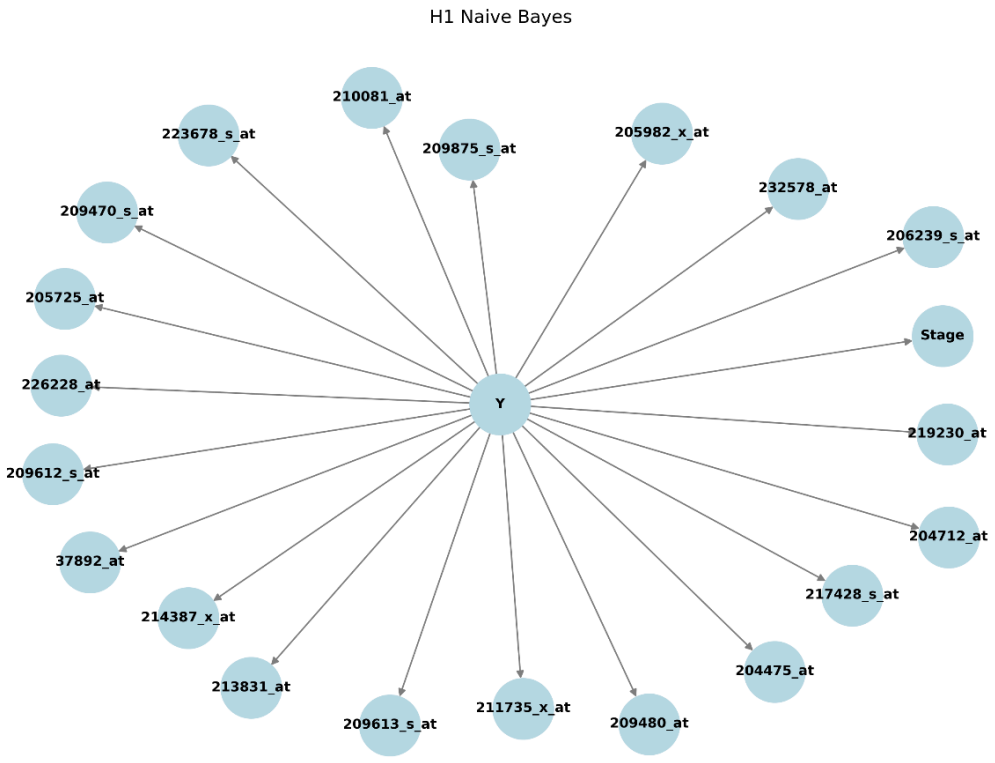
In every cross-validation fold, the model selected Y → Stage as the better-fitting direction, indicating that tumor status explains stage more strongly than the alternative in this dataset. I then ran a 5-fold stratified group cross-validation, where grouping by patient ID prevents leakage between training and testing sets. These are my cross-validation results below:

```
Fold 1/5
Stage→Y BIC: -3596.96 | Y→Stage BIC: -995.27
  Accuracy: 1.0000
Fold 2/5
Stage→Y BIC: -3591.71 | Y→Stage BIC: -993.89
  Accuracy: 0.9167
Fold 3/5
Stage→Y BIC: -3638.95 | Y→Stage BIC: -1036.08
  Accuracy: 1.0000
Fold 4/5
Stage→Y BIC: -3645.55 | Y→Stage BIC: -1046.76
  Accuracy: 1.0000
Fold 5/5
Stage→Y BIC: -3673.03 | Y→Stage BIC: -1074.25
  Accuracy: 0.8750

H1 Results:
  Accuracy: 0.9583 ± 0.0527
  Pooled Confusion Matrix:
[[57  3]
 [ 2 58]]
```

These results suggest the H1 model is performing extremely well, with only a few misclassifications across all folds.

I also visualized the learned H1 model structure. The network shows Y as the parent of all selected genes, and Y → Stage as chosen by the BIC comparison, which is consistent with the Naive Bayes framework and my cross-validated results.



What I plan to do next

Now that the H1 Naive Bayes model is complete and validated, the next steps in my project are:

1. Implement H2: TAN (Tree-Augmented Naive Bayes)
2. Implement H3: Fully learned Bayesian Network structure
3. Compare H1, H2, and H3