**Notes from meeting with Omar**

Link to the project

<https://www.kaggle.com/code/karthik0809/dementia-classification-random-forrest>

The developer chose different imputation methods for the SES and MMSE columns based on the characteristics and common practices for each type of data:

**Imputing SES with the Mode:**

* The SES (Socioeconomic Status) column seems to have discrete values, often representing categories or classes.
* Using the mode (most frequent value) is appropriate for categorical or ordinal data. This method preserves the most common value in the column and avoids skewing the distribution.

**Imputing MMSE with the Mean:**

* The MMSE (Mini-Mental State Examination) column appears to contain continuous scores.
* Using the mean is a common practice for continuous variables, as it minimizes the effect of missing values without altering the overall distribution.
* The mean provides a central tendency that reflects the overall dataset’s average score, making it suitable for filling in gaps in this context.

**In summary:**

* Mode for SES because it’s likely categorical/ordinal, where using the most frequent value is typical.
* Mean for MMSE because it’s continuous, and using the mean preserves the average trend in the data.

Based on the data, the EDUC column likely represents years of education. In research datasets, particularly those related to cognitive assessments and aging (like in dementia studies), EDUC commonly denotes the number of years an individual has completed formal education. This interpretation is supported by values such as 12, 14, and 18, which align with typical educational milestones (e.g., high school completion around 12 years, college around 16 years, etc.).

SKlearn Library   
<https://scikit-learn.org/stable/supervised_learning.html>

**Steps to be DONE**

### 1. Choose a Dataset and Justify

* **Dementia Dataset**: Focused on structural brain features from MRI and cognitive scores. It could be useful for conditions where age, brain volume, and cognitive function scores (like MMSE) are critical.
* **Parkinson's Dataset**: Based on voice features, helpful for distinguishing between affected and non-affected individuals via vocal variability.
*  Decide on either the Dementia or Parkinson's dataset, or both if feasible.
*  Justify your choice based on aspects like feature type (MRI for Dementia or voice features for Parkinson's), complexity, and how well each dataset supports probabilistic analysis.

For report:

Justifying your choice could involve discussing data relevance, feature complexity, or ease of processing for probabilistic reasoning.

### 2. Data Preprocessing (Data Cleaning)

* **Format and Discretize**: Since Bayesian Networks often work with discrete variables, you may need to **discretize continuous variables** (like age or eTIV) into bins (e.g., age groups or brain volume ranges). **Discretization vs. Continuous Processing**: Decide whether to discretize data (for discrete Bayesian Networks) or keep it continuous (for Gaussian Networks or Gaussian Processes). For example, age or MRI measurements could be discretized into ranges if you’re focusing on discrete Bayesian Networks.
* **Handle Missing or Noisy Data**: If data has missing values, you could use methods like **mean imputation** or **Mode imputation** to fill gaps.

### 3. Select Probabilistic Methods

* **Discrete Bayesian Networks**: Good for discrete or categorical variables, where conditional dependencies can be represented in a network structure (e.g., influence of MMSE on dementia diagnosis).
* **Gaussian Bayesian Networks**: If many variables are continuous, these networks model variables as Gaussian distributions, enabling continuous data usage with some discretization.
* **Gaussian Processes**: Useful for regression-based tasks or if predicting specific continuous outcomes (though often less common in straightforward classification tasks like dementia diagnosis). Might be used for **Parkinson's Dataset**
* **Dementia Dataset**:
  + **Feature Types**: This dataset has a mix of categorical (e.g., Group, M/F, Hand) and continuous variables (e.g., Age, MMSE, eTIV, nWBV).
  + **Structure**: Some variables, like MMSE and CDR, may represent scores or assessments that could have meaningful probabilistic relationships with dementia status.
  + **Potential Model**: A **Discrete Bayesian Network** could work well if we discretize continuous variables, but if preserving the continuous nature of some features (e.g., Age, eTIV) is essential, then a **Gaussian Bayesian Network** would be more suitable.
* **Parkinson’s Dataset**:
  + **Feature Types**: This dataset predominantly consists of continuous variables (e.g., various frequency, jitter, and shimmer measures). The target variable, status, is binary (1 for Parkinson's, 0 for control).
  + **Distribution**: Given the continuous nature and likely Gaussian-like distributions of features, a **Gaussian Bayesian Network** or **Gaussian Processes** could be ideal. If there’s any time-series data, Gaussian Processes would also allow for temporal modeling, which could capture patterns over time if applicable.

**Recommendation**

For the **Dementia Dataset**, a **Gaussian Bayesian Network** is likely best if you wish to preserve the continuous nature of certain features and model probabilistic dependencies between them without discretizing.

For the **Parkinson's Dataset**, **Gaussian Processes** would be effective for handling continuous features, particularly if there’s a need to predict or understand changes over time. However, if time-series is not a factor, a **Gaussian Bayesian Network** would be an excellent fit due to the continuous nature of the features.

### 4. Data Visualization

**\*this can be done before step 3 \***

 **Univariate Analysis**: Visualize each variable independently.

* Use **histograms** for distributions of continuous variables (e.g., Age, MMSE).
* Use **bar charts** for categorical variables (e.g., Group status).

 **Bivariate Analysis**: Explore relationships between pairs of variables.

* **Scatter plots** to examine relationships between continuous variables (e.g., Age vs MMSE).
* **Box plots** or **violin plots** to compare distributions across groups (e.g., MMSE scores across Group categories).
* **Heatmaps** for correlations among continuous variables, which can reveal dependencies or multicollinearity.

 **Multivariate Analysis**: Look at interactions between three or more variables.

* **Pair plots** for several continuous variables to visualize pairwise relationships.
* **Facet grids** to create subsets by a category and visualize how another variable behaves within those subsets.

 **Document Findings**: Note down any patterns, trends, or anomalies that appear in the visuals for inclusion in your report.

### 5. Write Code for the Solution

* You can use libraries like **pgmpy** or **bnlearn** (for Python) for Bayesian Networks or **scikit-learn** for Gaussian Processes.
* For instance, **pgmpy** allows you to set up a Bayesian Network, define conditional probability tables, and run inference based on the queries given.

### Example Output for Queries

For **Dementia**:

* Once the model is trained, you can use it to calculate conditional probabilities like P(Group=nondemented | provided values)P(\text{Group=nondemented | provided values})P(Group=nondemented | provided values), showing the likelihood based on input parameters.

### 6. ****Training and Testing Using Cross-Validation****

* If you choose a different number of splits, provide a justification (e.g., computational resources, dataset size).

### 7. ****Model Performance Metrics (Evaluating)****

* **Disease Classification Accuracy**: This measures the percentage of correctly classified cases (e.g., demented vs. nondemented).
* **Execution Time**: Measure training and inference times for each model to compare efficiency.

**These might not be used**:

* **AUC (Area Under the Curve)**: An important metric to assess the model’s performance across different thresholds in binary classification.
* **Statistical Distances**: Metrics like the **Kullback-Leibler Divergence** (measuring similarity between distributions) and the **Brier Score** (assessing probabilistic predictions) will add depth to your evaluation.

### 8. ****Comparative Analysis and Reporting (report writing, IEEE)****

* **Comparisons**: You could compare:
  + Discrete vs. continuous methods on the same dataset.
  + Different structure learning algorithms for Bayesian Networks.
  + Different implementations of the same algorithm (e.g., your own code vs. a library).
* **Justification**: For each choice (algorithms, metrics, discretization strategy), provide a clear justification in your report. For instance, if Gaussian Bayesian Networks perform better with continuous variables, explain the impact on the results.

### 9. ****Presentation in the Report (3 minutes demo)****

* **Introduction**: Summarize the purpose, dataset choice, and overview of methods.
* **Methodology**: Describe your data preprocessing steps, model selection, and training process.
* **Results**: Present performance metrics (accuracy, AUC, Kullback-Leibler, etc.), and compare across models. Use tables or charts to summarize findings.
* **Discussion and Conclusion**: Interpret your results, discuss the strengths and weaknesses of each approach, and suggest potential improvements or further research.

### 10. Steps Summary

Here’s a concise step-by-step outline for the assignment:

1. **Select Dataset(s) and Justify Choice**:
   * Choose Dementia (MRI data) or Parkinson (voice data), or both.
   * Justify selection based on data type, feature complexity, and compatibility with probabilistic methods.
2. **Preprocess Data**:
   * **Discretize** continuous variables if using discrete Bayesian Networks.
   * **Handle Missing Data** with imputation methods if needed.
3. **Implement Probabilistic Models**:
   * **Choose Methods**: Discrete Bayesian Network, Gaussian Bayesian Network, or Gaussian Processes.
   * **Structure Learning**: Use algorithms like Hill-Climbing or K2 for Bayesian Networks.
4. **Train and Evaluate Models with Cross-Validation**:
   * Use **5-Fold Cross-Validation** for training/testing to ensure robust performance estimates.
   * Adjust fold number if needed and justify any changes.
5. **Compute and Compare Performance Metrics**:
   * Key metrics: Classification Accuracy, AUC, Kullback-Leibler Divergence, Brier Score, Training/Inference Time.
6. **Conduct Comparative Analysis**:
   * Compare models on chosen metrics, focusing on discrete vs. continuous handling, structure learning algorithms, or implementations.
7. **Report Findings**:
   * **Introduction**: Describe objectives, dataset choice, and selected methods.
   * **Methodology**: Summarize data preprocessing, models, and training.
   * **Results**: Present metrics (tables/charts) and analyze differences across models.
   * **Discussion and Conclusion**: Interpret results, highlight model strengths and weaknesses, and suggest improvements.