- An individual’s annual income results from various factors. Intuitively, it is influenced by the individual’s education level, age, gender, occupation, and etc.
- This is a widely cited KNN dataset. I encountered it during my course, and I wish to share it here because it is a good starter example for data pre-processing and machine learning practices.
- **Fields**
  The dataset contains 16 columns
  -- The income is divide into two classes: <=50K and >50K
  Number of attributes: 14
  -- These are the demographics and other features to describe a person
- We can explore the possibility of predicting income level based on the individual’s personal information.
- Can we use a K-Nearest Neighbors (KNN) model to predict an individual's income level based on their personal information, such as age, workclass, education, occupation, gender, etc., using the provided dataset?

- How accurately can a K-Nearest Neighbors (KNN) model predict an individual's income level (<=50K or >50K) based on their demographics and personal attributes, using the given dataset?

- What is the optimal value of K (number of neighbors) to achieve the highest accuracy in predicting income levels using the K-Nearest Neighbors (KNN) algorithm with the provided dataset?

- Can we improve the accuracy of income level predictions by performing feature selection or dimensionality reduction techniques on the dataset before applying the K-Nearest Neighbors (KNN) algorithm? If so, which features or combination of features contribute the most to the accuracy improvement?

- Is there a significant difference in the performance of the K-Nearest Neighbors (KNN) algorithm when predicting income levels between different subgroups based on gender or race using the provided dataset?

- Can we improve the accuracy of the K-Nearest Neighbors (KNN) model for predicting income levels by applying data preprocessing techniques, such as handling missing values, outlier detection, or feature scaling, to the given dataset? If so, which preprocessing techniques have the most impact on improving the model's performance?

- How does the performance of the K-Nearest Neighbors (KNN) model for predicting income levels vary when using different distance metrics, such as Euclidean distance, Manhattan distance, or cosine similarity, on the given dataset? Which distance metric yields the best performance?

- Can we achieve better predictive performance for income level classification by applying ensemble techniques, such as bagging or boosting, to a combination of K-Nearest Neighbors (KNN) models trained on different subsets of the given dataset? How does the ensemble approach compare to using a single KNN model?