Mini-Project1

Know Your Data, Naive Bayesian Classifier, and k-fold cross validation

By

Karuna Nayak

CSC -869

SFSU

Submitted to

Dr. Hui Yang

CS dept

SFSU

1. Compile and run instructions

The project files contain two Jupyter Notebooks

- NB_EqualBinWidth.ipynb : Implements the Naïve Bayesian with equal binning width method.
- NB_withGaussian.ipynb : Implements the Naïve Bayesian with Gaussian Distribution assumption for continuous attributes.

Steps:

1. Download both jupyter notebook files and *adult.data* file and save these in the same folder.
2. Open the notebook files and go to **Cell** and click on **Run All**
3. Upon completing, last two cells in both files will display avg accuracy, F1-score and Matthews correlation coefficient.

- In NB_EqualBinWidth.ipynb , Naive_Bayesian_model(data, k, remove_missing_values) function takes parameters as Census data set, value of k for K-fold cross validation and Boolean True or False for remove_missing_values.
  →If value passed to remove_missing_values is True, algorithm removes the records with missing values and call the functions to train and predict the dataset and returns the avg accuracy, avg F1-score and avg Matthews correlation coefficient.
  →If value passed to remove_missing_values is false, algorithm replaces the missing values with mode of the attributes and perform the train, predict and returns the avg accuracy, avg F1-score and avg Matthews correlation coefficient.
- In NB_withGaussian.ipynb, same procedure is taken care with Naive_Bayesian_model_Gauusian(data, k, remove_missing_values) function.

2. Data Analysis

The goal of the project to predict whether income exceeds $50K/yr based on census data.
The census dataset has 14 attributes, in which 6 attributes are continuous and 8 are discrete.
Upon looking at the data, missing values were found in three attributes, which are discrete.

number of records of missing values in attributes

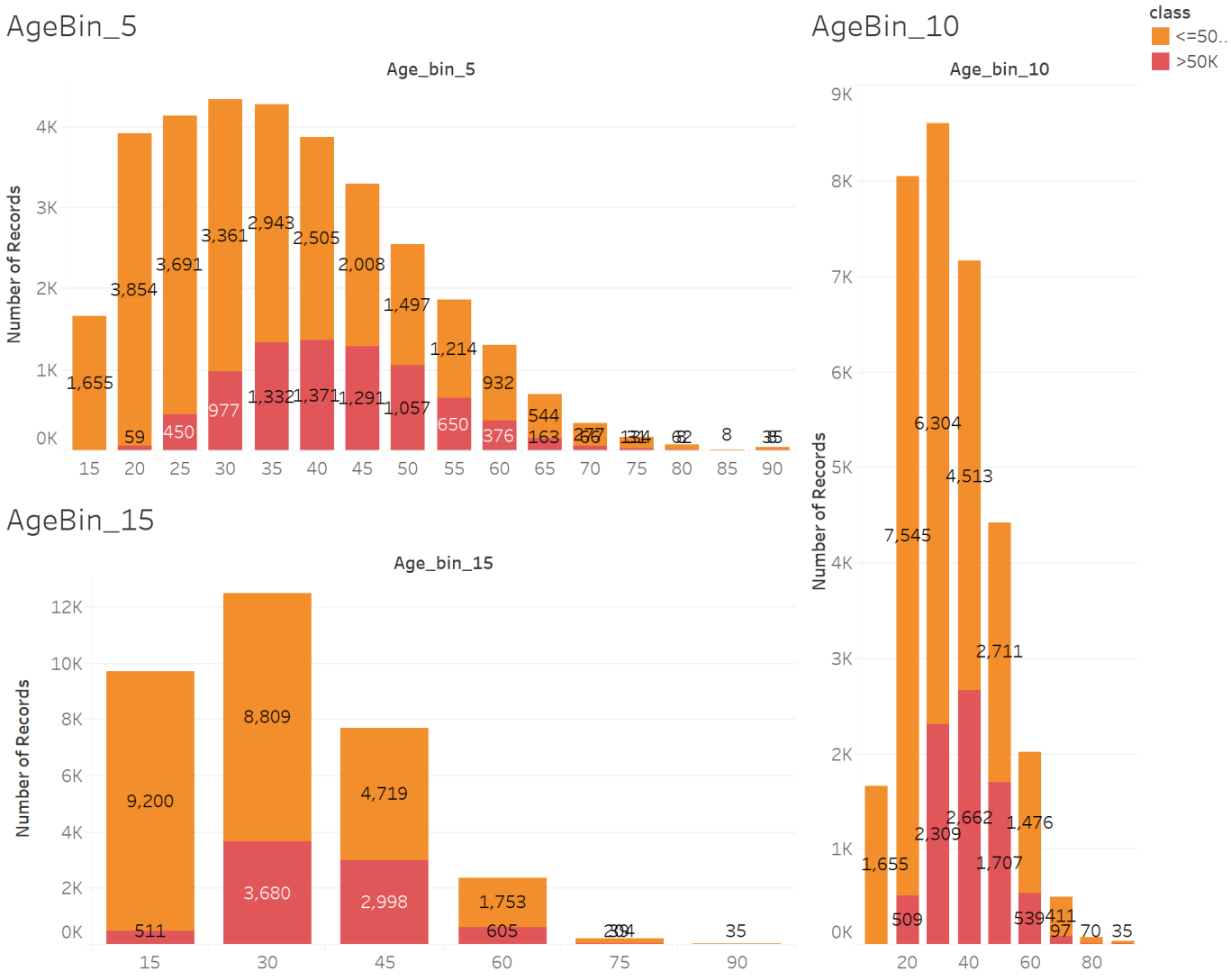
age	0
workclass	1836
fnlwgt	0
education	0
education-num	0
marital-status	0
occupation	1843
relationship	0
race	0
sex	0
capital-gain	0
capital-loss	0
hours-per-week	0
native-country	583
class	0

Using Tableau, data visualization is done for relationship between attributes and class. The 6 attributes - 'age','fnlwgt','education-num','capital-loss','capital-gain' and 'hours-per-week' are the continuous values and need to be discretized for classification. Discretization of continuous attributes are handled in two ways.

1. Discretize using equi-width bins
2. Assuming continuous attributes follows Gaussian distribution.

Equi-width method:

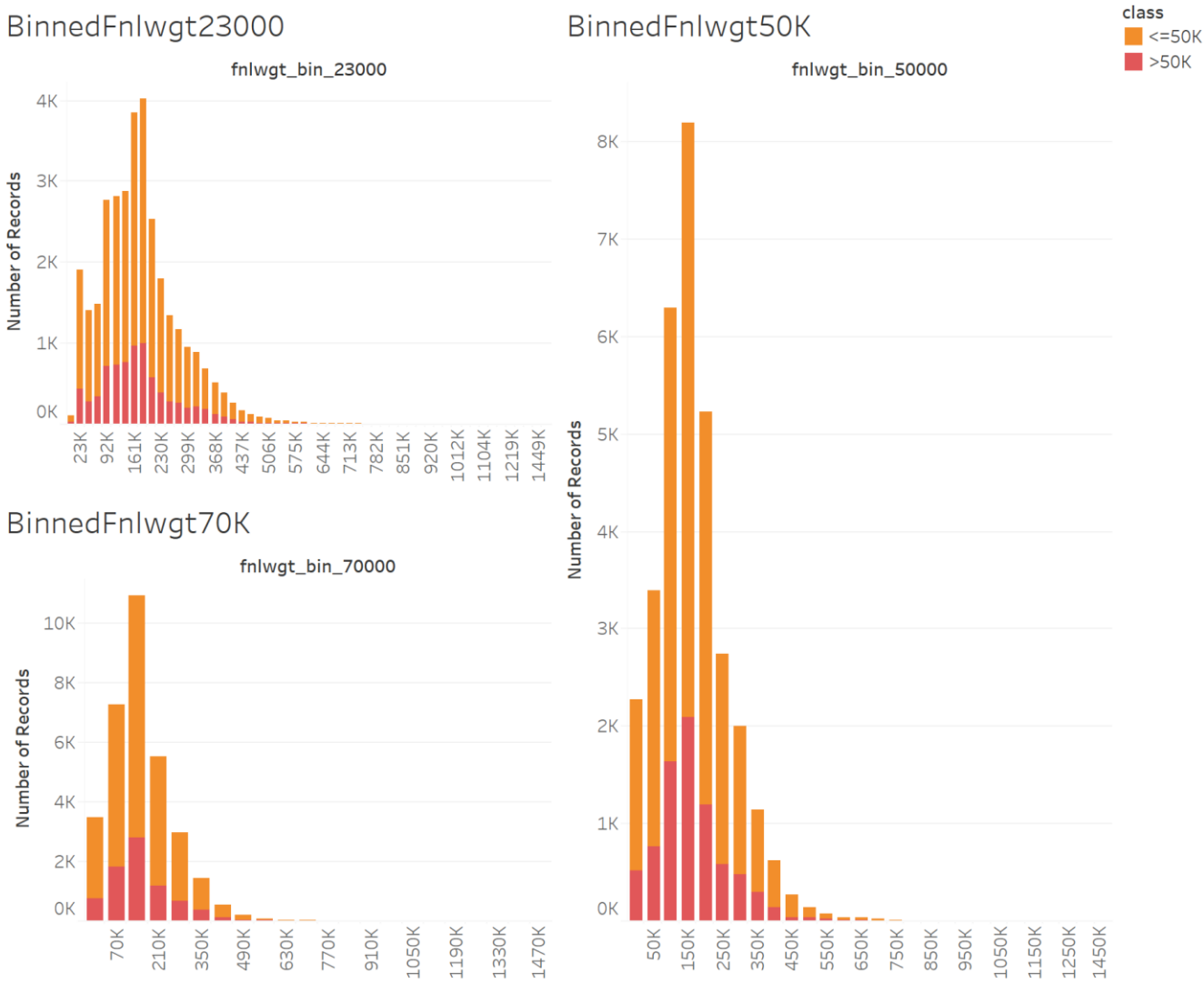
The continuous attributes are converted into discrete using equi-width binning strategy. To determine the suitable binning width, 3 different sizes were tried for each continuous attribute.



2.1

1. Age:

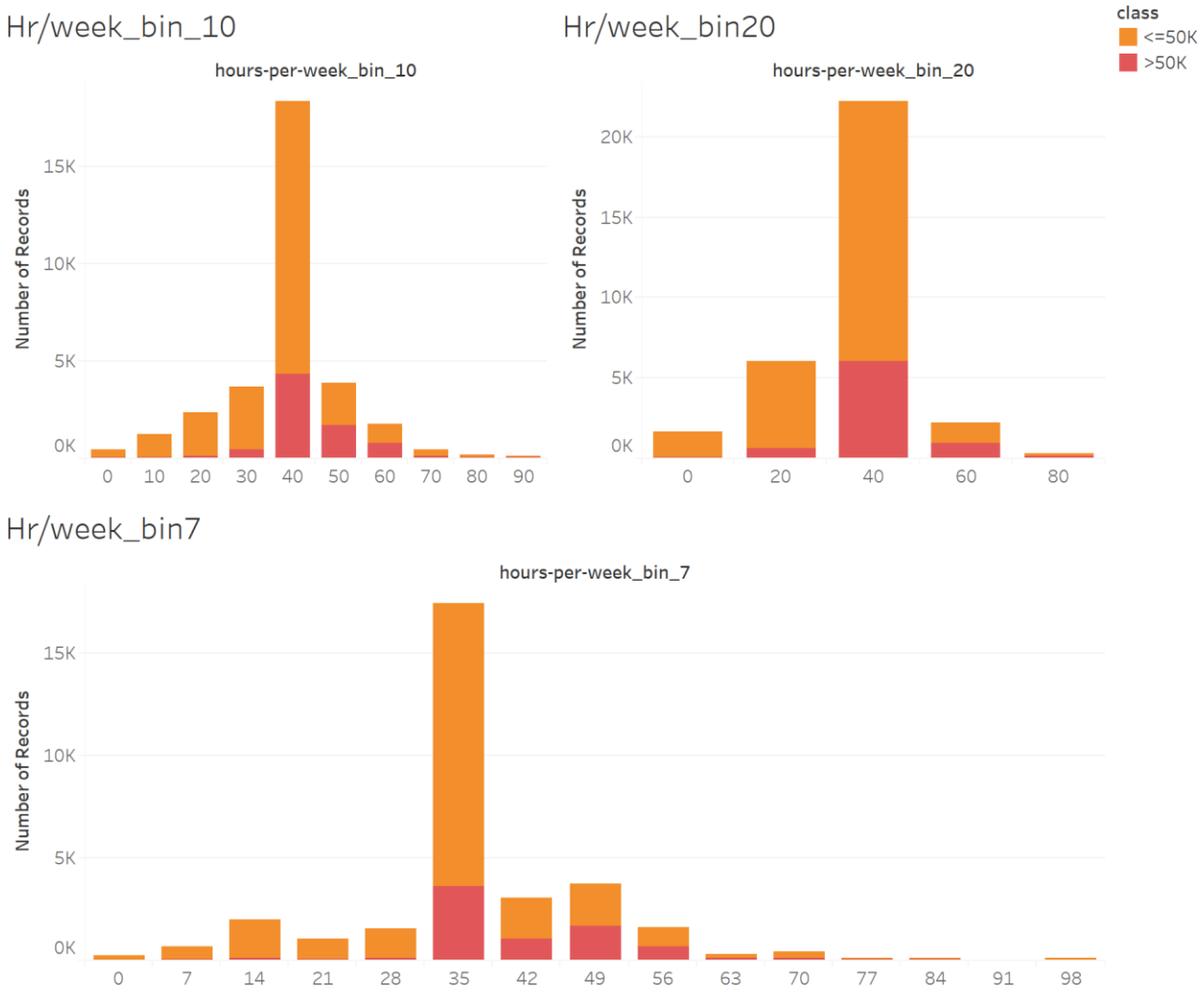
The image 2.1 shows the 3 discretization of Age attribute using different binning widths- 5, 10 and 15. Age attribute has the values ranging from 17 to 90. Looking at the all 3 plots, the values looks little left skewed. Higher the binwidth, more data details were lost. Bin width 5 is chosen for classification task so that not to loose many data details.



2.2

2. Fnlwgt:

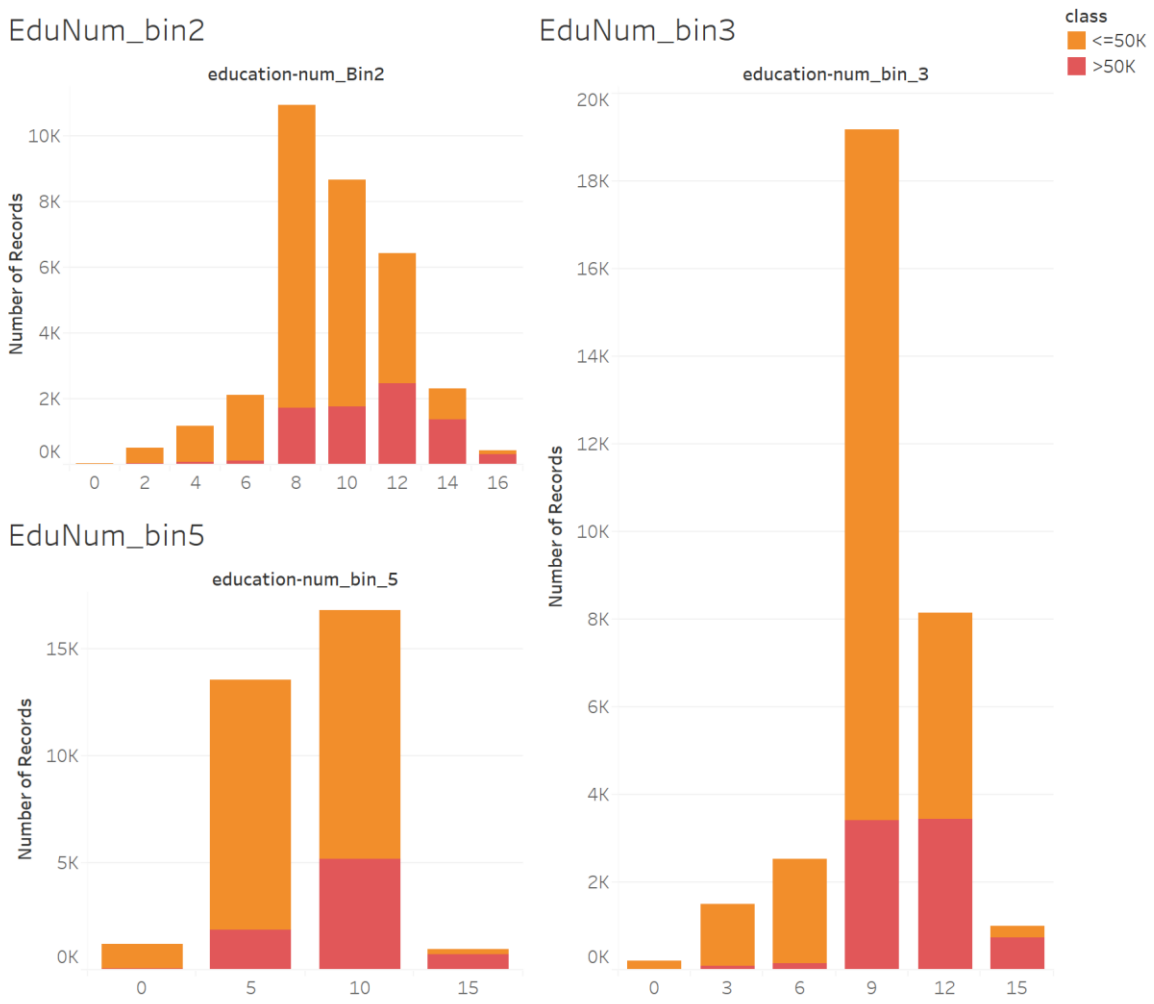
   The above image 2.2 shows the plots using different binning widths. The continuous attribute fnlwgt has values ranging from 12,285 to 1,484,705. Since bin width 50000 and 70000 is losing many details, 23000 is chosen as bin width for fnlwgt.



2.3

3. Hour-per-week

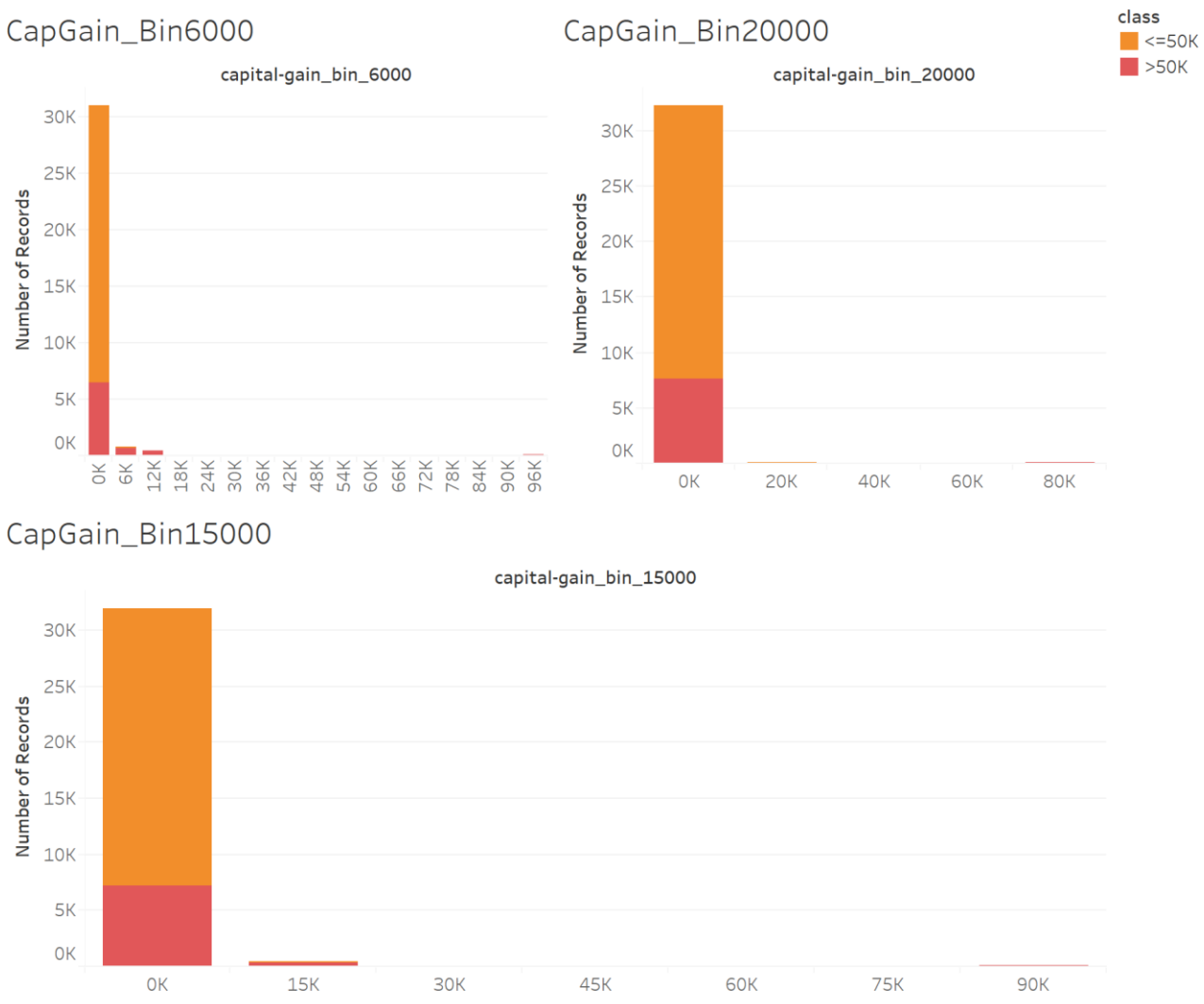
Hour-per-week continuous attributes values ranges from 1 to 99. Looking at the dashboard snapshot, binwidth 20 looses many data details, where as few bins in 7-binwidth plot has very less to zero number of records. The bin width 10 plot looks almost similar to normal distribution and hence bin width for hour-per-week is chosen as 10.



2.4

4. Education-num

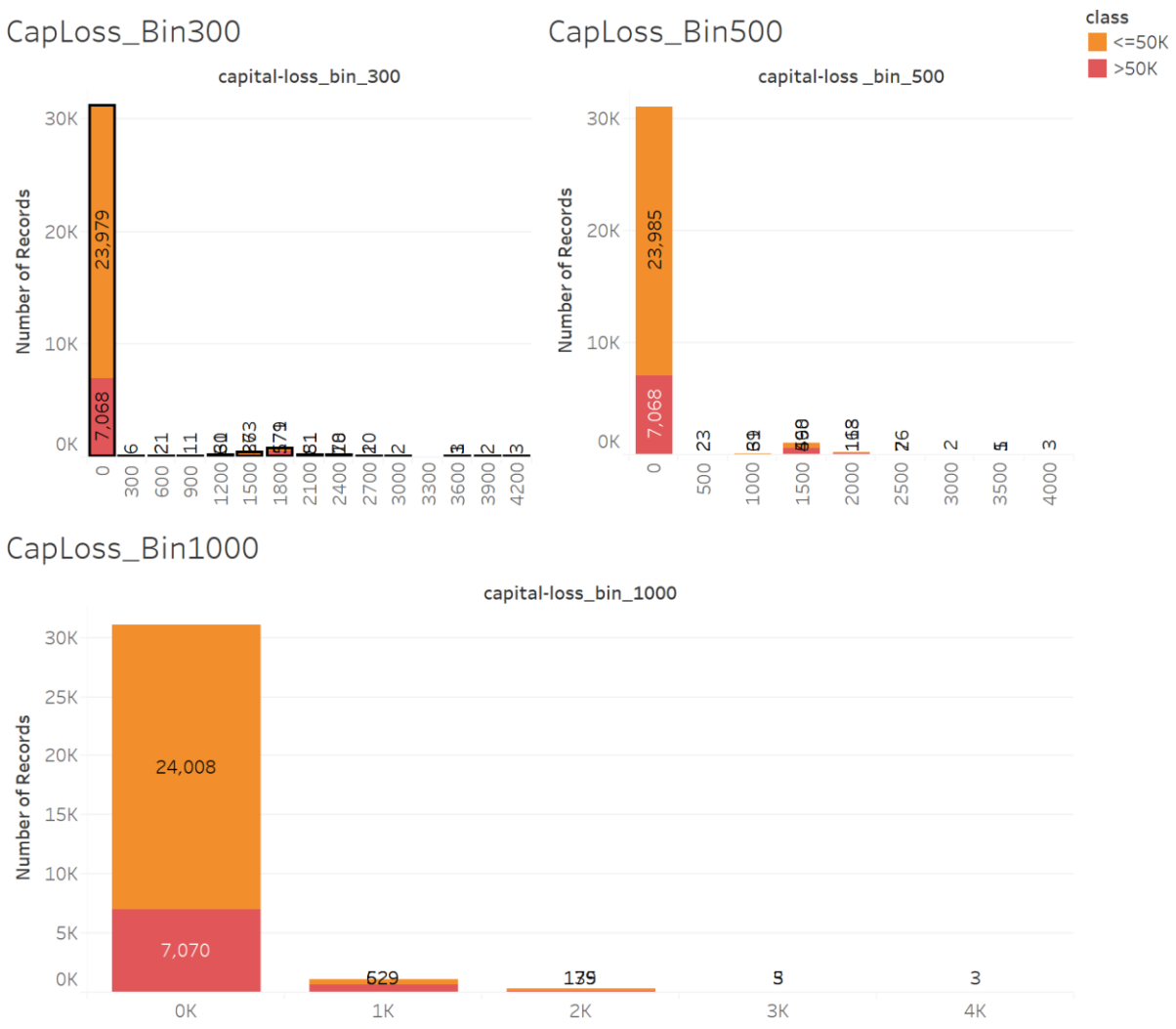
Education number is the one of the continuous attributes in the dataset. The bin-width size tried were 2, 3 and 5. Looking at the dashboard snapshot 2.4, bin width 2 gives more details in the data and was chosen while doing classification task.



2.5

5. Capital-gain

For continuous attribute Capital-gain, most of the records have value as 0. Looking at the dashboard snapshot, bin width can be taken as 20000, since majority of bins will have zero records for lower bin widths.



2.6

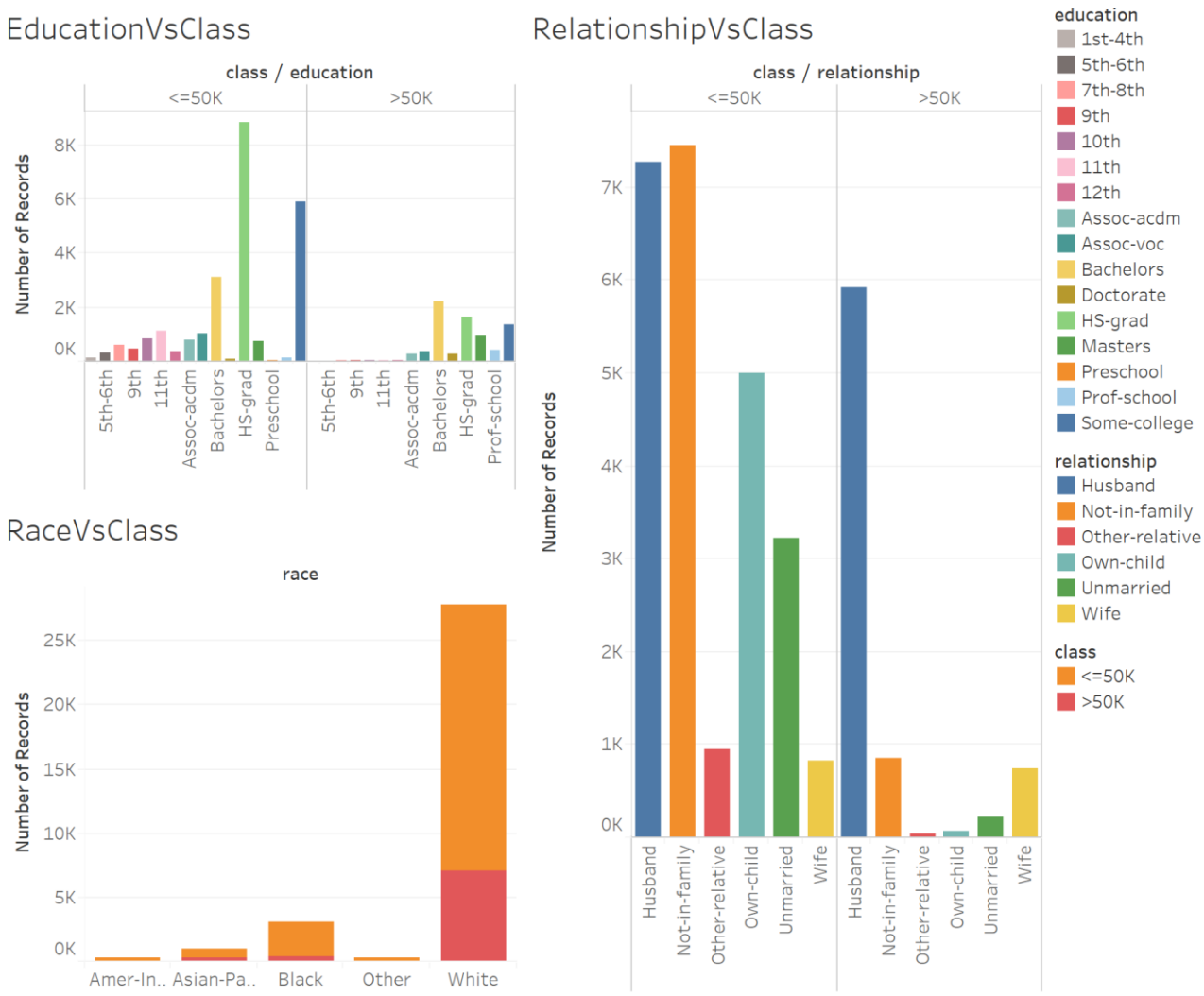
6. Capital Loss

For continuous attribute Capital-loss, most of the records have value as 0. Looking at the dashboard snapshot, bin width can be taken as 500, since majority of bins will have very less number of records for lower bin widths.

Gaussian-distribution method:

The continuous attributes are assumed to be following Gaussian distribution and probability for each value is calculated using Probability Distribution Function formula with mean and standard deviation of the attribute.

Discrete attributes:



For discrete variables, race, education and relationship, class distribution is visualized.

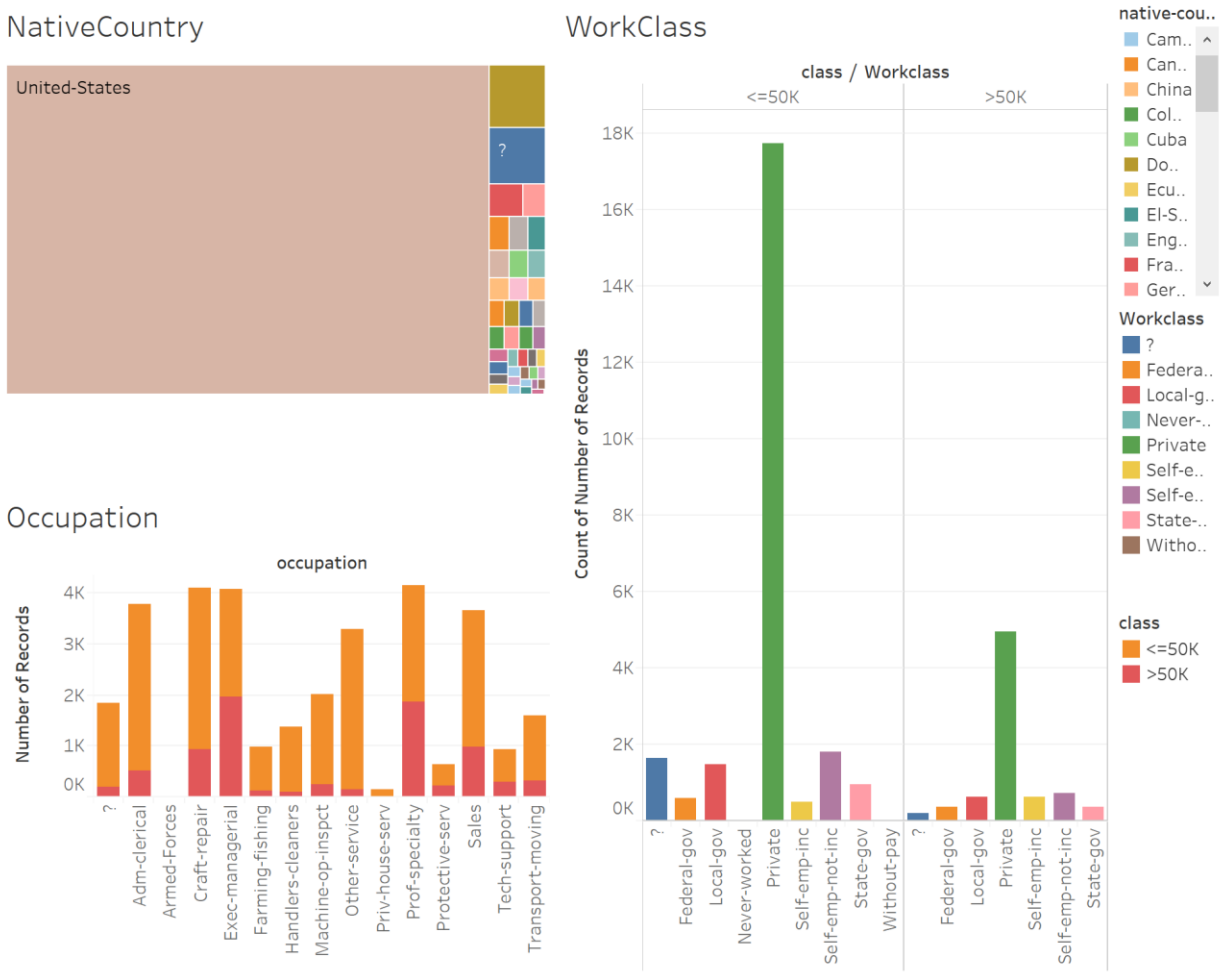
Majority of records are with race White, where as Hs-grad has more records for education. For class, <=50K, not-family and Husbands having more records and for class >50K, Husband is having more number of records.

For sex attribute, male has more records for both >50K and <=50K classes, where as in marital status attribute, never-married has more records in <=50K and married-civ-spouse has similar number of records for both classes.

Handling missing values:

Below Tableau dashboard snapshot shows the relation between class and attributes of missing values. The missing value are represented as "**?**". Clearly, for native-country, occupation and workclass, the number of records of missing values is very small compared to the total number of records in the dataset (0.017 to 0.056 ratio).

In NativeCountry, major set of records are for United-Sates and In workclass, major set of records are for private, which are the mode of these attributes. For occupation, craft-repair, exec-managerial and prof-specialty has almost similar number of records. However, prof-specialty has the highest count, hence mode for occupation was given as prof-specialty.



To handle the missing values, two different approaches are taken.

1. Replace the missing values with mode of the attributes.
   Below table shows the mode for attributes which have missing values.

Attribute	**Mode**
Workclass	Private
Occupation	prof-specialty
NativeCountry	United-States

2. Remove the records with missing values.

Since the number of records of missing values is very small compared to the total number of records in the dataset (0.017 to 0.056 ratio), the missing value records are dropped from dataset in the second approach.

Implementation:

There are two parts of code in the project. The first part contains the implementation using equi-width binning method. Missing values are handled by replacing it with mode of the attribute or removing records as per the parameter passed by the user. The data set in divided into k(10) folds and in each iteration, 1 subset of data is considered as test and other 9 parts are considered as training set. In the training part of the process, binning is performed on continuous variables and training set is sub-divided for both class labels. The training set is iterated to store the sum of records of each value in the dictionary. In the prediction part, the bins are created for continuous attributes of test dataset and for every value on the row, probability is calculated using counts stored in the dictionary. The predicted label is compared with actual label. Accuracy, F1-score and Matthews correlation coefficient are calculated for each k – fold and avg accuracy, avg F1-score and avg MCC is calculated.

For second part, continuous attributes are assumed to be following Gaussian distribution and in training part, mean and standard deviation of continuous attribute is calculated and stored in dictionary. While predicting the labels, for continuous value, probability is calculated using PDF formula. At the end of K-fold, avg accuracy, F1-measure and MCC is calculated.

Results:

Evaluation Strategies:

In K-fold cross validation, for each fold, accuracy, F1-measure and Matthews Correlation Coefficient is calculated and avg accuracy, F1 score and MCC is returned for each model.

The above results are for bin widths:

['age'] = 15 ,['fnlwgt'] = 70000 , ['education-num'] = 5, ['capital-loss'] = 1000 , ['capital-gain'] = 20000 ,['hours-per-week'] = 20

1. Equi-bin width strategy with removed missing values

Accuracy	81 %
F1-score	0.66
Matthews Correlation coefficient	0.54

2. Equi-bin width strategy with missing values replaced with mode of the attributes

Accuracy	81.2%
F1-score	0.66
Matthews Correlation coefficient	0.54

Changing the binwidths of all continuous attributes to lower values increased the all 3 measures Accuracy, F1-score and MCC. The lower bin width values gives the more bins so that the dataset contains more details, and hence the improvement in the measures.

['age'] = 5 , ['fnlwgt'] = 23000, ['education-num'] = 2, ['capital-loss'] = 300 , ['capital-gain'] = 6000 , ['hours-per-week'] = 10

1. Equi-bin width strategy with removed missing values

Accuracy	82.7 %
F1-score	0.69
Matthews Correlation coefficient	0.58

2. Equi-bin width strategy with missing values replaced with mode of the attributes

Accuracy	82.6
F1-score	0.69
Matthews Correlation coefficient	0.58

Below tables lists the measures for Gaussian distribution method:

3. Gaussian distribution assumption with removed missing values

Accuracy	82.98%
F1-score	0.60
Matthews Correlation coefficient	0.51

4. Gaussian distribution assumption with missing values replaced with mode of the attribute

Accuracy	82.8%
F1-score	0.59
Matthews Correlation coefficient	0.50

Improving results:

Data set is highly imbalanced. If the data set can be under-sampled or over-sampled to balance the classes, measures could be improved more. One more strategy could be used is while doing K fold, divide the data set such a way that each sub set contains same ratio of both classes.