Data Mining Final Project

CISC 4631: Data Mining Fall 2022

By: Claude Batrouni, Romel Hakim,
William Shrewsbury, Kevin Yagar

Professor: Yijun Zhao

Introduction

Using census-income.data.csv, a dataset containing 32,561 entries obtained from the UCI machine learning lab, we aim to predict whether or not an individual's income will be greater than $50,000 per year based on key attributes from the census data.

The dataset used was from adult.csv, which is the exact same dataset as census-income.data.csv, however with better formatting. Adult.csv is a dataset from the UCI repository, and is a dataset that we have permission to use as per Professor Zhao. Naming has only changed to census-income.data for naming consistency with the original dataset provided to us in the original instructions.

Below are the listed features that were used in examining the data distribution, as well as executing our algorithms.

age: continuous.
Age of the subject
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Classification of the type of occupation
fnlwgt: continuous.
Weight assigned by the US census bureau for each subject
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
Highest education level completed by the subject
educational-num: 1-15
Conversion of educational level into integer values
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
Marital status of the subject (NOTE: AF stands for Armed Forces)
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical,

Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

Occupation of the subject

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

Relationship status of the subject

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

Race of the subject

sex: Female, Male.

Sex of the subject

capital-gain: continuous.

Profit made by the subject by selling an investment

capital-loss: continuous.

Loss made by the subject by selling an investment

hours-per-week: continuous.

Amount of work hours per week done by the subject

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

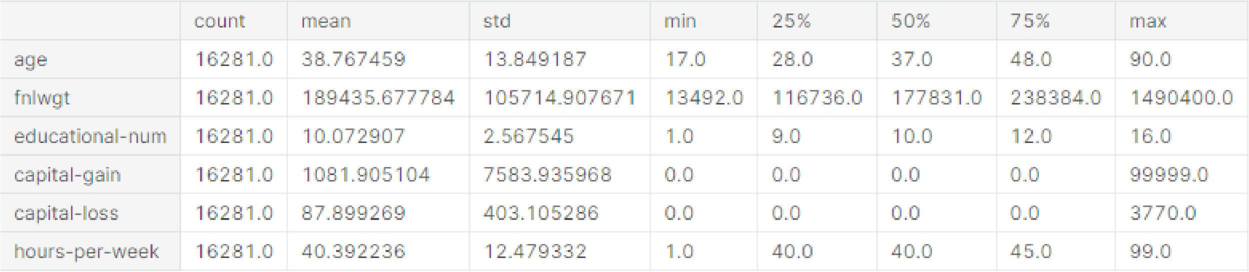
Native country of the subject

income: >50k or <=50k.

Whether or not the subject’s income is above or below 50k

Summary of Initial Data

Initially, our dataset can be summarized as follows for our continous attributes:



	count	mean	std	min	25%	50%	75%	max
age	16281.0	38.767459	13.849187	17.0	28.0	37.0	48.0	90.0
fnlwgt	16281.0	189435.677784	105714.907671	13492.0	116736.0	177831.0	238384.0	1490400.0
educational-num	16281.0	10.072907	2.567545	1.0	9.0	10.0	12.0	16.0
capital-gain	16281.0	1081.905104	7583.935968	0.0	0.0	0.0	0.0	99999.0
capital-loss	16281.0	87.899269	403.105286	0.0	0.0	0.0	0.0	3770.0
hours-per-week	16281.0	40.392236	12.479332	1.0	40.0	40.0	45.0	99.0

Figure 1

For our categorical data, it can be summarized within the following figures:

The Number of Countries Listed in the Dataset

United States	14662	Guatemala	24
Mexico	308	Greece	20
?	274	Vietnam	19
Philippines	97	Ecuador	17
Puerto-Rico	70	Iran	16
Germany	69	Peru	15
Canada	61	Nicaragua	15
India	51	Taiwan	14
El-Salvador	49	Ireland	13
China	47	Thailand	12
Cuba	43	Hong Kong	10
England	37	Outlying-US	9
South	35	France	9
Dominican Republic	33	Scotland	9

Italy	32	Cambodia	9
Haiti	31	Trinidad & Tobago	8
Portugal	30	Yugoslavia	7
Japan	30	Honduras	7
Poland	27	Hungary	6
Columbia	26	Laos	5
Jamaica	25		

Figure 2

Private	11210
Self-emp-not-inc	1321
Local-gov	1043
?	963
State-gov	683
Self-emp-inc	579
Federal-gov	472
Without-pay	7
Never-worked	3

Figure 3

Prof-specialty	2032
Exec-managerial	2020
Craft-repair	2013
Sales	1854
Adm-clerical	1841
Other-service	1628
Machine-op-inspct	1020
?	966
Transport-moving	758
Handlers-cleaners	702
Tech-support	518
Farming-fishing	496
Protective-serv	334
Priv-house-serv	93

Armed-Forces	6

Figure 4

Married-civ-spouse	7403
Never-married	5434
Divorced	2190
Widowed	525
Separated	505
Married-spouse-absent	210
Married-AF-spouse	14

Figure 5

Male	10860
Female	5421

Figure 6

White	13946
Black	1561
Asian-Pac-Islander	480
Amer-Indian-Eskimo	159
Other	135

Figure 7

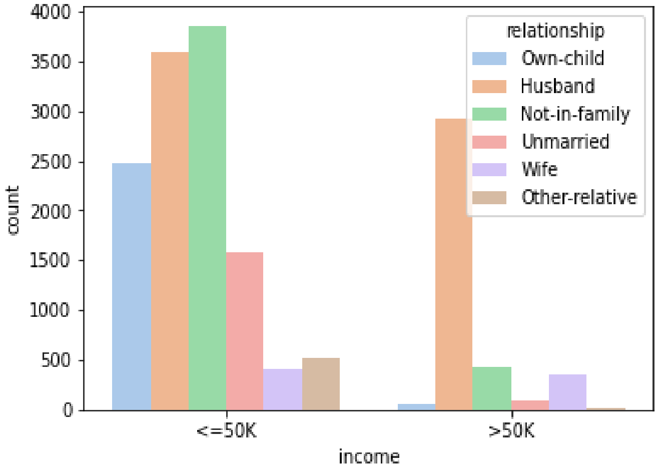


Figure 8

Number of people making either above or below $50k by relationship status

Looking at this graph, one key trend we can observe is that there are far more husbands represented in the dataset compared to wives in both categorical topics. One could make the argument that this is due to social normalities such as it is more common to find wives as a stay at home parent compared to husbands. A similar argument could be made that husbands make in general more money than wives. However, that argument is not conclusive in our data because even though husbands are the second highest counted in the above 50k attribute, husbands are actually significantly represented in the below 50k attribute. This likely ties in to the fact that male subjects are far more prevalent in our dataset as confirmed by figure 6, meaning the listed common arguments to be made from the dataset can not be proven or disproven.

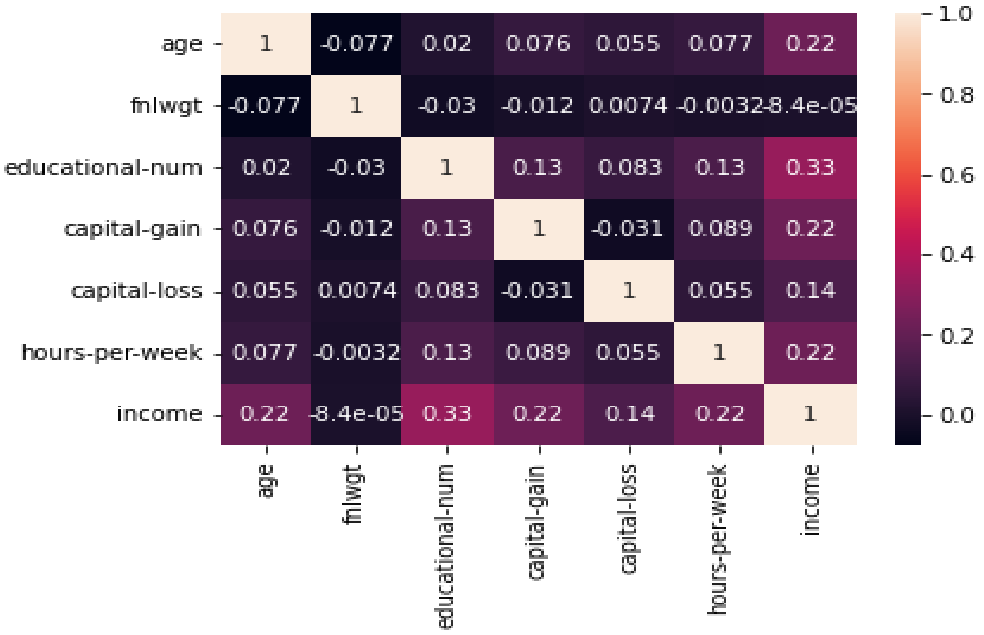


Figure 9

Heatmap displaying correlation between features

As seen in figure 9, there is no strong correlation between the features listed. The one correlation that shows through is the common thought that the level of education (*educational-num)* completed affects income. But even so, the correlation is not strong standing at a value of .33.

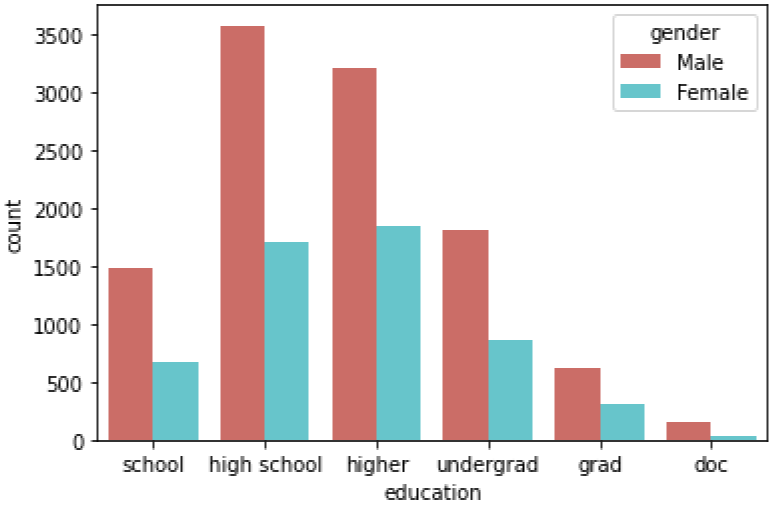


Figure 10

Number of males and females in each education level

As seen in this graph, the data states that there is a generally higher count of educated Males in every spectrum. We cannot make arguments based off figure 10 for the same reason why we could not make arguments based off figure 8. There are double the amount of males compared to females which skewers the data to represent males (figure 2).

Percentage of individuals in each education level

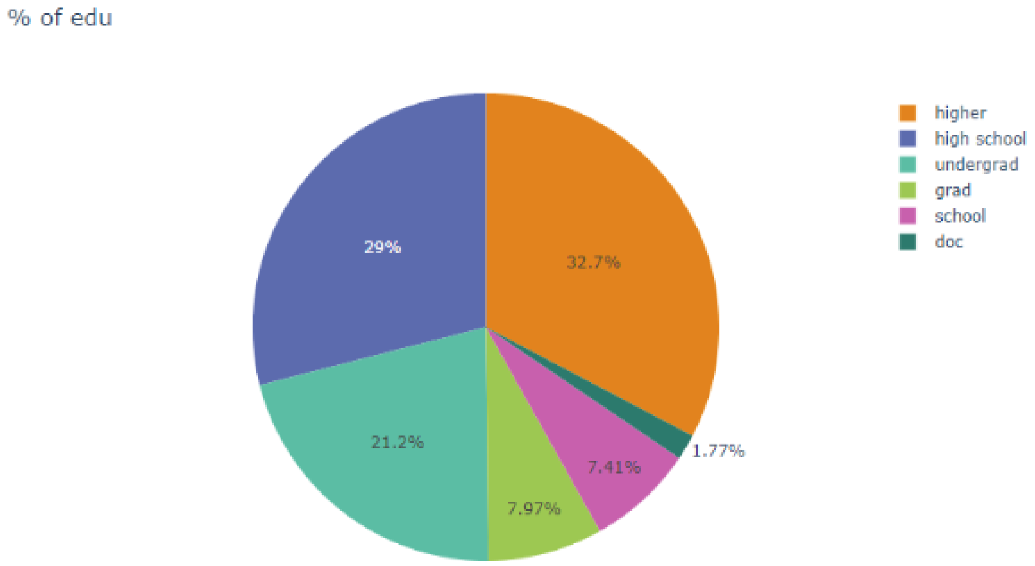


Figure 11

Missing Data

Taking a look at our census data prior to applying any sort of algorithmic work, we found that our dataset had missing values. All three of the missing values belonged to the following categorical values: *native_country, workclass* and *occupation*. Initially, we thought to simply remove the entire row as we did not want unknown variables to affect our predictions. However, we decided against this method as there would simply be far too many rows of data that would have to be removed. Each of these rows has their own *fnlwgt* value which in total represent a significant amount of data taken away from our data set if we chose to remove them. Instead, what we did was replace the unknowns values with the mode for that attribute.

Take for example figure 2, a chart listing all the categorical variables listed for *native-country.* The unknown data set occupied just 0.01% of the total data for that attribute. However, due to the *fnlwgt* of the various subjects that have an unknown value for *native-country,* those subjects being removed could have a far greater impact on the accuracy than we could account for. Instead, we replaced the unknown value with the mode for the *native-country* which was the United States.

Analysis of Initial Data

Mapping out our initial data allows us to see the data fully visualized before any algorithms are applied. Analyzing this data distribution gives us a general understanding of the dataset. When applying the algorithms, we can find correlations using Naive Bayes and our ensemble.

Boxplot displaying the data distribution

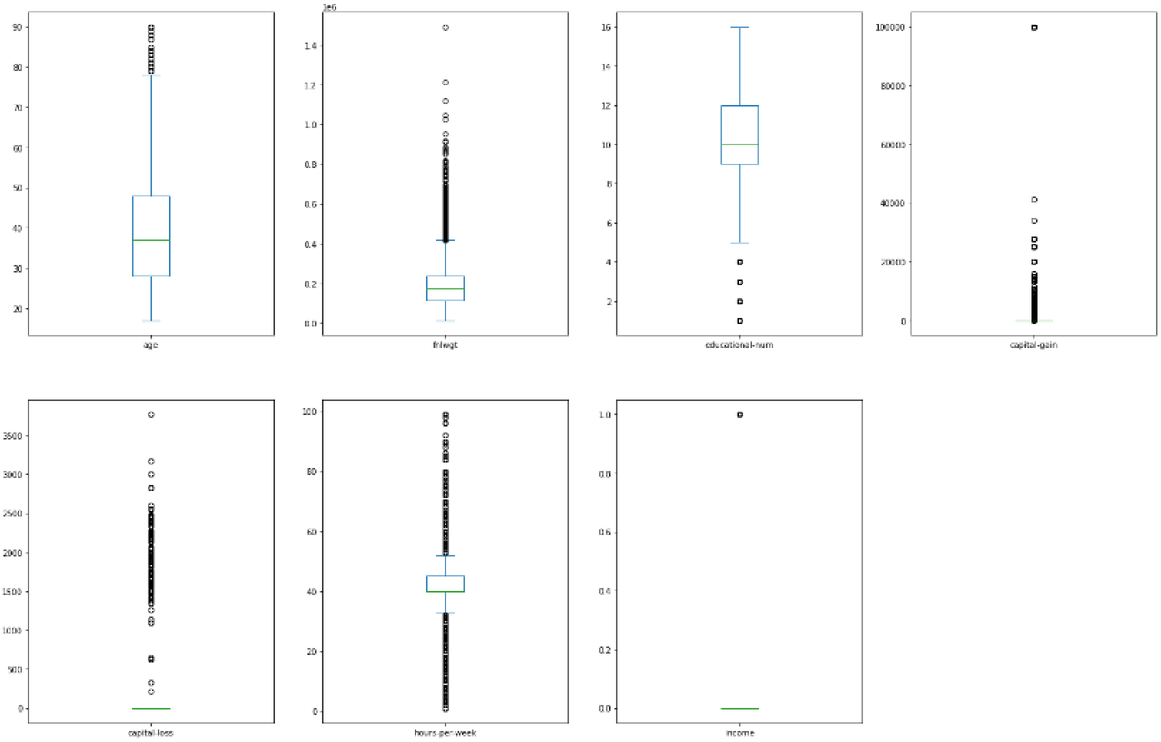


Figure 12

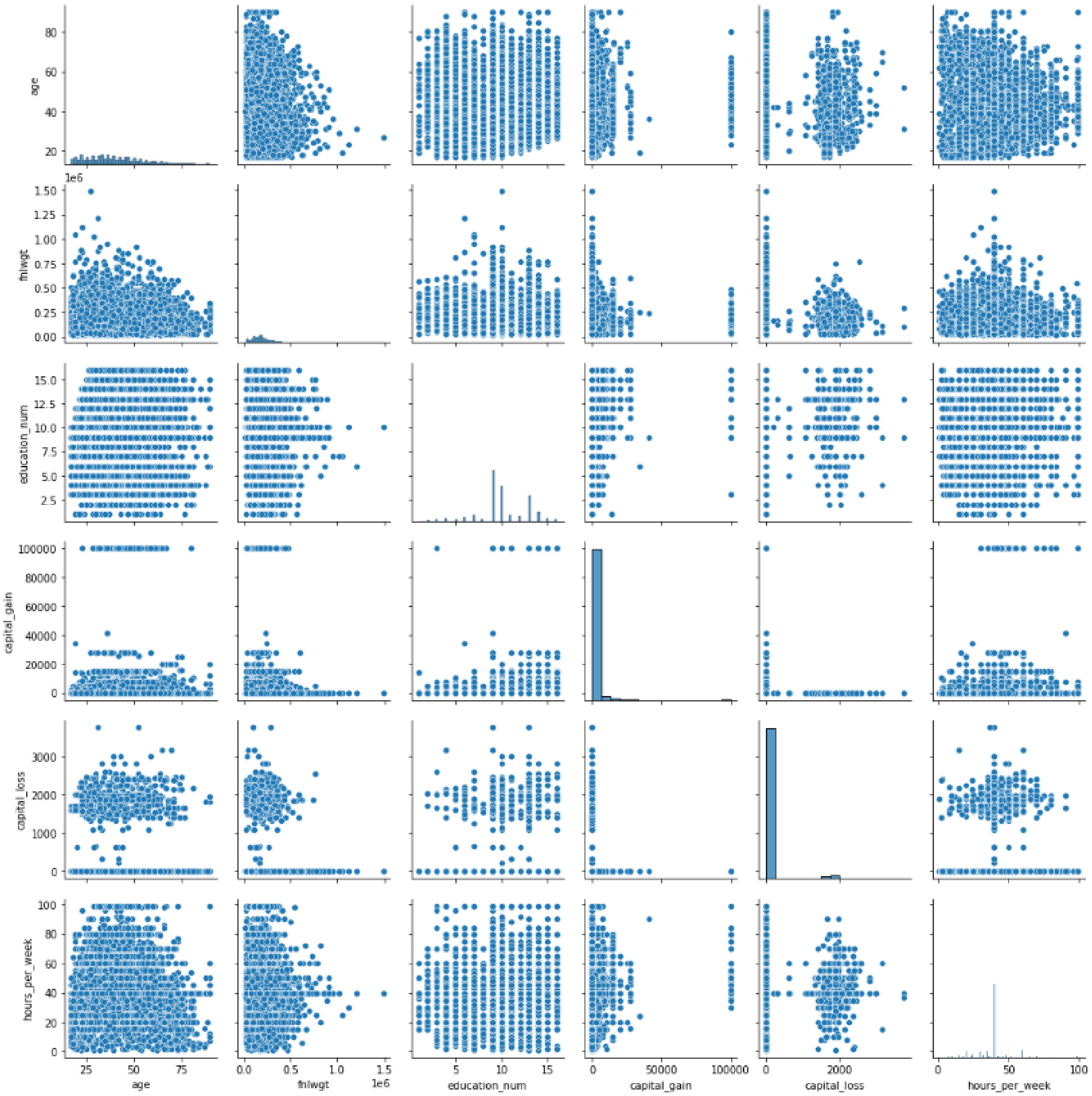


Figure 13

Total distribution of the dataset.

Amongst these graphs we can notice several trends in our data. For instance, when education level increases, capital gain also increases. There is also a sharp decline in the hours_per_week as age increases.

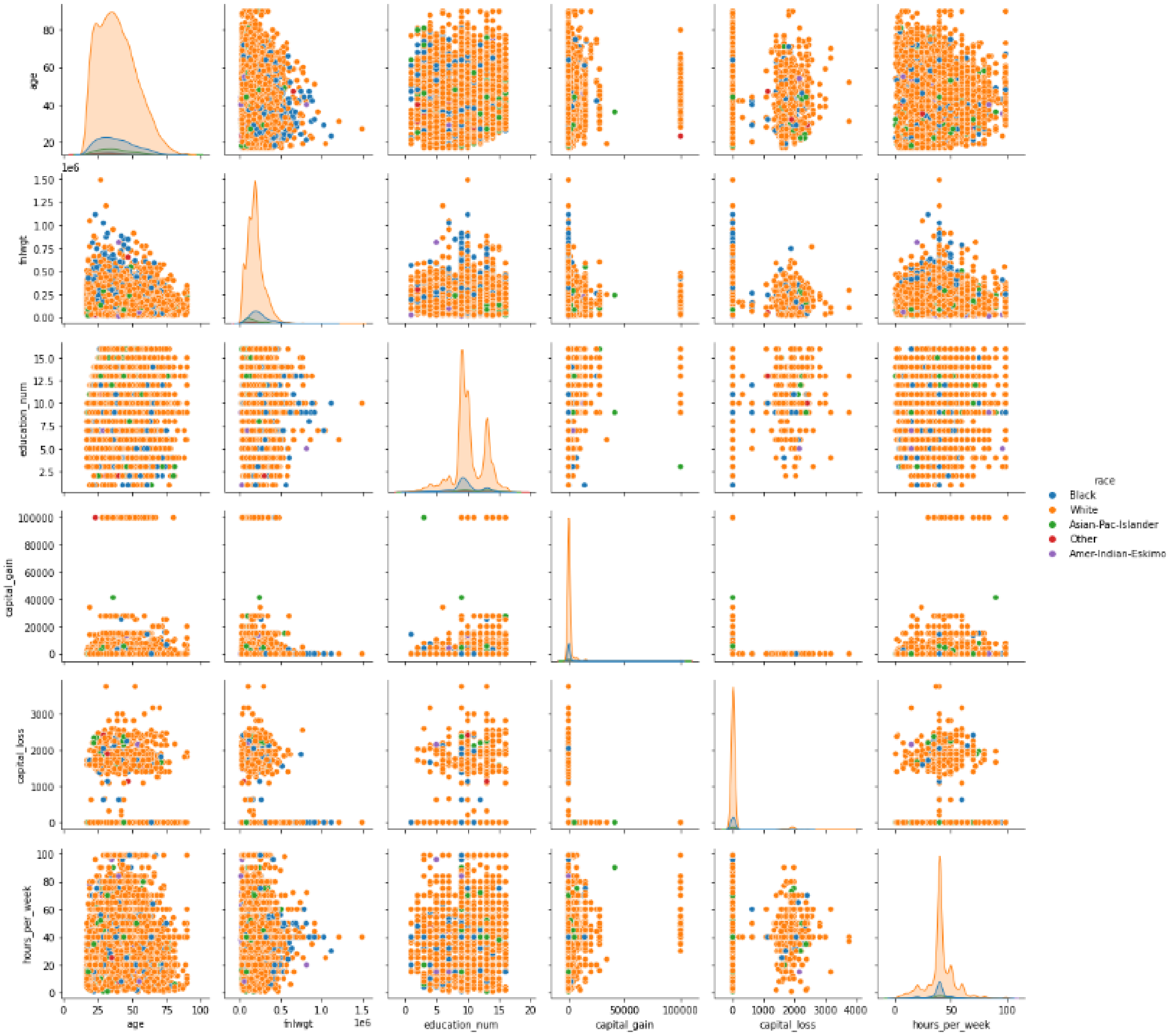


Figure 14

Total distribution of the dataset given race

The total distribution of the dataset given race depicts the disproportionately large sample size of race:White in our data. Qualitatively, race:White make up 85.7% of the data.

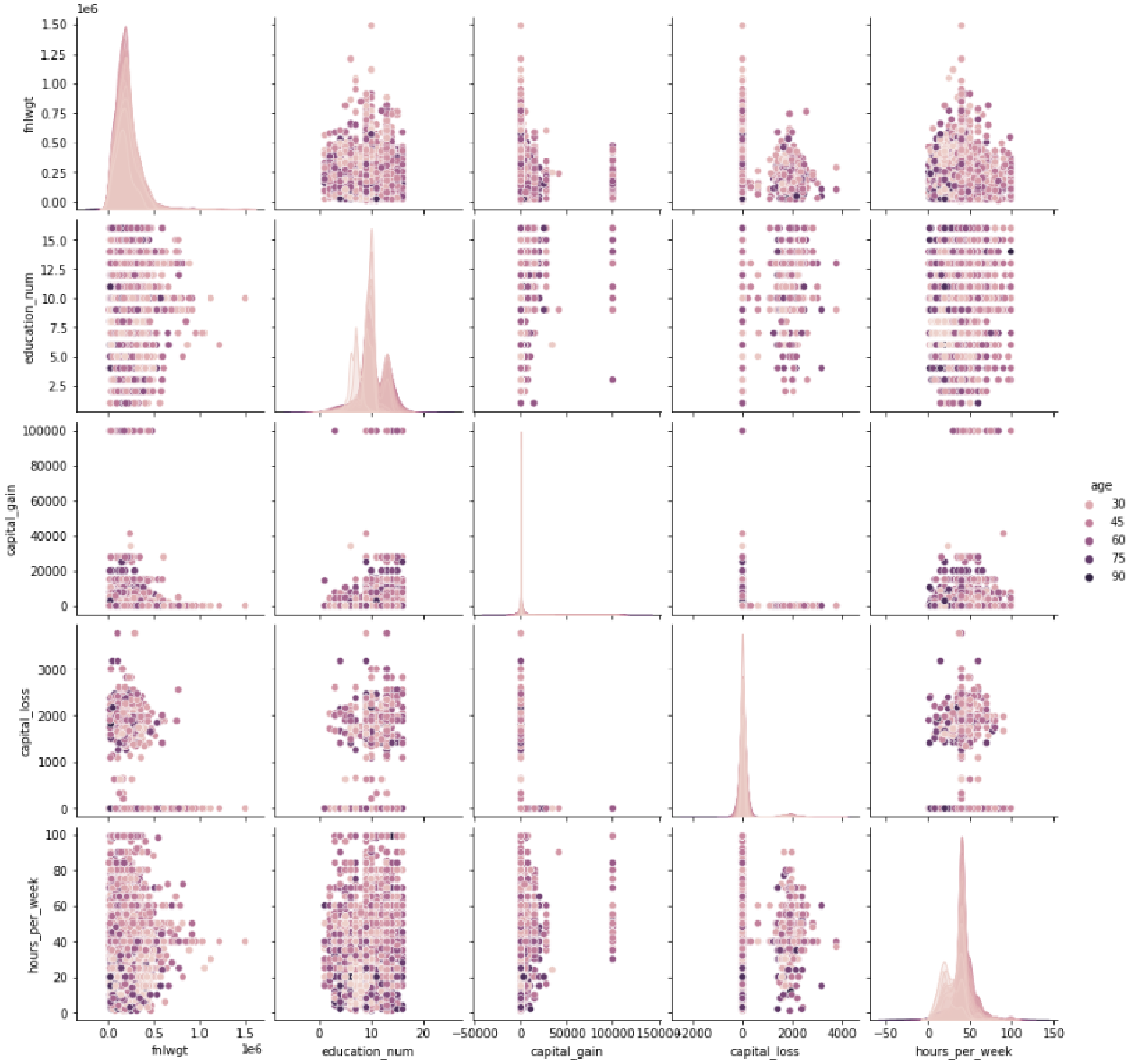


Figure 15

Total distribution of the dataset given age

The total distribution of the dataset given age reveals that age has a more equal distribution compared to factors such as race. We can see that individuals in an older age bracket work less hours a week regardless of education level.

Classifiers

Our group has decided to create an ensemble model consisting of the following classifying algorithms: logistic regression, naive bayes, and random forest. We have decided to take this route because by utilizing the ensemble method and combining multiple classifiers, rather than using a single algorithm, we hope to diminish overfitting. We ultimately hope that we end up with a more precise model than if we had only used a single classifying algorithm. Our ensemble model will be located in a Jupyter file.

We have also decided to test our ensemble model against a singular classification algorithm. We decided to use an algorithm already within our ensemble learning, so we choose to use the Naive Bayes classification algorithm. Our implementation of Naive Bayes will be in a separate Jupyter file than our ensemble model.

An assumption we had beforehand is that our ensemble model will yield better results and be a much better predictor than the singular classification algorithm we implemented, Naive Bayes. Moreover, taking into account time complexity, we can see there is a stark difference. Naive Bayes complexity is O(d*c) whereas an ensemble of Naive Bayes, Random Forest, and Logisitical Regression has a greater time complexity. The ensemble also comes at a greater cost, however, the trade off we received was more accurate results.

We also ran into problems with the run times of other algorithms we tested, for example SVM took a very long time to compile, so much so to the point where we disregarded it as a potential option. This is because the time complexity of the algorithm is heavily dependent on the size of the dataset.

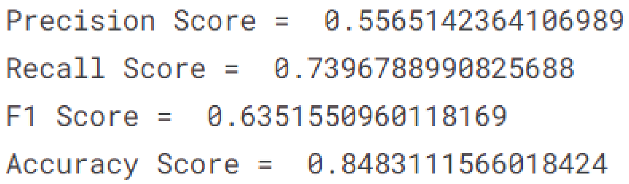
Ensemble Run

After running our first ensemble run, we obtained the following results:

tn	fp	fn	tp
3499	514	227	645

Figure 16

With these results, the following scores were obtainable:



```
Precision Score =  0.5565142364106989
Recall Score =  0.7396788990825688
F1 Score =  0.6351550960118169
Accuracy Score =  0.8483111566018424
```

Baseline model accuracy

During research we came across a way to predict base model accuracy based on TrueNegative, FalsePositive, FalseNegative, and TruePositive (TN, FP, FN, TP). The most common occurrence given these values for TN,FP,FN, and TP would be a model that only predicts people who make 50K or less per year. Based on the current run of our ensemble method program displayed on Kaggle, our values are 3499, 514, 227, 645 (TN, FP, FN, TP respectively). To calculate the base model accuracy we take all the occurrences of 50K or less (TN + FP) 3499 + 514 = 4013. We then divide this by the total occurrences, 4885, to get our base model accuracy. 4013/4885 = .8214943705, or 82%. We will use this base model and compare it to the accuracy of our ensemble method.

Results

When running our singular Naive Bayes algorithm on the dataset, we were given an accuracy score of 79% (rounded from: 0.7858751279426817). As we initially suspected, when we ran our ensemble model on the dataset, we were given a better accuracy score of 85% (rounded from: 0.8483111566018424). Our ensemble method is 6 percentage points greater, and 7.6% ((.85-.79)/.79) improved performance than the singular Naive Bayes algorithm. The accuracy we obtained from our ensemble method is 85% compared to the baseline model of 82% for a net increase of three percentage points, and 3.6% ((.85-.82)/.82) improved performance over the baseline model.