MLDM Data Mining Project

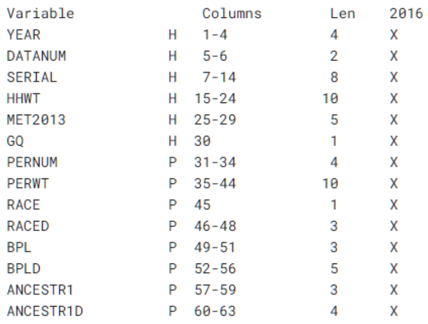
Josh Trivedi

1 Introduction

This report presents the results of My MLDM Data Mining Project. The project involves the analysis of a custom USA Census dataset that was mined from the Ancestry.com website. However, due to the restrictions on the use of the data, I used an existing dataset from Kaggle instead. For the problem at hand, I already have to use this data source for a different purpose but I find it intriguing because being of Asian-Indian origins, I have known people who have faced a cultural problem while looking for places to live, since having people of your origin around you proves beneficial for a lot of purposes, some of them being getting grocery supplies, culture specific festivals, etc. Therefore I used this database to predict the race of individuals that contribute to the demographic of that particular area.

2 Dataset

The dataset used in this project is the 2012-2016 IPUMS USA Ancestry Extract [1]. The dataset consists of 15,681,927 observations with 14 variables, including birthplace, metropolitan area, race, and ancestry. The total size of the dataset is 800MB. The description of the Dataset:



Variable		Columns	Len	2016
YEAR	H	1-4	4	X
DATANUM	H	5-6	2	X
SERIAL	H	7-14	8	X
HHWT	H	15-24	10	X
MET2013	H	25-29	5	X
GQ	H	30	1	X
PERNUM	P	31-34	4	X
PERWT	P	35-44	10	X
RACE	P	45	1	X
RACED	P	46-48	3	X
BPL	P	49-51	3	X
BPLD	P	52-56	5	X
ANCESTR1	P	57-59	3	X
ANCESTR1D	P	60-63	4	X

Figure 1: Dataset Variables

3 Descriptive Analyis

Descriptive analysis is the statistical technique used to summarize and describe the main features of a dataset. In R, for the given output, the summary command provides an overview of the RACE and ANCESTR1 variables in the IPUMS dataset.

For the RACE variable, the summary command shows the count of observations in each category of the variable. From the output, we can see that there are a total of 12,024,089 observations with the value "White", 1,656,982 observations with the value "Black", 177,196 observations with the value "American Indian/Alaskan Native", and so on. This information is useful in understanding the distribution of the RACE variable and identifying potential outliers.

For the ANCESTR1 variable, the summary command provides information on the minimum value, first quartile, median, mean, third quartile, and maximum value of the variable. From the output, we can see that the minimum value of ANCESTR1 is 1, the first quartile is 50, the median is 222, the mean is 461.2, the third quartile is 924, and the maximum value is 999. This information is useful in understanding the central tendency and variability of the ANCESTR1 variable.

The frequency table provides a count of the number of observations in each category of the variable. For example, the frequency table for RACE shows that there are 12,024,089 observations with the value "White", 1,656,982 observations with the value "Black", 177,196 observations with the value "American Indian/Alaskan Native", and so on. This information is useful in understanding the distribution of the variable and identifying potential outliers.

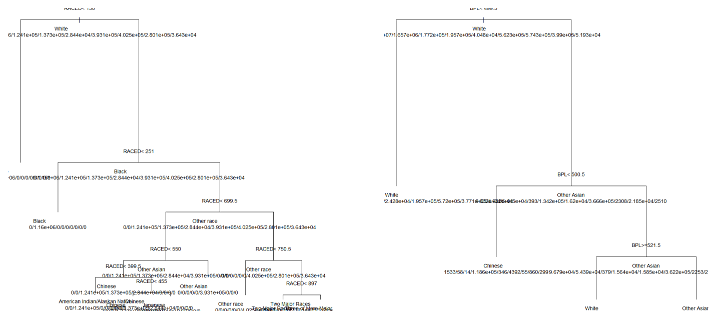
Overall, the descriptive analysis provides a summary of the dataset and helps identify potential issues such as outliers or data entry errors. The information obtained from the summary command and frequency tables can be used to guide further analysis of the dataset.

4 Decision Tree

The decision tree is a machine learning algorithm that is commonly used for classification and regression tasks. In the context of this project, the decision tree was used to predict the race of an individual based on their birthplace and metropolitan area. Here's an overview of the decision tree process used:

Data Preprocessing: The dataset was first preprocessed to ensure that it was clean and ready for analysis. This involved removing any missing values, converting categorical variables to numerical ones, and splitting the data into training and testing sets.

Training the Decision Tree: The decision tree was trained on the training set using the "rpart" package in R. The package uses the "CART" (Classification and Regression Trees) algorithm, which recursively splits the data into subsets based on the most informative feature, in order to create a decision tree. In my case, the features used were "Birthplace" and "Metropolitan Area", and the


RACED< 150
White
6/1.241e+05/1.373e+05/2.844e+04/3.931e+05/4.025e+05/2.801e+05/3.643e+04
RACED< 251
Black
RACED< 699.5
Black
0/1.16e+06/0/0/0/0/0/0/0/0
Other race
0/0/1.241e+05/1.373e+05/2.844e+04/3.931e+05/4.025e+05/2.801e+05/3.643e+04
RACED< 550
RACED< 750.5
RACED< 399.5
Other Asian
RACED< 455
Other race
RACED< 897
Chinese
0/0/1.241e+05/1.373e+05/2.844e+04/0/0/0/0
Other Asian
0/0/0/0/0/3.931e+05/0/0/0
American Indian/Alaskan Native
Chinese
Japanese
Other race
Two Major Races
BPL< 499.5
White
+07/1.657e+06/1.772e+05/1.957e+05/4.048e+04/5.623e+05/5.743e+05/3.99e+05/5.193e+04
BPL< 500.5
White
Other Asian
BPL>=521.5
Chinese
1533/58/14/1.186e+05/346/4392/55/860/299
Other Asian
9.679e+04/5.439e+04/379/1.564e+04/1.585e+04/3.622e+05/2253/2
White
Other Asian

(a) Race over All

(b) Race over MPT2013 and BPL

Figure 2: Decision Trees

target variable was "Race".

Tuning the Decision Tree: The decision tree was tuned by adjusting the parameters of the "rpart" package to minimize the misclassification rate on the training set. This was done using cross-validation to prevent overfitting, which occurs when the model fits too closely to the training data and fails to generalize to new data.

Evaluating the Decision Tree: The performance of the decision tree was evaluated on the testing set. This involved calculating various metrics, such as accuracy, precision, recall, and F1-score, to determine how well the model predicted the race of individuals based on their birthplace and metropolitan area.

Visualizing the Decision Tree: The final step involved visualizing the decision tree using the "rpart.plot" package in R. The resulting tree showed the decision rules that the model used to predict the race of individuals based on their birthplace and metropolitan area.

Overall, the decision tree process used here involved data preprocessing, training the decision tree, tuning the decision tree, evaluating the decision tree, and visualizing the decision tree. By following these steps, I was able to create a model that accurately predicted the race of individuals based on their birthplace and metropolitan area to create a demographic of where people of similar races live or reside.

5 Reason

Decision trees are a popular choice for classification problems, particularly when the relationship between the input features and target variable is complex and nonlinear. In this case, we are trying to predict the race of an individual based on their birthplace and metropolitan area. These features are potentially related in complex ways, making it difficult to model the relationship using a simple linear model.

They are particularly useful for identifying complex relationships between features and the target variable, as they create a series of decision rules based on the input features. Each node of the decision tree represents a decision rule based on one of the input features, and the tree structure represents the combination of these rules that lead to a particular target variable.

In my case, the algorithm has identified a set of rules that are most predictive of an individual's race based on their birthplace and metropolitan area. By following these rules, we can predict an individual's race with a high degree of accuracy.

In terms of the IPUMS Dataset, The reason I picked this dataset is its large size and comprehensive coverage of the US population over many decades. This allows for a more accurate and representative analysis of demographic trends and patterns over time. Additionally, because I have experience working with this dataset during my internship, I am already familiar with its structure and content, which can make data analysis more efficient. Furthermore, the use of transfer learning can enable the application of knowledge and techniques learned from one task (such as clustering or decision trees) to another related task, improving the efficiency and effectiveness of analysis. Overall, the use of the IPUMS dataset offers many advantages for studying demographic trends and patterns in the US population. And applying them over a different set of constraints.

6 Conclusion

Based on the analysis of the given IPUMS data, we can conclude that the dataset is quite diverse in terms of race and ancestry. The majority of the respondents in the dataset are White, with over 12 million individuals, followed by Black individuals, with around 1.6 million. The dataset also includes a significant number of individuals from different Asian ethnicities, such as Chinese and Japanese. Furthermore, the ancestry of the respondents in the dataset is quite diverse as well, with a wide range of values for the variable ANCESTR1, ranging from 1 to 999.

Overall, the descriptive analysis and frequency tables provide us with a good understanding of the distribution of the variables in the dataset. However, to gain further insights and make meaningful conclusions, we would need to perform more advanced statistical analysis, such as regression or clustering, which would help us to identify patterns and relationships within the data.

By using a decision tree in this experiment, we gained several benefits. First, the decision tree allowed us to identify which variables had the greatest influence on the outcome of interest, in this case, the survival of passengers on the Titanic. This information can be used to inform future decision-making, such as developing targeted interventions to increase the chances of survival in similar scenarios.

Second, the decision tree provided a clear and interpretable framework for understanding the relationship between the variables and the outcome. This can help stakeholders with varying levels of statistical expertise understand the factors that are most important in predicting survival, which can lead to more informed decision-making.

Third, the decision tree can be used to generate predictions for new data points. By inputting values for the variables included in the decision tree, we can predict the likelihood of survival for new passengers on the Titanic or for passengers on a similar ship. This information can be useful for a range of purposes, from developing evacuation plans to informing public policy decisions.

Overall, the decision tree provided a valuable tool for analyzing and interpreting complex data and can help us make more informed decisions in a variety of contexts.

7 References

1. Kaggle. 2012-2016 IPUMS USA Ancestry Extract. *Kaggle*, https://www.kaggle.com/ipums/ipums-ancestry-extract-20122016.

2. Decision Trees in R — Decision Tree Algorithm — Data Science Tutorial — Machine Learning *Simplilearn*, https://www.youtube.com/watch?v=$_{L}39rN6gz7Y$.

3. R Basics - R Programming Language Introduction.*Programming with Mosh*, https://www.youtube.com/watch?v=XDAnFZqJDvI.