**PROJECT 1 REPORT**

**FREQUENT PATTERNS AND**

**ASSOCIATION RULES**

**Submitted To**

Dr. Longzhuang Li

**Submitted By**

Vinay Datta Pinnaka

Naveen Chada

1. **OBJECTIVES**

* **Domain**: US census data extracted from 1994 to 1995.
* **Benefits derived from Association rule mining**: The potential benefits derived from association rule mining are:

1. Employment status of the entire population in United states.
2. Education levels of normal US citizen.
3. Taxable income amount range for individuals.
4. Female Entrepreneurs in United states
5. **DATA SET DESCRIPTION**

No of Instances: 199523

No of Attributes: 17

The below are the list of attributes.

**Age:** This attribute depicts the age of the individual, this attribute is continuous

**Class of the worker:** Fallowing are the different class labels for the workers.

***Labels:*** *Not in universe, Federal government, Local government, never worked, Private, Self-employed-incorporated, Self-employed-not incorporated, State government, Without pay*

**Education:**  Different education levels of the individuals, Infants who are not able to go to school are labeled as children.

***Labels:*** *Children, 7th and 8th grade, 9th grade, 10th grade, High school graduate, 11th grade, 12th grade no diploma, 5th or 6th grade, Less than 1st grade, Bachelor’s degree(BA AB BS), 1st 2nd 3rd or 4th grade, Some college but no degree, Master’s degree(MA MS MEng MEd MSW MBA), Associates degree-occup /vocational, Associates degree-academic program, Doctorate degree(PhD EdD), Prof school degree (MD DDS DVM LLB JD)*

**Marital status:** Marital status indicate whether a person is married or divorced. It also indicate other labels like Widowed, Never married etc,. Following are the full list of lables for this attribute.

***Labels:*** *Never married, Married-civilian spouse present, Married-spouse absent, Separated, Divorced, Widowed, Married-A F spouse present.*

**Occupation:**  Label name for each individual who are working, if a person is not working either because he is expired or he is unable to work the label mentioned is ‘Not in Universe**.**

***Labels:*** *Not in universe, Professional specialty, Other service, Farming forestry and fishing, Sales, Adm support including clerical, Protective services, Handlers equip cleaners etc, Precision production craft & repair, Technicians and related support, Machine operators assmblrs & inspctrs, Transportation and material moving, Executive admin and managerial, Private household services, Armed Forces.*

**Race:** To which race an individual belongs to, a total of 5 different labels are used to classify each individual.

***Labels:*** *White, Black, Other, Amer Indian Aleut or Eskimo, Asian or Pacific Islander.*

**Hispanic origin:**  This attribute indicates which Hispanic group an individual belongs to. The following are the labels for this attribute.

***Labels:*** *Mexican (Mexicano), Mexican-American, Puerto Rican, Central or South American, all other, Other Spanish, Chicano, Cuban, Do not know, NA.*

**Sex:** This attribute happens to describe about the gender of individuals.

***Labels:*** *Male and Female****.***

**Member of a labor union:**  This attribute tells if any person belongs to any labor union or not.

***Labels: Not associated no,yes.***

**Reason for un employment:** This attribute tells the reason for the person being unemployed.

***Labels:*** Not in universe, Re-entrant, Job loser - on layoff, New entrant, Job leaver, Other job loser.

***Full or part time status:***  *This attribute says weather the individual is employed or not.*

***Labels****:* Children or Armed Forces, Full-time schedules, Unemployed part- time, Not in labor force, Unemployed full-time, PT for non-econ reasons usually FT, PT for econ reasons usually PT, PT for econ reasons usually FT.

**Tax filer stat:**  This attribute indicates whether an individual files the tax returns for his family or as single

***Labels:*** Non filer, Joint one under 65 & one 65+, Joint both under 65, Single, Head of household, Joint both 65+.

**Detailed household summary in household:**  This attribute indicates the family details of the individual.

***Labels:*** detailed household summary in household: Child under 18 never married, Other relative of householder, Nonrelative of householder, Spouse of householder, Householder, Child under 18 ever married, Group Quarters- Secondary individual, Child 18 or older.

**Instance weight:** This is a continuous attribute.

**Country of birth self:** This attribute specifies the birth place of the individual.

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

**Citizenship:** This attribute mentions about the citizenship of the individual.

***Labels:*** Native- Born in the United States, Foreign born- Not a citizen of U S , Native- Born in Puerto Rico or U S Outlying, Native- Born abroad of American Parent(s), Foreign born- U S citizen by naturalization.

**Own business or self-employed:**  This is attribute mentions about an individual saying if he has own business or an employed person.

***Labels***: 0, 2, 1.

**Tax Income:**  This attribute mentions about the maximum taxable income for an individual.

***Labels:*** *>50000, <50000, =50000.*

* 1. **Data Preprocessing**
* In this step we deal with the missing values, As the dataset is huge we find missing values in many instances. If we take a decision to remove the suze of the datset will be reduced to avoid lose of information we filled the empty values with most estimated label fo that particular attribute.
* Another choice we made is to remove the instance which has duplicate values for different attributes. The reason behind this is the FP Growth algorithm that is implemented in HPC cluster accepts only unique values. Data lost in this step is very minimal because most of the attributes in the dataset are nominal and pocesses different lables.

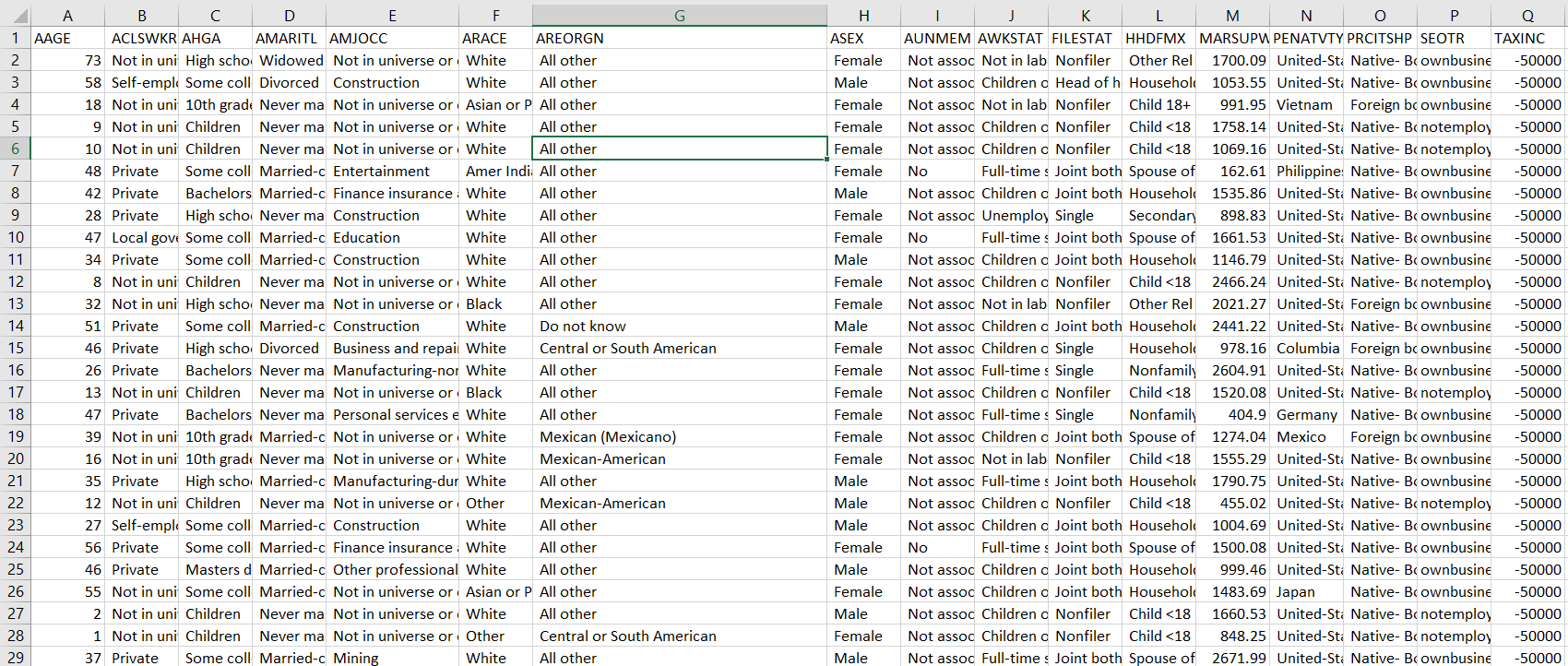


Figure 1: US-census data after preprocessing with all the 20 attributes

1. **Rule mining process**

**3.1 WEKA rule mining process**

Applying numeric to binary filter, to convert all the numeric attributes to binary format. Any continuous attribute remove before applying the filters because to binaries a continuous data it takes lot of time.

Before applying FP-Growth set the parameters such as confidence and the support.

The following images describe the rule mining process and resulting rules:

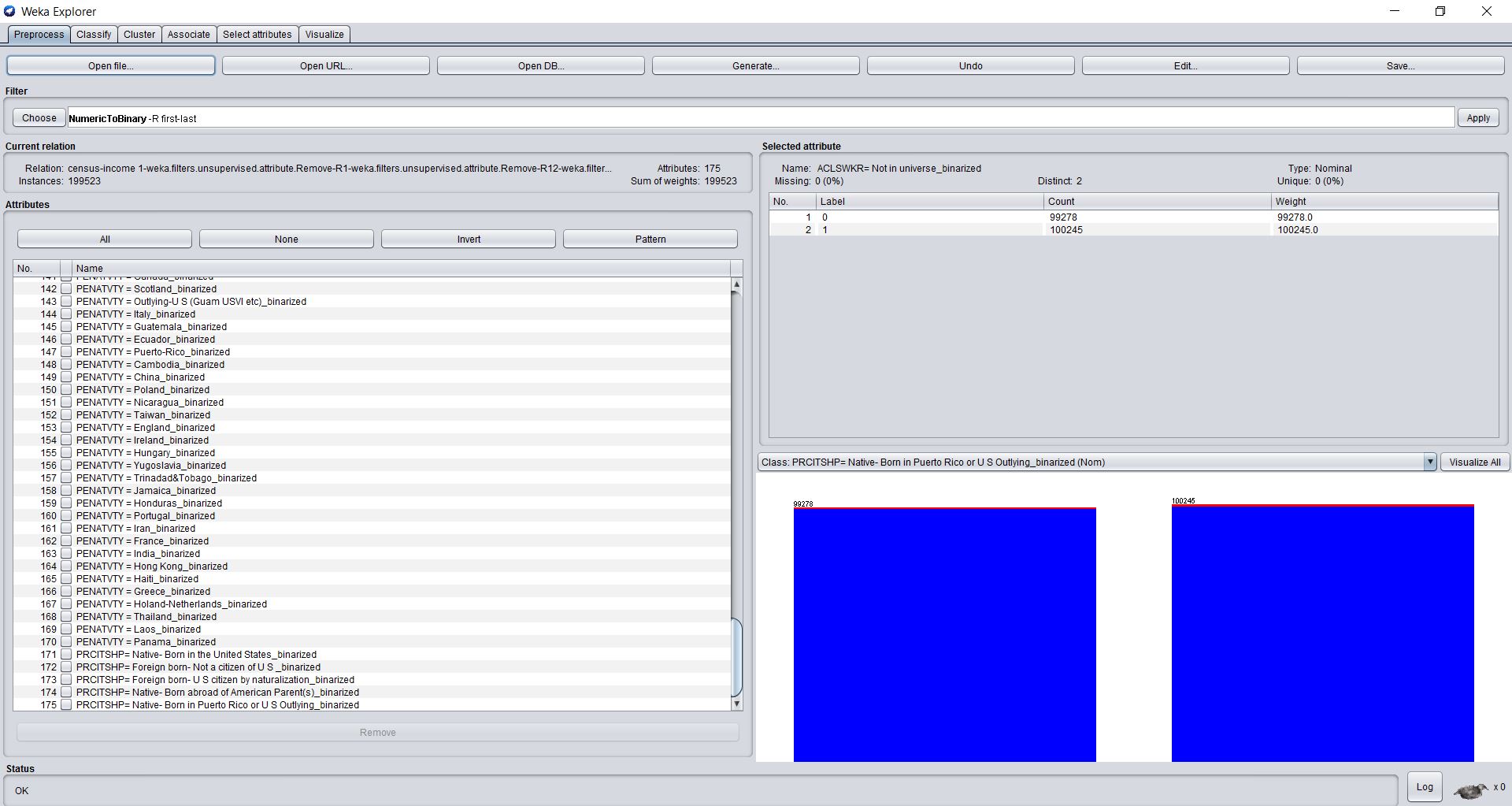


Figure 2: Data after applying filter

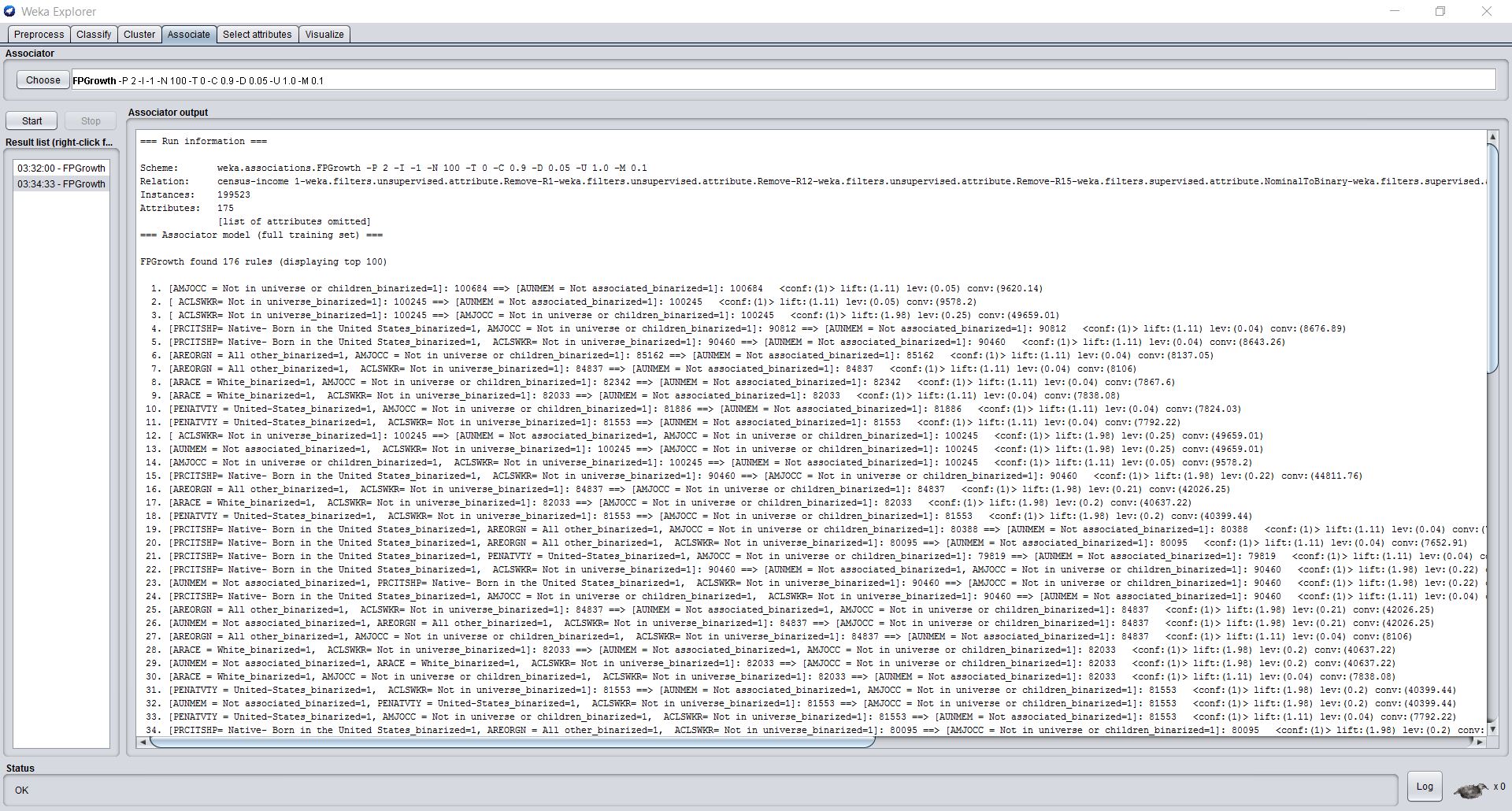


Figure 3: Rules obtained after FP-Growth

**3.2 HPC cluster rule mining process**

On HPC cluster four different ways are followed to obtain the association rules. The following are the machine specifications and type of deployment technique used.

Local machine using maven One node with default settings

Local machine using spark submit One node default settings

Spark standalone cluster with client deploy 20 cores, 4GB memory

YARN cluster using spark submitting process Sbatch job submission



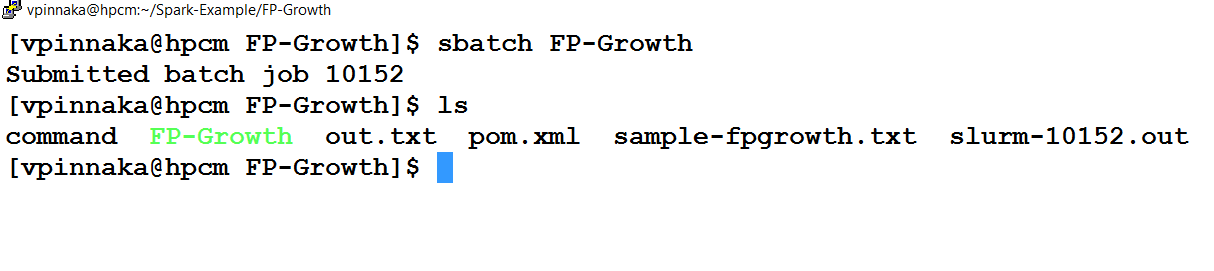


Fig: YARN cluster using sbatch submit

Fig : YARN cluster sbatch job submit.­­­­

**Steps to follow on cluster**

**1.** Transfer the entire zip file of FP-Growth source folder and the data set to the HPC cluster using sftp protocol

**2.** Unzip the file and set the hadoop path in .bashrc file, build the maven package.

**3.** Execute each command from the command file on the cluster.

**4.** Note the execution times and copy the association rules to a text file.

1. **Resulting rules**

The resulting rules are 22000 in number but following are identified as relevant to our expected benfits.

Rule 1: Employment details and type of job each individual is associated with.

[[ Male, ownbusiness, Native- Born in the United States]], 60392

[[ Private, Native- Born in the United States]], 61833

[[ Private, ownbusiness, Native- Born in the United States]], 61422

[ Male, Native- Born in the United States, Not associated, -50000] => [ United-States], 0.9035487464375099

Rule 2: Education levels of US individual

[[ High school graduate, Native- Born in the United States]], 43440

[ High school graduate, Native- Born in the United States] => [ United-States], 0.9311003683241252

Rule 3: Taxable income for each individual

[[ Nonfiler, Never married, Not associated, -50000]], 61012

[ Married-civilian spouse present, United-States, -50000] => [ownbusiness], 0.9860935319508

Rule 4: Female in business

[ Female, ownbusiness, United-States, All other, Native- Born in the United States] => [-50000], 0.9644250706110921

1. **Findings and Recommendations**

The following are some of the findings we obtained from data.

* Most of the united states population works in the private sector. Among the working population most of them are male.
* Most of the US citizens are High school graduates.
* Taxable income is less than 50000.
* Female Entrepreneurs are taxable less than $50000.
* Based on our findings we recommend The US government to focus on Education for citizens and reduce the taxes to encourage more women in Business.

1. **Performance comparison**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tool used** | **WEKA local machine** | **Local machine using maven**  **[one node]** | **Local machine using spark submit**  **[one node]** | **Spark standalone cluster with client deploy**  **[20 cores 4G]** | **Spark standalone cluster with client deploy**  **[40 cores 4G]** | **YARN cluster using spark submit** |
| **Execution time** | **Appx. 3 min {reading, applying filter and rule mining}** | **24.207 sec** | **23.449sec** | **16.945 sec** | **15.372 sec** | **9.449 sec** |

The following table indicates performance comparison between cluster and WEKA.