1 Data Description

1.1 Data Source

The data used for this assignment is the 'Pima Indians Diabetes Database' originally the National Institute of Diabetes and Digestive and Kidney Diseases, but retrieved from Kaggle.[1] The data is comprised of information from 768 individuals taken in 1990, all patients being females of at least 21 years old of Pima Indian heritage.

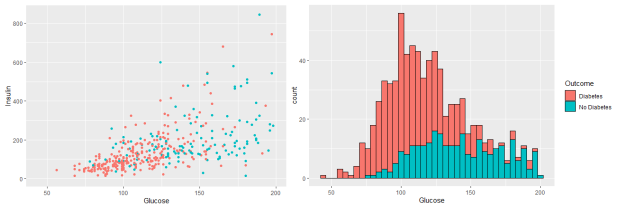

Figure 1: Glucose Levels vs. Insulin Levels (Left), Count of Glucose Levels (right)

1.2 Variable Details

The data set is made up of 8 features and 1 outcome variable. All of the features are numeric variables, and the outcome is a binary variable denoted whether a patient has diabetes or not. For the blood pressure variable, 35 zero values were found which is biologically impossible. These values were replaced with the median blood pressure from the respective outcome of that patient. The same was done for the 11 zero values for BMI, and the 5 zero values for glucose. The variables used in the analysis are described here:

- **Pregnancies:** Number of pregnancies the patient has had (0-17)
- **Glucose:** Plasma glucose concentration after 2 hours in an oral glucose tolerance test (44-199)
- **Blood Pressure:** Diastolic blood pressure (24-122 mm Hg)
- **Skin Thickness:** Triceps skin fold thickness (0-99 mm)
- **Insulin:** 2-Hour serum insulin (0-846 mU/mL)
- **BMI:** Body mass index (18.2-67.1)
- **Diabetes Pedigree Function:** Function scoring family history of diabetes (0.078-2.42)
- **Age:** Age of the patient (21-81)
- **Outcome:** Whether a patient has diabetes (1) or not (0)

images/IncSex.png

Figure 2: Representation of sex in census grouped by income level

1.3 Descriptive Analysis

For this descriptive analysis we will focus on some biological factors that are most common when dealing with diabetes. These will include glucose, insulin, pregnancies and BMI.

[1] **Dua:2019**.

Figure 3: Representation of race in census grouped by income level

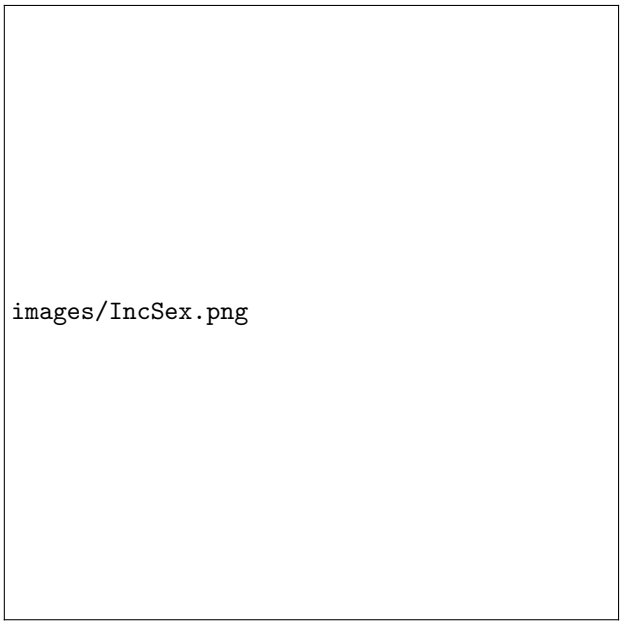

Figure 4: Representation of sex in census grouped by income level

2 Problem Description

The aim of this analysis is to determine what are the major factors which determine the level of income of individuals. Specifically, the analysis will also look at how factors such as sex and race affect other variables, especially income. Through this analysis we will look to see if there are visible signs of income inequality based on the demographics of this population.

3 Technique Description

Association rules are used commonly for the analysis of transactional data. Transactional data means that each variable of the data can be coded exclusively in binary variables.[2] Association rules can find frequently occurring relationships in the data between sets of antecedents and consequents, finding which variables are likely to occur with others. To quantify this, metrics such as support, confidence, and lift are used.

$$\textbf{Support: } s(A \Rightarrow B) = P(A, B)$$
$$\textbf{Confidence: } c(A \Rightarrow B) = P(A|B) = \frac{P(A,B)}{P(A)}$$
$$\textbf{Lift: } L(A \Rightarrow B) = \frac{c(A \Rightarrow B)}{P(B)} = \frac{P(A,B)}{P(A)P(B)}$$

From these we can aim to rank rules based on standardized lift, which is determined by the minimum and maximum confidence thresholds.

$$\textbf{Std. Lift: } \mathcal{L}(A \Rightarrow B) = \frac{L(A \Rightarrow B) - \lambda}{\upsilon - \lambda}$$
$$\upsilon = \frac{1}{max\{P(A), P(B)\}}$$
$$\lambda = max\{\frac{P(A)+P(B)-1}{P(A)P(B)}, \frac{4s}{(1+s)^2}, \frac{s}{P(A)P(B)}\frac{c}{P(B)}\}$$

This now gives us a comparative measure for which to evaluate the strength of the relationships between antecedents and consequents.

4 Result Description

The association rules analysis described in Section 3 was carried out using the 'arules' package in R.[3] Two interesting analyses were conducted, shown in Table 1 and Table 2. The first analysis, shown in Table 1 was generated using association rules based on the fixed right hand side of 'Income: >$50k', with a minimum support of 0.1, a minimum confidence of 0.3, and rule length between 2 and 4. Interesting results from this table are that the strongest antecedents to earning more than $50k are being married, white, or a male (or a combination of these). Another interesting result is that working more than 40 hours a week does relate to making more money. There were many redundant rules such as 2,3,4,5,7,9,10 due to the first rule being that an individual is married.

The second analysis, shown in Table 2 was generated using association rules based on the fixed left hand side with every category of race and gender (see Section 1.2), with a minimum support of 0.1, a minimum confidence of 0.3, and rule length between 2 and 3. Interesting results from this table are that the strongest consequent to being Black, female, or Asian are that you make less than $50k annually. There were a few redundant rules such as 3,5,9 some being due to combinations of the groups mentioned above.

5 Conclusions

Our analysis has successfully accomplished our goal of determining the leading variables that are related

[2] Prof. Sharon McNicholas. *CSE780: Association Rules*. McMaster University, School of Computational Science & Engineering. 2021.

[3] Michael Hahsler et al. *arules: Mining Association Rules and Frequent Itemsets*. R package version 1.6-6. 2020. URL: https://CRAN.R-project.org/package=arules.

to the income level of an individual. We have seen a relationship in the data indicating a large connection between making more than $50k per year and being any combination of White, married, and male. On the other side, we have seen that being Non-White and female tend to lead to earning less than $50k annually. These findings highlight differences in income that could be explained by access to opportunity, socioeconomic factors between demographics, as well as concepts such as the gender wage gap.

Table 1: High Income Result Table

Id.	Rule	Supp.	Conf.	Lift	$\mathcal{L}$
1	{Married} $\Rightarrow$ {>50K}	0.207	0.437	1.814	0.656
2	{Married, White} $\Rightarrow$ {>50K}	0.189	0.448	1.859	0.544
3	{Married, Male} $\Rightarrow$ {>50K}	0.183	0.441	1.829	0.504
4	{Married, White, Male} $\Rightarrow$ {>50K}	0.168	0.448	1.860	0.432
5	{Private, Married} $\Rightarrow$ {>50K}	0.130	0.423	1.755	0.216
6	{White, Male} $\Rightarrow$ {>50K}	0.187	0.318	1.319	0.161
7	{Private,Married,White} $\Rightarrow$ {>50K}	0.120	0.437	1.814	0.139
8	{>40 hrs.} $\Rightarrow$ {>50K}	0.118	0.402	1.671	0.131
9	{Private, Married, Male} $\Rightarrow$ {>50K}	0.116	0.429	1.781	0.115
10	{Middle-aged, Married} $\Rightarrow$ {>50K}	0.114	0.434	1.802	0.102

Table 2: Demographics Result Table

Id.	Rule	Supp.	Conf.	Lift	$\mathcal{L}$
1	{Black, Female} $\Rightarrow$ {<=50K}	0.0450	0.942	1.241	0.884
2	{Female} $\Rightarrow$ {<=50K}	0.295	0.891	1.173	0.781
3	{White, Female} $\Rightarrow$ {<=50K}	0.234	0.881	1.161	0.762
4	{Black} $\Rightarrow$ {<=50K}	0.0841	0.876	1.154	0.752
5	{Black, Male} $\Rightarrow$ {<=50K}	0.039	0.810	1.068	0.621
6	{Male} $\Rightarrow$ {Married}	0.415	0.621	1.312	0.585
7	{Asian-Pac-Islander} $\Rightarrow$ {<=50K}	0.023	0.734	0.967	0.468
8	{White, Female} $\Rightarrow$ {Private}	0.193	0.727	1.043	0.454
9	{White, Male} $\Rightarrow$ {Married}	0.375	0.637	1.344	0.450
10	{Female} $\Rightarrow$ {Private}	0.238	0.720	1.033	0.440