**CSCI 317 Lab 7 – Association Rule Analysis**

1. **Describe the data set. What is the data about, attributes of the data set, special features if applicable, ... (10’)**

The dataset is about diabetes (Pima Indians Diabetes Database), specifically focusing on predicting whether a patient has diabetes based on various diagnostic measurements. The attributes include:

* preg : Number of times pregnant
* plas: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
* pres: Diastolic blood pressure (mm Hg)
* skin: Triceps skin fold thickness (mm)
* insu: 2-Hour serum insulin (mm U/ml)
* mass: Body mass index (weight in kg/(height in m)^2)
* pedi: Diabetes pedigree function
* age: Age (years)
* class: Class variable (tested\_positive or tested\_negative)

Each entry represents an individual patient assessment for diabetes. There are 768 entries, with no null values present. The class attribute serves as the target variable for predictive modeling/classification tasks.

1. **What is the purpose of your analysis? (20’)**

The primary purpose of this analysis is to discover association rules that may exist between the various attributes related to the diagnosis of diabetes. This analysis could reveal interesting patterns, such as risk factors that increase the likelihood of diabetes when combined. This analysis could be useful for predictive healthcare and patient management.

1. Show each step of your analysis process.(40’)

**Using Weka**

1. **Loading Data into Weka**

* Open Weka -> Click `Explorer`
* In the `Preprocess` tab, click Open File -> load diabetes.arff

A screenshot of a computer

Description automatically generated

1. **Preprocessing Data**

* Numeric attributes might need to be converted to nominal as Weka’s association rule miners require nominal data.
* Select all `numeric` attributes in list -> filter -> unsupervised -> attribute -> NumericToNominal -> Apply

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Running Association Rule Analysis**

* `Associate` tab -> choose `Apriori` (I chose the default settings for Apriori) -> `Start`

A screenshot of a computer

Description automatically generated

**Using R**

* Each step is described in the R Script

1. What rules are most interesting and why? (20’)

**From Weka**

* + All three top rules involve the condition **skin=0**, which leads to **insu=0** with a confidence of 1 (100%). This means whenever the triceps skin fold thickness is recorded as 0, the 2-hour serum insulin measurement is also 0.
  + This might be related to data entry, where skin thickness measurement was not taken (or not available – due to which it is 0), insulin measurements were probably not available too.
  + These rules have a lift of 2.05, which means that the likelihood of both **skin=0** and **insu=0** occurring together is about twice as likely as it would be if they were independent.
  + The **lev** (leverage) statistic shows the proportion of additional instances covered by both parts of the rule above expected if they were independent. For example, the first rule's leverage of 0.15 indicates a relatively high dependency between the attributes.
  + **conv** (conviction) has a value of 116.46 which means that it would be very reliable in predicting **insu=0** given **skin=0**.
  + In a practical healthcare setting, this kind of rule, while statistically strong, might not offer direct clinical insights but does indicate the importance of complete data collection.

**From R**

* Similar association to the one in Weka where skin not being measured Is linked to insulin not being measured, with a confidence of 100%
* Rule [8] **{plas=Low, age=Young} => {class=tested\_negative}** with a confidence of 98.06% and a lift of 1.51 indicates that younger individuals with low plasma glucose levels are highly likely to test negative for diabetes.
* Rule [9] **{mass=Low, age=Young} => {class=tested\_negative}** has a high confidence of 97.85% and a lift of 1.50, suggesting that lower body mass in younger individuals is linked to a negative diabetes test.
* Rule [10] **{plas=Low, insu=Low} => {class=tested\_negative}** has a confidence of 95.06% and a lift of 1.46, indicating that lower levels of both plasma glucose and insulin are strong predictors of a negative diabetes test outcome.

**Result from R Script**

> # Association rule analysis using apriori

> # The thresholds/settings were set similar to the one used in Weka

> rules <- apriori(data, parameter = list(supp = 0.1, conf = 0.9, target = "rules", maxlen = 10))

**Apriori**

**Parameter specification:**

**confidence minval smax arem aval originalSupport maxtime support minlen maxlen target ext**

**0.9 0.1 1 none FALSE TRUE 5 0.1 1 10 rules TRUE**

**Algorithmic control:**

**filter tree heap memopt load sort verbose**

**0.1 TRUE TRUE FALSE TRUE 2 TRUE**

**Absolute minimum support count: 76**

**set item appearances ...[0 item(s)] done [0.00s].**

**set transactions ...[33 item(s), 768 transaction(s)] done [0.00s].**

**sorting and recoding items ... [33 item(s)] done [0.00s].**

**creating transaction tree ... done [0.00s].**

**checking subsets of size 1 2 3 4 done [0.00s].**

**writing ... [16 rule(s)] done [0.00s].**

**creating S4 object ... done [0.00s].**

> rules.sorted <- sort(rules, by = "confidence")

> # Inspecting the top 10 rules

> inspect(head(rules.sorted, n = 10))

**lhs rhs support confidence coverage lift count**

**[1] {skin=Not Measured} => {insu=Not Measured} 0.2955729 1.0000000 0.2955729 2.053476 227**

**[2] {skin=Not Measured,**

**pedi=Low} => {insu=Not Measured} 0.1015625 1.0000000 0.1015625 2.053476 78**

**[3] {skin=Not Measured,**

**mass=Low} => {insu=Not Measured} 0.1002604 1.0000000 0.1002604 2.053476 77**

**[4] {skin=Not Measured,**

**class=tested\_positive} => {insu=Not Measured} 0.1145833 1.0000000 0.1145833 2.053476 88**

**[5] {skin=Not Measured,**

**age=Middle-Aged} => {insu=Not Measured} 0.1835938 1.0000000 0.1835938 2.053476 141**

**[6] {skin=Not Measured,**

**class=tested\_negative} => {insu=Not Measured} 0.1809896 1.0000000 0.1809896 2.053476 139**

**[7] {skin=Not Measured,**

**age=Middle-Aged,**

**class=tested\_negative} => {insu=Not Measured} 0.1002604 1.0000000 0.1002604 2.053476 77**

**[8] {plas=Low,**

**age=Young} => {class=tested\_negative} 0.1315104 0.9805825 0.1341146 1.506175 101**

**[9] {mass=Low,**

**age=Young} => {class=tested\_negative} 0.1184896 0.9784946 0.1210937 1.502968 91**

**[10] {plas=Low,**

**insu=Low} => {class=tested\_negative} 0.1002604 0.9506173 0.1054688 1.460148 77**

> # Save the rules to a file

> write(rules, file="diabetes\_rules.txt")

> # View the dataset and rules for validation

> head(data)

**preg plas pres skin insu mass pedi age**

**1 Medium High Medium-Low High Not Measured Medium-High High Middle-Aged**

**2 Low Low Medium-Low Medium Not Measured Low Medium-Low Middle-Aged**

**3 High High Medium-Low Not Measured Not Measured Low High Middle-Aged**

**4 Low Low Medium-Low Medium Medium Medium-Low Low Young**

**5 None Medium-High Low High High High High Middle-Aged**

**6 Medium Medium-Low Medium-High Not Measured Not Measured Low Low Middle-Aged**

**class**

**1 tested\_positive**

**2 tested\_negative**

**3 tested\_positive**

**4 tested\_negative**

**5 tested\_positive**

**6 tested\_negative**

> inspect(rules.sorted[1:10])

**lhs rhs support confidence coverage lift count**

**[1] {skin=Not Measured} => {insu=Not Measured} 0.2955729 1.0000000 0.2955729 2.053476 227**

**[2] {skin=Not Measured,**

**pedi=Low} => {insu=Not Measured} 0.1015625 1.0000000 0.1015625 2.053476 78**

**[3] {skin=Not Measured,**

**mass=Low} => {insu=Not Measured} 0.1002604 1.0000000 0.1002604 2.053476 77**

**[4] {skin=Not Measured,**

**class=tested\_positive} => {insu=Not Measured} 0.1145833 1.0000000 0.1145833 2.053476 88**

**[5] {skin=Not Measured,**

**age=Middle-Aged} => {insu=Not Measured} 0.1835938 1.0000000 0.1835938 2.053476 141**

**[6] {skin=Not Measured,**

**class=tested\_negative} => {insu=Not Measured} 0.1809896 1.0000000 0.1809896 2.053476 139**

**[7] {skin=Not Measured,**

**age=Middle-Aged,**

**class=tested\_negative} => {insu=Not Measured} 0.1002604 1.0000000 0.1002604 2.053476 77**

**[8] {plas=Low,**

**age=Young} => {class=tested\_negative} 0.1315104 0.9805825 0.1341146 1.506175 101**

**[9] {mass=Low,**

**age=Young} => {class=tested\_negative} 0.1184896 0.9784946 0.1210937 1.502968 91**

**[10] {plas=Low,**

**insu=Low} => {class=tested\_negative} 0.1002604 0.9506173 0.1054688 1.460148 77**