Data Mining Project

Team

Daniel Hanspeter 6129

Thomas Schievenin 5701

# Datasets:

**Blood transfusion**

**Census-Income**

# PROJECT STEP ONE

## What is the data about?

1. Blood transfusion

This data is as the name says a collection of data taken from blood transfusions and to be more precise, it was collected at the Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The data is collected randomly from donors and is used to achieve a Frequency-Metric-Model of donations, focusing on March 2007.

1. Census-Income

This data is a census done by the US Census Bureau investigating about the income class of randomly taken people, taking in account their education, current living situation and so on.

## How many attributes describe the data? What are the types of these attributes?

1. Blood transfusion

The data collected focuses on the donor’s situation, namely on the amount of done donations before, of the total amount of donations and so on, for a total of 5 Attributes. The data collected is focused on March 2007, since one attribute collects exactly if one person donated in March 2007 or not. All collected data is numerical, except for the donation in March ’07 which is a bit / Boolean attribute.

1. Census-Income

The data is described by 14 attributes detailing each person’s situation (i.e. martial-status). The attributes type vary, since some are numerical values as the age, and others are explicit defined fields (nominal), such as the native country (i.e. United-States)

## Are there missing values? If, yes propose a method to deal with missing values.

1. Blood transfusion

No missing values in the Dataset.

1. Census-Income

Around 7% of Attributes are missing. To work with that data, we can easily set a default value, like for instance an average if the value is continuous or the most occurring in order to not affect statistics.

## If the data has numeric attributes, choose at least two attributes and define their distributions. Represent these distributions using Boxplots. Which kind of conclusions you derive from these representations? Are there any outliers?

1. Blood transfusion

|  |  |  |
| --- | --- | --- |
| Label | Frequency (times) | Monetary (c.c. blood) |
| Min | 1 | 250 |
| Q1 | 2 | 500 |
| Median | 4 | 1000 |
| Q3 | 7 | 1750 |
| Max | 50 | 12500 |
| IQR | 5 | 1250 |
| Upper Outliers | 45 | 45 |
| Lower Outliers | 0 | 0 |
| *For the Box (IQR and Median)* | |  |
| Q2-Q1 | 2 | 500 |
| Q3-Q2 | 3 | 750 |
| *For the Whiskers* | |  |
| Q3+1.5\*IQR | 14,5 | 3625 |
| Q1-1.5\*IQR | -5,5 | -1375 |
| Upper Whisker | 14,5 | 3625 |
| Lower Whisker | 1 | 250 |
| Wupper-Q3 | 7,5 | 1875 |
| Q1-Wlower | 1 | 250 |
| *For the Outliers* | |  |
| Max | 50 | 12500 |
| Min | #N/A | #N/A |

As derivable from the Boxplots in both collected values there are Positive outliers.

For the Monetary (c.c. blood) Boxplot we can say that the values are located in a very short range, by noticing that the whiskers are just below/above the box ranges

1. Census-Income

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Labels | Education-num | hours-per-week |  |  |  |  |  |  |  |  |  |  |  |
| Min | 1 | 1 |  | | | | | | | | | | |
| Q1 | 9 | 40 |
| Median | 10 | 40 |
| Q3 | 12 | 45 |
| Max | 16 | 99 |
| IQR | 3 | 5 |
| Upper Outliers | 0 | 3492 |
| Lower Outliers | 1198 | 5516 |
| *For the Box (IQR and Median)* | | |
| Q2-Q1 | 1 | 0 |
| Q3-Q2 | 2 | 5 |
| *For the Whiskers* | |  |
| Q3+1.5\*IQR | 16,5 | 52,5 |
| Q1-1.5\*IQR | 4,5 | 32,5 |
| Upper Whisker | 16 | 52,5 |
| Lower Whisker | 4,5 | 32,5 |
| Wupper-Q3 | 4 | 7,5 |
| Q1-Wlower | 4,5 | 7,5 |
| *For the Outliers* | |  |
| Max | #N/A | 99 |
| Min | 1 | 1 |
|  |  |  |

We can deduce from the Boxplots that there are Minima and Maxima located largely outside the 50% of the data. “Hours-per-week” gives us an example of a complete Boxplot, having also a Max and Min Outlier.

## Define how to measure the similarity between the data objects according to the attribute types of your datasets?

1. Blood transfusion

Since all values (except the binary value for doing transfusion in march 07) in this Dataset are numeric ones, to compare on similarity we can calculate the distance using the Minkowski distance, by first cleaning up the data by bringing attributes to a unit-less form.

1. Census-Income

Here we have in addition to numeric values also nominal values. In this case we can do a simple matching or we could create a binary mapping for those values

## Define the data mining tasks that can be performed on the chosen datasets:

1. Blood transfusion

For this Dataset we can only do a prediction on the amount of blood donated and the times a person donated and vice versa. So we can build a classifier. The other attributes, like the monthly distance towards the last donation, can’t be connected to the other attributes in a reasonable way.

1. Census-Income

This Dataset gives us more fields to play with. In fact, we can create classifiers, decision trees make predictions etc. upon all attributes, pointing to the income class for instance. For example we could construct a tree which tells us if a 20-24 or a 25-30 with different workclass can be mapped in a >50K income class.

E.G Age + Work 🡺 Class income: >50K /<=50K

25-30

20-24

Yes

No

Yes

No

# PROJECT STEP TWO