Stroke Prediction Knowledge Discovery

## Dataset Description

In order to be able to apply all the knowledge discovery methods presented, in this project I will use a dataset called - Stroke Prediction Dataset, which contains important features that can predict the occurrence of a stroke. This dataset contains 5111 entries and contains 11 features based on patients' lifestyles to see if they have had a stroke or are prone to having one. The distinguishing properties of the dataset are:

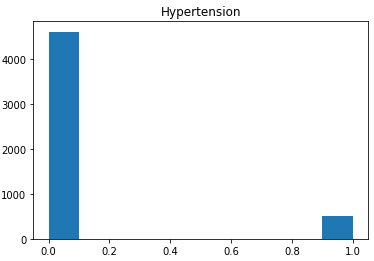
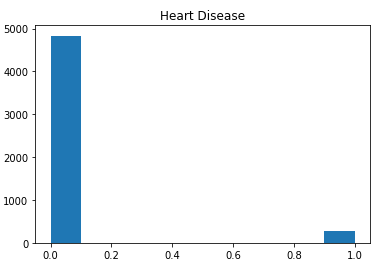
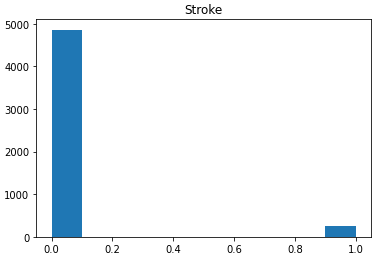
* Binary
  + **Hypertension** - indicate whether the patient suffers from the condition of hypertension
  + **Heart\_disease** - indicate whether the patient suffers from the condition of heart disease
  + **Ever\_married** - Has the patient ever been married?
  + **Stroke** - indicate whether the patient has ever suffered a stroke
* Multi-Valued
  + Label based
    - **Gender** – gender of the patient - female or male
    - **Work\_type** – indicates the work the patient does on a daily basis - self\_employed, private, govt\_job, never\_worked
    - **Residence\_type** - where the patient lives – rural or urban
    - **Smoking\_status** – indicates if the patient’s smoking status – formerly\_status, never\_smoked, smokes
  + Numeric based
    - **Age** – patient’s age
    - **Avg\_glucose\_level** – patient’s average glucose level in blood
    - **Bmi** – body mass index of the patient

## Data Pre-Processing

The dataset is currently unusable for formal concept analysis. In the trials that followed, the dataset was preprocessed to produce several formal contexts for analysis. Based on the previously mentioned qualities, the attributes from the generated contexts will be determined.

The data from the original dataset was examined before the contexts were made. The imbalance of three columns, namely hypertension, heart\_disease and stroke is the first issue we ran into. The majority of patients do not suffer from hypertension, as shown in Figure 1, nor from the heart disease shown in Figure 2, which can be interpreted in the following way - none of these medical conditions influence the occurrence of a stroke. Also, although it is an extremely well-known disease, out of 5111 entries in the dataset representing the number of patients, the number of those who have suffered a stroke is extremely low - as shown in Figure 3. What is evident in terms of the interpretation and

pre-processing of the dataset, is the most severe inconsistency in the columns presented above, however, no column was removed from the dataset based on the inconsistencies presented.

*Figure 3 – Unbalanced set of stroke*

*Figure 2 – Unbalanced set of heart disease*

*Figure 1 – Unbalanced set of hypertension*

Dealing with missing or insufficient data and preparing numerical data for conceptual scaling were some other preparation tasks. The formal ideas did not include the entries that contained the patient's missing information. In order to make it simple to create ordinal scales and include the patients in larger clusters, we rounded the values for the body mass index to the first integer and eliminated the null values from this column. Because the consequences of smoking cigars last a lifetime, we combined the rows that were labeled as smokers or former smokers when determining smoke status. We also deleted the rows that included 'unknown' information. After removing all the inconsistencies and null values from the dataset, it contains 3426 entries.

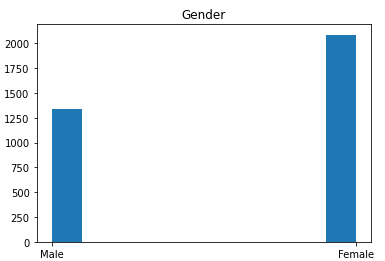
In order to extract knowledge regarding stroke diagnosis, the key features (procedures and personal information) that indicate the susceptibility, and the traits that meaningfully alter the outcomes, we developed several formal contexts in the experiments. We start with a basic context created exclusively from binary features, from which we will extract the initial layer of knowledge, and then we will further examine the concepts discovered using conceptual scaling and the derived context. We will use attribute exploration to determine the implications between attributes based on the formal contexts from which we gathered the most significant knowledge. We will use a variety of attributes as criteria in the final context, which is a triadic formal context.

## Extracting First Layer of Knowledge

### **Dataset Analysis**

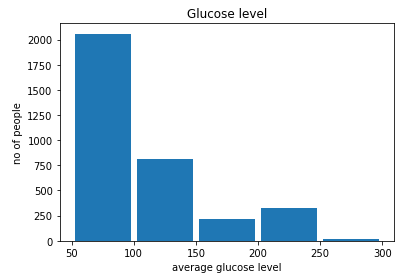
The main causes of stroke are smoking, diabetes, gender and age. For these reasons, we evaluated these columns in the dataset in order to apply knowledge discovery to predict whether these patients are prone to stroke.

A person's gender is another main symptom of this serious condition. Men have a higher risk of stroke than women. They are usually older when they have a stroke and are more likely to die of stroke than men. In the dataset used, it is not so obvious because the number of women participating is higher, as it can be seen from Figure 4.



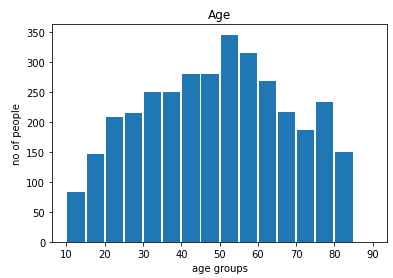
*Figure 4 – Plot showing gender breakdown*

Having diabetes is a high-risk factor for stroke. The normal blood glucose level should be less than 100 mL. In terms of the dataset used as can be seen from Figure 5, many of the patients involved in the study are within the normal range, but there are still many who are above the range and even at worrying values.



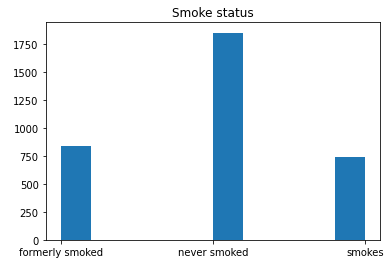
*Figure 5 - Graph displaying blood glucose ranges*

According to several studies, age plays an extremely important role in the main causes of stroke. It is known that when you get older, you are prone to different diseases, but this stroke affects people from a much younger age - people over 55 are at higher risk. We can categorize the patients into various age groups, as seen in Figure 6. We took into account the following age groups to achieve an equitable distribution: 10-15, 20-25, 30-35, 40-45, 50-55, 60-65, 70-75, 80-85. Since there are different numbers of patients in each category, we can pinpoint those patients who are more likely to be stroke-prone. The conceptual scales utilized in the following sections will be based on the age ranges for these groups.



*Figure 6 - Age ranges for patients*

Smoking or exposure to cigarette smoke (passive smoker) is another risk factor, but one that is treatable and as long as it is kept under control can reduce the risk of stroke. Regarding the dataset utilized, as shown in Figure 7, it is encouraging that after processing and eliminating the null values, the number of non-smokers is higher.



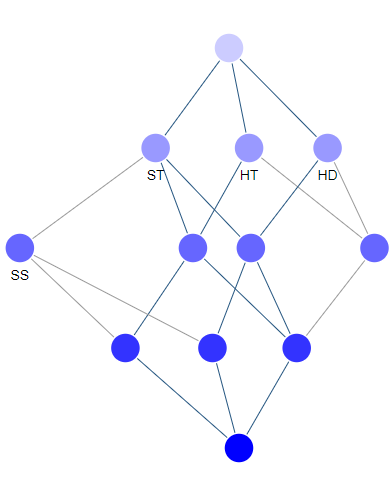
*Figure 7 – Smoke status histogram*

### **Knowledge Extraction Using FCA**

As was already indicated, the binary characteristics from the resulting dataset are used in the initial formal context. The set of objects G for this context (and the ones that follow) comprises of the patients from whose data was gathered. Among the attributes in the set M1 are: heart disease (HD), hypertension (HT), stroke (ST) and smoke status (SS). The patient's status with regard to the property is shown by the relation set I. Table 1 displays a sample from the formal context K 1=(G,M,I), and to visually assess them, we also plotted the idea lattice (Figure 8).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | HD | HT | ST | SS |
| o\_1 |  |  |  |  |
| o\_2 |  |  |  |  |
| o\_3 |  |  |  |  |
| …. |  |  |  |  |
| o\_ 3425 |  |  |  |  |
| o\_ 3426 |  |  |  |  |

*Table 1 – a representative set of the formal context's objects and attributes*



*Figure 8 - Concept lattice that matches the formal context shown in Table 1*

We found that many individuals had heart disease or are smokers when we isolated the ideas with the characteristic "ST" (stroke). This information is supported by experts in the field, however by themselves, these characteristics do not reliably indicate the likelihood of having a stroke. Additionally, there are no theories in which individuals have just the two characteristics and are also experiencing a stroke; thus, even if they may increase the likelihood of having a stroke, these parameters are unreliable for estimating the chance. The quantity of concepts' objects with the "ST" attribute shows how many attributes are frequently taken into account while diagnosing a stroke in patients. One technique of identification is insufficient for determining the probability of having a stroke in the future, as was found in the previous section.

In the last section, we used the assumption that people with heart disease also had hypertension most frequently. The number of items with the qualities HD and HT is rather low among the detected ideas, proving that the assumption was incorrect. As cardiac disease and smoking are frequently found combined in patient histories, we therefore draw the conclusion that there may be a correlation between the two.

In our professional setting, there are many patients who only had a stroke. This is because we didn't include the multi-valued properties in the initial trial. After applying conceptual scaling in the following sections, more information can be gleaned from the resulting contexts.

## Conceptual Scaling

We presented the dataset and noted in the sections before that some features can have more than two values. Smoking status, gender, and features referring to age and body mass index (bmi) will use values from discrete labels, whereas smoking status, gender, and features referring to smoking status will use continuous values. In the sections that follow, the nominal and ordinal types of scales will be discussed. By combining them into derived and nested diagrams for knowledge extraction, we will also be able to develop new formal contexts by utilizing the other set of attributes.

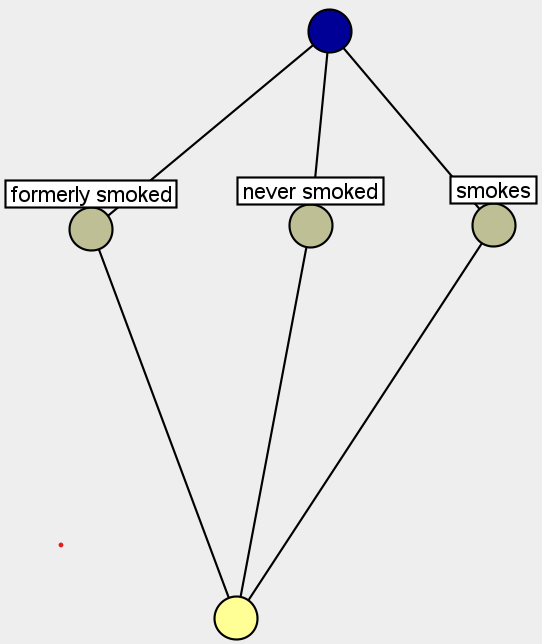
### **Nominal Scaling**

Three smoking status assessments are included in our dataset: smokes, never smokes, and formerly smoker.

We developed a formal context for the conceptual scale (SN1) represented in Table 2 and the associated concept lattice in Figure 9 using ToscanaJ. All of the smoking statuses that the tree may have been represented in the attribute set. The dataset's smoke status is fully included in the attribute set. We are trying to learn more about the patient's smoking status using this scale.



*Table 2 – Smoke Status Conceptual Scale*



*Figure 9 - Smoke Status Concept Lattice*

The next nominal scale (SN2) employed in the research provides details about the patient's line of work so that a pattern in their life can be seen. Private, never worked, self-employed, government job, and children are among the values present in the list of traits. Similar to the scale previously stated, the initial scale we found places all concepts on the same level on the concept lattice. Table 3 depicts the formal context, and Figure 10 shows the concept lattice that goes along with it. However, employing this conceptual scale in actual practice, we were unable to glean very much information from the resulting contexts.

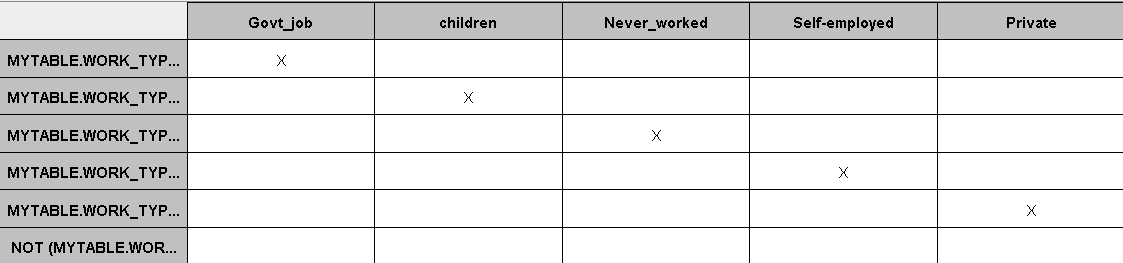


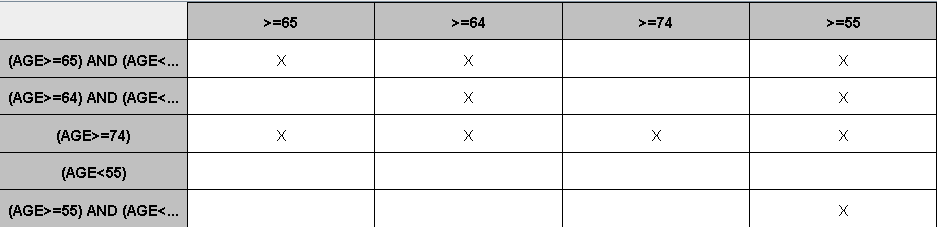
Table 3 - *Work Type Conceptual Scale*

### 

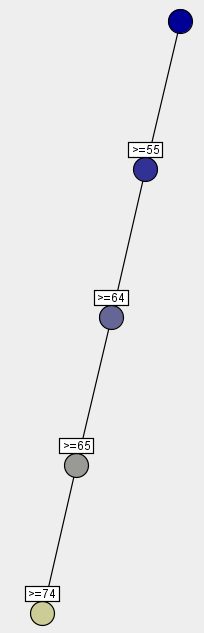
*Figure 10 – Corresponding Work Type Concept Lattice*

### **Ordinal Scaling**

For two characteristics—the patient's age and body mass index—we developed ordinal scales. In terms of age, we adhered to the ranges established in the data analysis section: 50-55, 56-64, 65-74, and >=75. Age increasing ordinal scale formal context. The first scale (SO1) taken into consideration is an ordinal scale with four attributes: >=55, >=64, >=65, and >=74. The formal context that was produced is shown in Table 4 and its lattice in Figure 11.

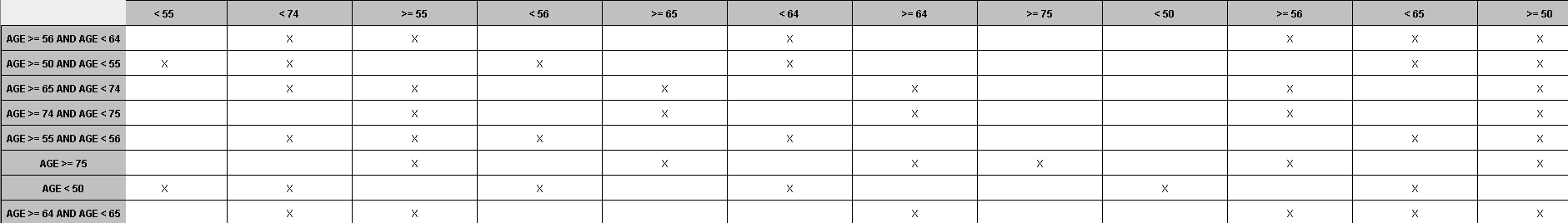


*Table 4 - Age increasing ordinal scale formal context*



*Figure 11 - Age increasing Concept Lattice*

We wanted to understand more during the experiments about age groups' chances of experiencing a stroke in their lifetime. In order to do this, we primarily employed an interval scale (SO2) adding the attributes: <75, <64, <55, and <50. The formal context that results is shown in Table 5 and its concept lattice is shown in Figure 12.



*Table 5 - Age interval ordinal scale formal context*

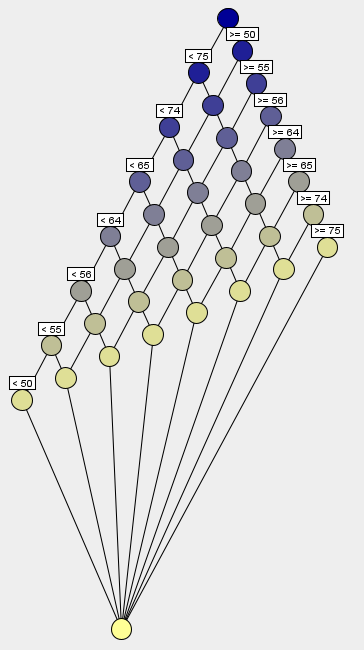
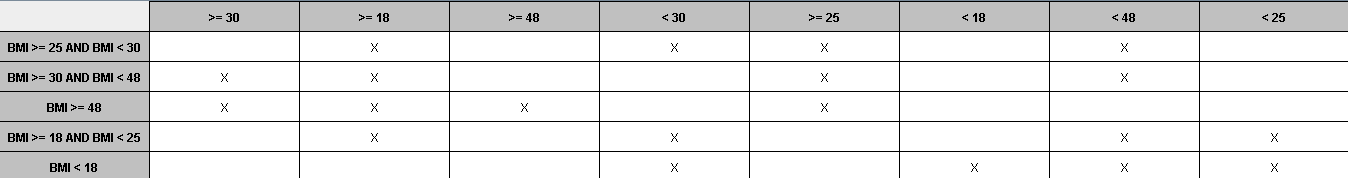
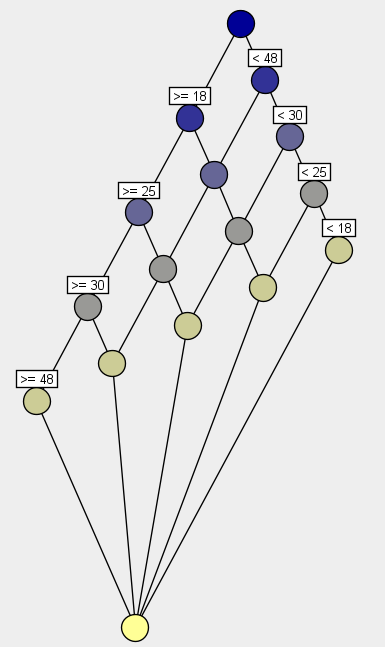


Figure 12 - *Age interval concept lattice*

We attempted to categorize the patients based on their BMI values and used the specialists' findings to create the Body Mass Index ordinal scale (SO3). In other words, people with a BMI under 18 are considered underweight, people with a BMI of 18 to 25 are regarded healthy, people with a BMI of 25 to 30 are considered overweight, and people with a BMI of 30 to 48 are deemed for this, we constructed an interval scale similar to SO2 with the following attributes: =48, =30, =25, =18, >18, >25, >30, >48. The resulting formal context is shown in Table 6 and its concept lattice may be seen in Figure 13.



*Table 6 – BMI ordinal scale formal context*



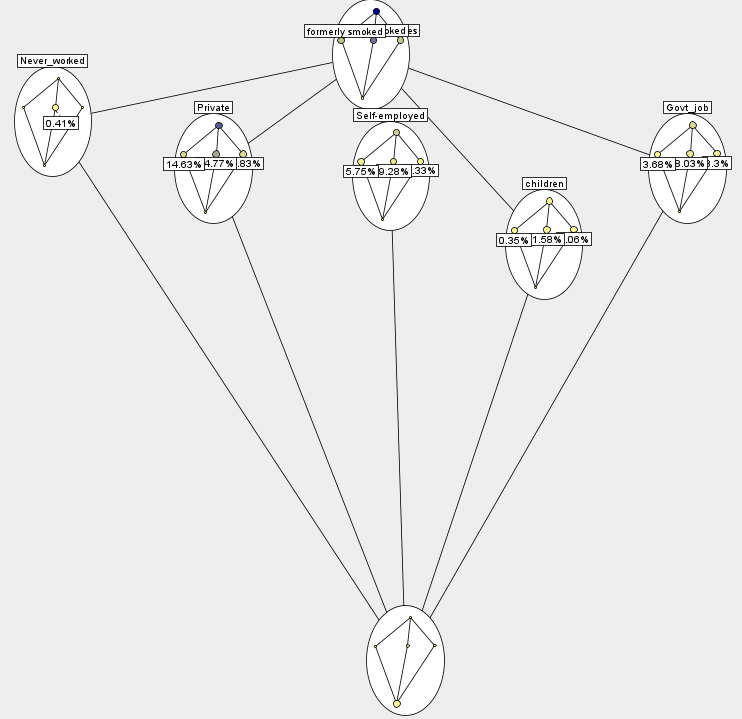
*Figure 13 - BMI interval concept lattice*

### **Knowledge Extraction Using Conceptual Scales**

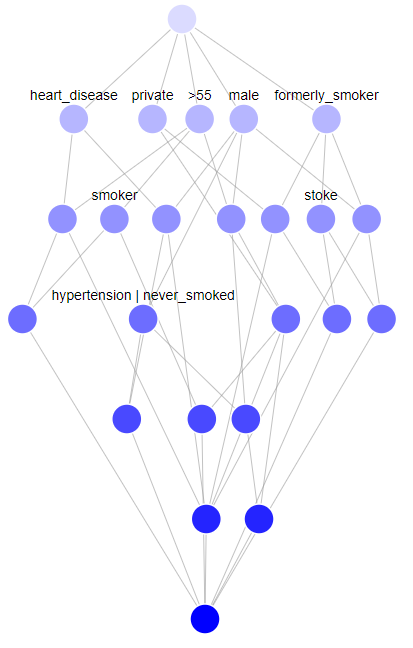
The derived contexts and layered diagrams that were constructed using the conceptual scales created in the earlier sections were used to extract new information about stroke prediction. First, we integrated SN1 and SN2 to understand how smoking status and work environment can affect one another and raise the risk of having a stroke. Figure 14 depicts the resulting layered diagram, which demonstrates that patients who have smoked in the past or who presently smoke are more likely to experience a stroke. If we consider all the major risks that can lead to a stroke are: gender - men being more prone, age - people over 55 are more likely, stressful environment and suffering from various chronic diseases, which can expose the patient to stroke. In order to see if the patients involved are prone to stroke, we combine SN1 with SN2 and the M = {heart disease, hypertension, gender, age, stroke}. The experiment's object set (G) is the same as it was before. Table 7 depicts the incomplete formal environment, and we used Figure 15's concept lattice for knowledge extraction. We collected 21 concepts to use in the concept lattice.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Male | >55 | Private | Hear\_disease | Hypertension | Formerly\_smoker | Never\_smoked | smoker |
| o1534 |  |  |  |  |  |  |  |  |
| o1535 |  |  |  |  |  |  |  |  |
| o3089 |  |  |  |  |  |  |  |  |
| o404 |  |  |  |  |  |  |  |  |
| o3419 |  |  |  |  |  |  |  |  |
| o3420 |  |  |  |  |  |  |  |  |
| o3421 |  |  |  |  |  |  |  |  |

*Table 7 – Derived context using SN1, SN2 and M*

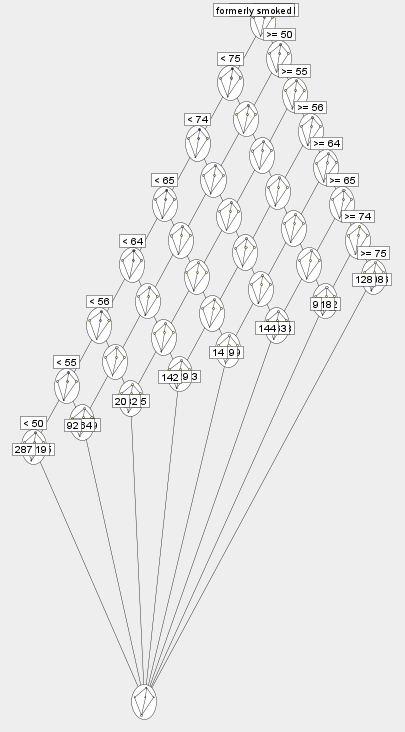


*Figure 14 – Nested diagram of smoke status and work type*



*Figure 15 - Concept lattice of the derived context represented in Table 8*

As we all know, both smoking and older age pose major risks in terms of the possibility of having a stroke. That's why we thought we'd see, through a nested diagram, how they influence each other, in order to draw more conclusions. From what we could see the most affected age range is 56 -65 years old, who are also smokers as it can be seen in Figure 14.

*Figure* *14 - Nested diagram of smoke status and average age interval*

Combining the scales SN2 and SO3 yielded the same information. The overweight categories (30-48 and 25-30) have the greatest proportion of patients with a lifetime risk of having a stroke since BMI evaluates the body fat in relation to height and weight. Figure 15 shows that experts say persons who work in private firms and have greater body fat levels are more likely to experience a stroke.

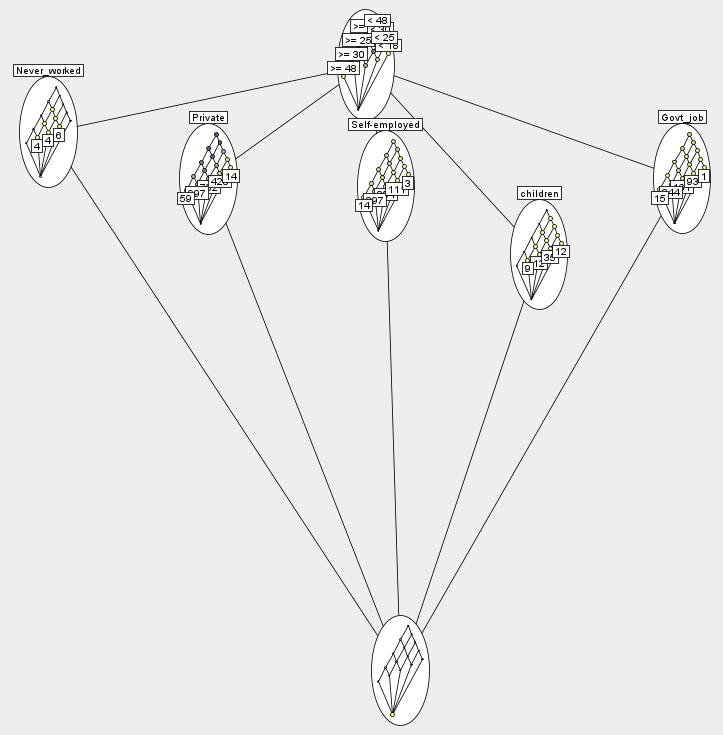
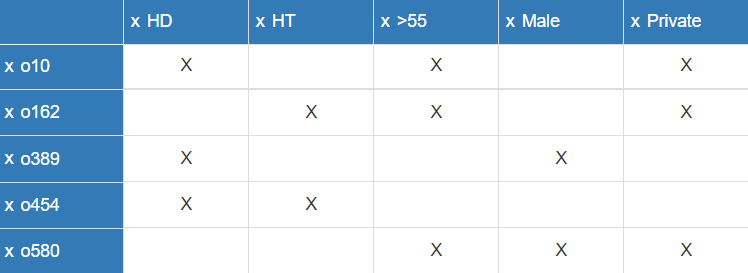


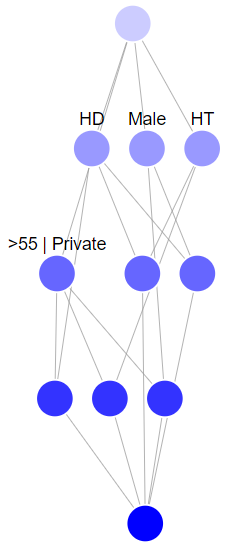
Figure 16 - N*ested diagram of SN2 and SO3*

### Attribute Exploration

Exploring attributes begins with a straightforward formal context and is frequently used for data collection. For our tests, we established the formal context K2 = (G2,M1,I) by extracting a sample from our dataset. This context is shown in Table 8. In this instance, G2 is a randomly selected subset of G's objects. By using FCA, we were able to generate 11 concepts, and Figure 17 shows the resulting concept lattice.



*Table 8 – Formal context for attribute exploration*



*Figure 17 – Concept lattice corresponding to Table 8*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A** | **L(A)** | **Ajj** | **Ajj = Aii**  **A -> Ajj** | **L** |
| ∅ | ∅ | ∅ |  |  |
| male | male | male | no |  |
| male, private | hypertension, … |  |  |  |
| >55 | private | private, >55 | yes | private -> >55 |
| >55, private | >55, private |  |  |  |
| >55, male | hypertension, … |  |  |  |
| hypertension | hypertension |  |  |  |
| hypertension, >55 |  | private, hypertension, >55 | no |  |
| hypertension, >55, male |  | private, hypertension, >55, male | no |  |
| heart\_disease | heart\_disease | heart\_disease |  |  |
| heart\_disease, hypertension |  | heart\_disease, hypertension |  |  |
| heart\_disease, >55 |  | private, heart\_disease, >55 | no |  |
| heart\_disease, male |  | heart\_disease, male |  |  |
| heart\_disease, private |  | >55, heart\_disease, private | yes | heart\_disease, private -> >55 |

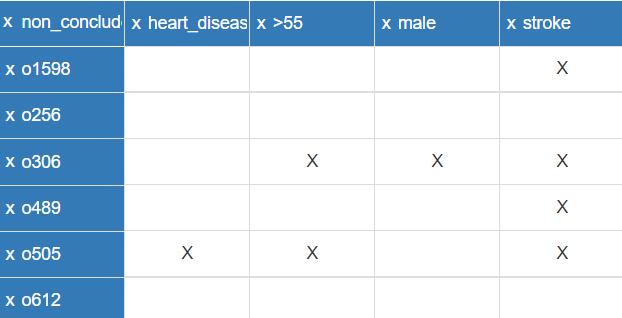
*Table 9 – Attribute Exploration*

Table 11 demonstrates the process of attribute discovery as well as the conclusions reached and verified by the experts. Analyzing the consequences, confirmed patients typically have a stroke condition because it is necessary for accurate and complete predictions. Additionally, the correlation between cardiac disease and hypertension typically points to a favorable outcome. Although age is not a definite characteristic of stroke, it is a crucial factor in determining the outcome. In reality, numerous additional variables are taken into account when deciding the final result, and it is frequently challenging to determine whether a set can accurately predict it. The experts assumed that best practices were followed when determining whether the implications are true or not.

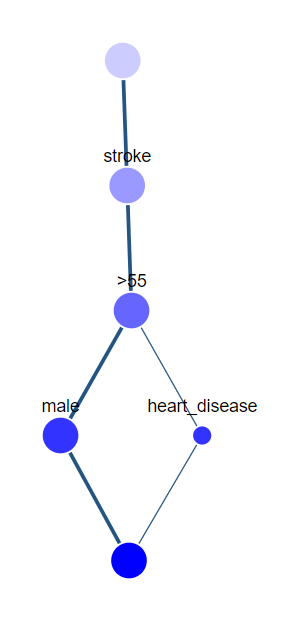
### Triadic Dataset

We developed a triadic framework using the available dataset to glean additional stroke information. We took the same group of study patients into consideration for the set of items (G). We identified a narrow set of factors M=heart disease (HD), hypertension (HT), age (>55), gender (male), and work type (private) that we employed in the majority of our prior research and for identifying significant information about the risks of stroke but we needed a set of circumstances in order to establish a triadic context.

To see if there are any associations between age and the workplace, we chose the heart disease condition for this experiment. We therefore took the set A = {positive, non concludent , and negative } into consideration. Table 10 provide samples from the resulting formal contexts. We accomplished local navigation by "locking" conditions to the detected concepts, and utilizing the concept lattices, we extracted and evaluated the data.



*Table 10 – Formal Context conditioned by “non concludent”*



*Figure 18 - Concept lattice with a locked "not conclusive" condition*