**CSI4142 – Introduction to Data Science**

**Final Project**

**Canadian Disaster Database**

**By**

**Mathieu Armstrong – 6799182**

**Francis Maher - 7315853**

**April 3rd 2018**

**Table of Contents**

[**List of Tables** 3](#_Toc510108990)

[**List of Figures** 4](#_Toc510108991)

[**Physical Design** 5](#_Toc510108992)

[**Data Staging** 7](#_Toc510108993)

[**OLAP Queries** 7](#_Toc510108994)

[Drill Down Query 7](#_Toc510108995)

[Roll up Query 7](#_Toc510108996)

[Slice Query 7](#_Toc510108997)

[Dice Query 7](#_Toc510108998)

[Iceberg Query 8](#_Toc510108999)

[**Business Intelligence Dashboard** 8](#_Toc510109000)

[**Machine Learning** 8](#_Toc510109001)

[Classification 8](#_Toc510109002)

[Cluster Analysis 9](#_Toc510109003)

# **List of Tables**

Table 1 - Physical model of the Date dimension 5

Table 2 - Physical model of the Summary dimension 6

Table 3 - Physical model of the Disaster dimension 6

Table 4 - Physical model of the Costs dimension 6

Table 5 - Physical model of the Location dimension 6

Table 6 - Physical model of the Fact table 7

# **List of Figures**

Figure 1 - Summary from the C4.5 algorithm using WEKA 8

Figure 2 - Confusion matrix from the C4.5 algorithm on WEKA 9

Figure 3 - K-means algorithm, first iteration (1/2) 9

Figure 4 - K-means algorithm, first iteration (2/2) 10

Figure 5 - K-means algorithm, second iteration (1/1) 10

Figure 6 - K-means algorithm, third iteration (1/2) 10

Figure 7 - K-means algorithm, third iteration (2/2) 10

Figure 8 - K-means algorithm, fourth iteration (1/2) 11

Figure 9 - K-means algorithm, fourth iteration (2/2) 11

Figure 10 - EM algorithm result 12

# **Physical Design**

The following tables describes the physical implementation of each dimensions found in the logical model. The physical design was implemented in PostgreSQL.

Table 1 - Physical model of the Date dimension

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date\_dimension** | **Data Type** | **Nulls?** | **PK** | **Comment** |
| date\_key | INT | NO | 1 |  |
| date\_actual | DATE | NO |  |  |
| epoch | BIGINT | NO |  | Seconds since UNIX Epoch time (negative if date is before Epoch) |
| day\_suffix | VARCHAR(4) | NO |  | MON for Monday… |
| day\_name | VARCHAR(9) | NO |  |  |
| day\_of\_week | INT | NO |  |  |
| day\_of\_month | INT | NO |  |  |
| day\_of\_quarter | INT | NO |  |  |
| day\_of\_year | INT | NO |  |  |
| week\_of\_month | INT | NO |  |  |
| week\_of\_year | INT | NO |  |  |
| week\_of\_year\_iso | CHAR(10) | NO |  | Week of the year in ISO format |
| month\_actual | INT | NO |  |  |
| month\_name | VARCHAR(9) | NO |  |  |
| month\_name\_abbreviated | CHAR(3) | NO |  |  |
| quarter\_actual | INT | NO |  |  |
| quarter\_name | VARCHAR(9) | NO |  |  |
| year\_actual | INT | NO |  |  |
| first\_day\_of\_the\_week | DATE | NO |  |  |
| last\_day\_of\_the\_week | DATE | NO |  |  |
| First\_day\_of\_the\_month | DATE | NO |  |  |
| Last\_day\_of\_the\_month | DATE | NO |  |  |
| First\_day\_of\_the\_quarter | DATE | NO |  |  |
| last\_day\_of\_the\_quarter | DATE | NO |  |  |
| first\_day\_of\_the\_year | DATE | NO |  |  |
| last\_day\_of\_the\_year | DATE | NO |  |  |
| mmyyyy | CHAR(6) | NO |  |  |
| mmddyyy | CHAR(10) | NO |  |  |
| weekend\_indr | BOOLEAN | NO |  |  |
| is\_holiday | BOOLEAN | NO |  | Default value is FALSE |
| holiday\_text | VARCHAR(50) | NO |  | Name of the holiday |
| meteorological\_season | VARCHAR(10) | NO |  |  |

Table 2 - Physical model of the Summary dimension

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Summary\_dimension** | **Data Type** | **Nulls?** | **PK** | **Comment** |
| summary\_key | SERIAL | NO | 1 |  |
| summary | TEXT | NO |  |  |
| keyword\_1 | VARCHAR(20) |  |  |  |
| keyword\_2 | VARCHAR(20) |  |  |  |
| keyword\_3 | VARCHAR(20) |  |  |  |

Table 3 - Physical model of the Disaster dimension

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Disaster\_dimension** | **Data Type** | **Nulls?** | **PK** | **Comment** |
| disaster\_key | SERIAL | NO | 1 |  |
| disaster\_type | VARCHAR(30) | NO |  |  |
| disaster\_subgroup | VARCHAR(30) | NO |  |  |
| disaster\_group | VARCHAR(30) | NO |  |  |
| disaster\_category | VARCHAR(30) | NO |  |  |
| magnitude | DECIMAL(18, 1) |  |  | Only relevant if disaster is an earthquake |
| utility\_people\_affected | INT |  |  |  |

Table 4 - Physical model of the Costs dimension

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cost\_dimension** | **Data Type** | **Nulls?** | **PK** | **Comment** |
| cost\_key | SERIAL | NO | 1 |  |
| estimated\_total\_cost | INT | NO |  |  |
| normalized\_total\_cost | INT | NO |  |  |
| federal\_dfaa\_payments | INT | NO |  |  |
| provincial\_dfaa\_payments | INT | NO |  |  |
| provincial\_department\_payments | INT | NO |  |  |
| municipal\_cost | INT | NO |  |  |
| ogd\_cost | INT | NO |  |  |
| insurance\_payments | INT | NO |  |  |
| ngo\_cost | INT | NO |  |  |

Table 5 - Physical model of the Location dimension

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Location\_dimension** | **Data Type** | **Nulls?** | **PK** | **Comment** |
| location\_key | SERIAL | NO | 1 |  |
| city | VARCHAR(190) | NO |  |  |
| province | VARCHAR(50) | NO |  |  |
| country | VARCHAR(30) | NO |  |  |
| canada | BOOLEAN | NO |  | True if the disaster was in Canada. |

Table 6 - Physical model of the Fact table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fact** | **Data Type** | **Nulls?** | **PK** | **Comment** |
| start\_date\_key | INT | NO | 1 | Links to a row in the Date\_dimension |
| end\_date\_key | INT | NO | 1 | Links to a row in the Date\_dimension |
| location\_key | INT | NO | 1 |  |
| disaster\_key | INT | NO | 1 |  |
| summary\_key | INT | NO | 1 |  |
| cost\_key | INT | NO | 1 |  |
| fatality\_number | NUMERIC |  |  |  |
| injured\_number | NUMERIC |  |  |  |
| evacuated\_number | NUMERIC |  |  |  |

# **Data Staging**

# **OLAP Queries**

The following five queries demonstrate how it’s possible to traverse concept hierarchies in OLAP databases.

## Drill Down Query

## Roll up Query

Determine the trends in fatalities over the last 100 years.

## Slice Query

Contrast the number of fatalities in Ontario due to wildfires during 1999 with the number of fatalities in Quebec due to wildfires during 1999.

## Dice Query

Determine the number of fatalities per province during 2000.

## Iceberg Query

Determine the 5 cities in Canada with the most riots.

# **Business Intelligence Dashboard**

# **Machine Learning**

## Classification

By using the classification algorithm C4.5 (or J48 in WEKA), we were interested to see if was possible to determine the Canadian province where a disaster occurred based on the following criteria:

* Disaster type
* Disaster group
* Disaster subgroup
* The 3 keywords
* Any associated cost with the disaster
* The amount of people affected (injuries, fatalities, etc)

We were able to retrieve the following result using the 10 folds cross-validation test option.

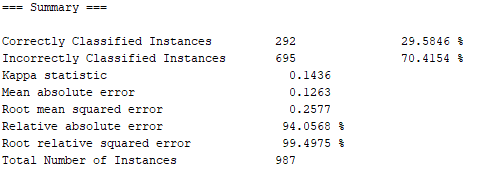


Figure 1 - Summary from the C4.5 algorithm using WEKA

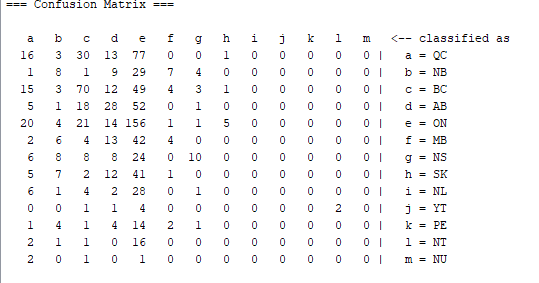


Figure 2 - Confusion matrix from the C4.5 algorithm on WEKA

Overall, we can see that the result is quite poor. In fact, the accuracy of the model is only of 29.58%. Just 292 of the 987 items were correctly classified. By simply looking at the confusion matrix, it’s easy to see that most of the items deviate from the main diagonal. This shows that the data we have to build the model is of low quality, the criteria chosen for this test were wrong or that we simply do not have enough data to find a correlation between location of a disaster and the criteria described earlier.

## Cluster Analysis

By doing a cluster analysis, we were interested to see if there were any similarities between the meteorological seasons and the type of disaster occurring across Canada since the beginning of the 20th century. We were expecting some of the results to be obvious, like having winter storms mostly in the winter season, but there might be something else hidden in the data.

We first used the K-means algorithm with k=4 hoping that we would find some relevant results. After 4 iterations using a different seed every time, we received the following results:

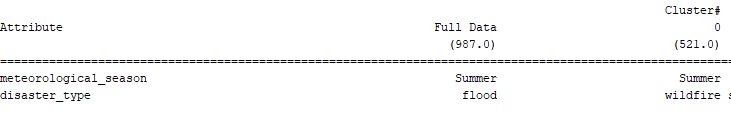


Figure 3 - K-means algorithm, first iteration (1/2)

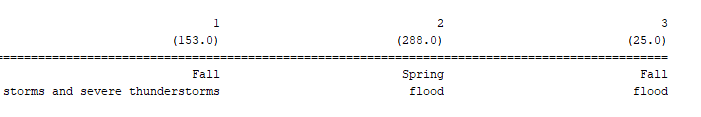


Figure 4 - K-means algorithm, first iteration (2/2)

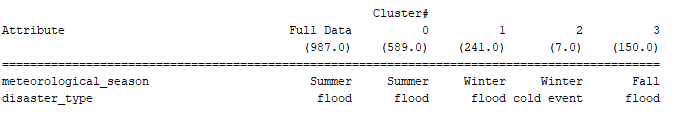


Figure 5 - K-means algorithm, second iteration (1/1)

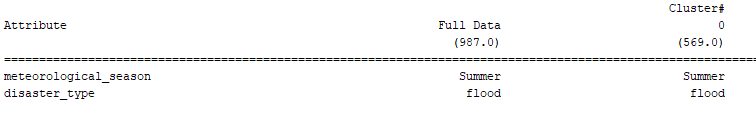


Figure 6 - K-means algorithm, third iteration (1/2)

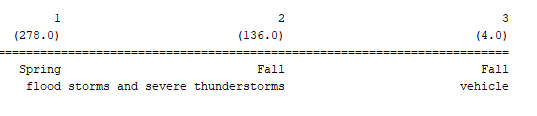


Figure 7 - K-means algorithm, third iteration (2/2)

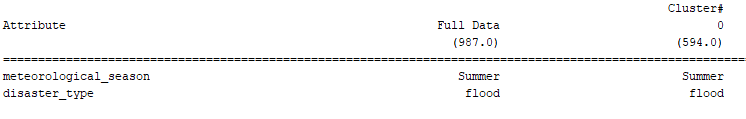


Figure 8 - K-means algorithm, fourth iteration (1/2)

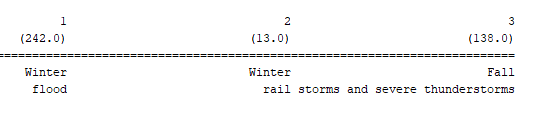


Figure 9 - K-means algorithm, fourth iteration (2/2)

Since the k-mean algorithm uses a random starting point every time it runs, the final result might be different. The algorithm runs until the two last iterations are the same. This means that there is a possibility that it finds a local optimal cluster or cluster groups instead of the global optimal cluster groups. This is exactly what happens in this scenario. Unless we run this algorithm with a different seed multiple times and averaging the results, we might find a relevant result but this is time consuming. Instead, using an Expectation-Maximization algorithm (EM) would solve this problem since the starting is not chosen randomly in this case. The following algorithm in WEKA provided this result:

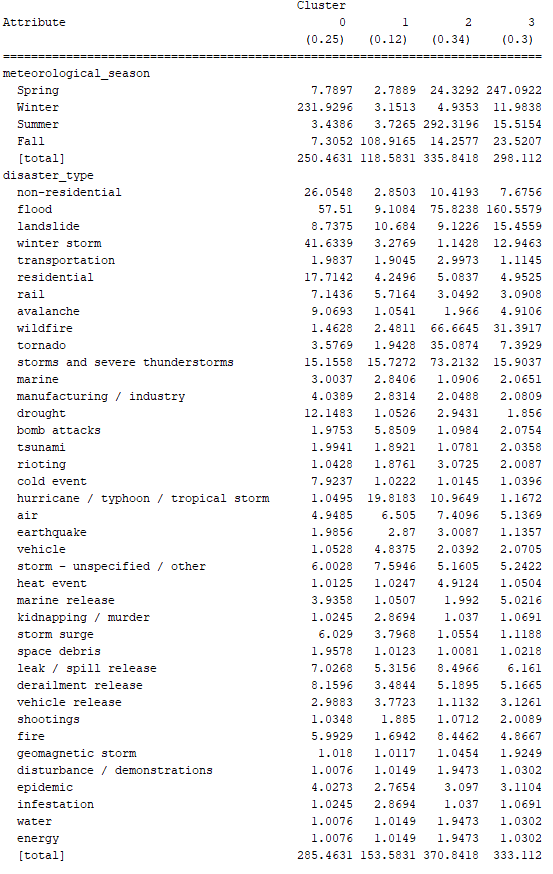


Figure 10 - EM algorithm result

We can clearly see that there are 4 major clusters, one for each season of the year. As predicted earlier, the highest occurrence of a winter storm disaster is within the cluster with the highest amount of disaster occurring in winter, cluster 0. Although an expected result, this shows that this model seems to make sense.

When looking at natural disasters, this model and dataset doesn’t seem to display anything of note. We see a similar result when looking at non-natural disasters: there doesn’t seem to be any direct correlation between a meteorological season and the type of disaster occurring anywhere in Canada.