

Deploying and Performing Business Intelligence with MS SQL Server using AWS Cloud

Case Study: National Highway Traffic Administration (NHTSA)

Abayomi Fashina

Department of Computer Science and Quantitative Methods, Austin Peay State University,
Clarksville, TN 37044, USA

Email: afashina@my.apsu.edu

ABSTRACT

This project takes into essence, the importance and benefits of deploying Business Intelligence (BI) on Cloud. Cloud computing is the availability of computer system resources through data storage and computing power accessible via the internet.

Performing and deploying BI using SQL Server 2012 on Amazon Web Service (AWS), I created AWS EC2 instance with Windows Server 2012 R2 Operating System. BI on MS platform will be conducted using different tools such as SQL Server Integration Service (SSIS), SQL Server Analysis Service (SSAS) and SQL Server Reporting Service (SSRS) to perform Data Integration, Data Analysis and Reporting Services. In conclusion I applied Data Mining predictive model such as Decision Tree, Naïve Bayes, Clustering and Neural Network to my case study – Fatality Reporting Systems, which I downloaded from National Highway Traffic Administration (NHTSA) to predict what factor influence the rate of fatality on our roads in the United States. This dataset only considers 2018 records of accident.

I then looked at the cost benefits of deploying BI on the cloud over deploying it on a physical infrastructure for any organization. This report is divided into sections, section 1 talks about basic introduction to Data, Information, BI and BI Lifecycle. Section 2 discusses similar works in retrospect to Business Intelligence on Cloud. Section 3 describes the Extract, Transform and Load (ETL) process which is done by SSIS. I will also carry out SSAS on the data in this section. Section 4 handles the Data Mining predictive model using Decision Tree, Naïve Bayes, Clustering and Neural Network Algorithms. Section 5 is the concluding part which talks about results and conclusion.

KEYWORDS

Data, Information, Data Mining, SSIS, SSAS, Decision Tree, MS SQL Server 2012, Windows Server 2012 R2, Business Intelligence, Amazon Web Services (AWS)

1. INTRODUCTION

Business intelligence is divided into two components; business and intelligence. Business Intelligence in simple terms means converting data into information for a non-technical user to understand. BI can also be defined as a process of transforming data into a meaningful information so that the end-user or stakeholders can look at the information and make proper decision or carry out forecasting if need be.

Business involves organized activities put together by an individual or a group of individuals to produce and sell goods and services for profit. These business activities generate data from customers, purchases, orders, suppliers, inventory and so on. All these data can be analyzed and mined using special techniques and tools to generate patterns and intelligence to indicate how the organization is performing.

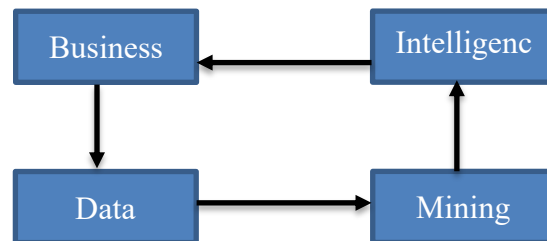


Figure 1: Business Intelligence and Data Mining Cycle (Business Intelligence and Data Mining by Anil Maheshwari)¹

Business Intelligence (BI) is a broad set of information technology (IT) that includes tools for gathering, analyzing and reporting information to users about the performance of the organization and its environment (Business Intelligence and Data Mining by Anil Maheshwari)¹.

1.1 SOFTWARE

1. Microsoft SQL Server 2012 Management Studio.
2. MS SQL Data Tool (MS Visual Studio)
3. SQL
4. SQL Server Integrating Services (SSIS)
5. SQL Server Analysis Services (SSAS)

1.2 Major Elements of Business Intelligence Framework

The followings are the major element of Business Intelligence Framework:

Data: Data, they say, is the new “oil”. In this statement is the discovery of hidden value in data. Data is the heartbeat of any business, it can be collected and stored in a database. Data is observations and facts, in fact, anything that is recorded is data.

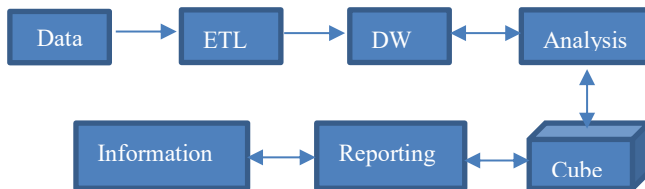


Figure 2: Business Intelligence Framework

Database: Is the modeled collection of data that is accessible in numerous ways within or outside an organization. Database Management Software Systems (DBMSs) are available today to aid storage, management, and manipulation of data (e.g MS SQL Server).

Data Warehousing: This can be referred to as the collection of data marts from all divisions of an organization. It can also be defined as the repositories where historical data of the organization can be intentionally designed to help in management decisions.

Data Mining: Is the process of finding valuable and imaginative designs from data. Data mining is all about looking for unexpected and unknown relationships amongst the data. It is a multi-disciplinary skill that uses statistics, database technology and machine learning.

Data Visualization: These techniques used to represent data graphically. In this project, I will be using Tableau to design dashboard which will serve as reports to the end-users or stakeholders of the business.

1.3 Cloud Concept

Cloud computing is the delivery of computing services including Servers, Storage, Databases, networking, Software, Analytics and Intelligence over the internet (cloud) to offer faster innovation, flexible resources, and economic scale.

1.3.1 Benefits of Cloud Computer

1. **Reduces:**
 - a. Hardware Cost
 - b. Operational Cost
 - c. Deployment Cost
2. **Increases:**
 - a. Resiliency
 - b. Performance
 - c. Capacity

1.3.2 Cloud Computing Models

1. Full Cloud Deployment: All components in the cloud
 - a. Databases
 - b. Processing
 - c. Storage
 - d. Application Logic
 - e. Nothing on premise
2. Hybrid Deployment: Some resources are internal; others are on the cloud.
 - a. Processing
 - b. Databases
 - c. Storage
 - d. Application Logic

1.3.3 Cloud Terminologies

1. Infrastructure as a Service (IaaS) (Kern et al. 2002)²:
 - a. Entire infrastructure on the cloud
 - b. Platforms and software run on other infrastructures
 - c. You must manage it all
2. Platform as a Service (PaaS).
 - a. You don't manage the infrastructure
 - b. Applications are deployed on the platform
3. Software as a Service (SaaS) (Benlian et al. 2009)³:
 - a. Someone else develops the software
 - b. You use the software on the cloud

1.3.4 AWS EC2

Amazon Elastic Compute Cloud (EC2)⁴ is a web service that enables secure pay as you use, resizable compute capacity on the cloud. An EC2 instance is a virtual server for running applications on AWS infrastructure⁴. You can use Amazon EC2 to launch as many virtual servers as you need, configure security, network and manage storage. It enables you to scale up or down using a feature called “Auto Scaling” which handles changes in requirements, reducing your need to forecast traffic.

1.3.5 Features of Amazon EC2⁴

Amazon EC2 provides the following features:

1. Virtual computing environments, known as instances

2. Preconfigured templates for your instances, known as Amazon Machine Images (AMIs), that package the bits you need for your server (including the operating system and additional software)
3. Various configurations of CPU, memory, storage, and networking capacity for your instances, known as instance types
4. Secure login information for your instances using key pairs (AWS stores the public key, and you store the private key in a secure place)
5. Storage volumes for temporary data that's deleted when you stop or terminate your instance, known as instance store volumes
6. Persistent storage volumes for your data using Amazon Elastic Block Store (Amazon EBS), known as Amazon EBS volumes
7. Multiple physical locations for your resources, such as instances and Amazon EBS volumes, known as Regions and Availability Zones
8. A firewall that enables you to specify the protocols, ports, and source IP ranges that can reach your instances using security groups
9. Static IPv4 addresses for dynamic cloud computing, known as Elastic IP addresses
10. Metadata, known as tags, that you can create and assign to your Amazon EC2 resources
11. Virtual networks you can create that are logically isolated from the rest of the AWS cloud, and that you can optionally connect to your own network, known as virtual private clouds (VPCs)

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status
Linux-Amazon-AMI	i-0b1150b47f0d4d48	t2.micro	us-east-1a	running	2/2 checks	None
Windows Server 2012	i-0b1150b47f0d4d48	t2.micro	us-east-1a	running	2/2 checks	None

Instance: i-0b1150b47f0d4d48 (Windows Server 2012)		Public DNS: ec2-54-196-230-132.compute-1.amazonaws.com
Description	Instance ID: i-0b1150b47f0d4d48	Public DNS (IPv4): ec2-54-196-230-132.compute-1.amazonaws.com
	Instance state: running	Private IP: 10.0.0.102
	Instance type: t2.micro	Public IP: -
	Platform: Amazon Linux AMI	Elastic IP: -
	Provisioning: Provisioned	Availability zone: us-east-1a
	Private DNS: ip-172-31-40-17.ec2.internal	Security groups: sg-7a51a200
	Private IP: 172.31.40.17	Subnet: subnet-4234a2d3
Secondary private IP	VPC ID: vpc-b7f7d3d1	Subnet: subnet-4234a2d3
Network interface	Network interface: eni0	Subnet: subnet-4234a2d3
Source/destination	Source/destination: check	Subnet: subnet-4234a2d3
Pay per name	Pay per name: Windows Server 2012	Subnet: subnet-4234a2d3

Figure 3: AWS EC2 Instance of Windows Server 2012 R2 64 Bit-Base

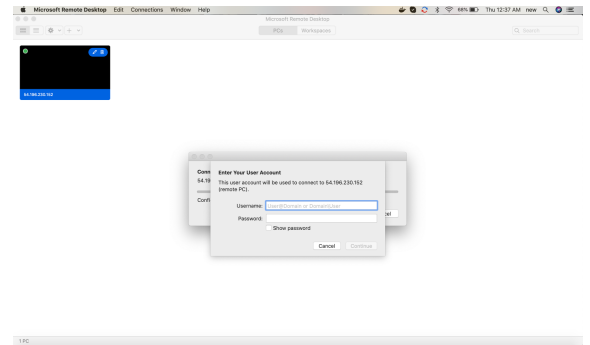


Figure 4: Microsoft Remote Desktop to connect to the instance on my physical laptop.



Figure 5: Windows Server 2012 R2 Operating System running on my Laptop

2. RELATED WORKS

According to Olszak C. M[5], the idea of cloud computing has been explored for several years, unfortunately the investigations on Business Intelligence on cloud are only partly addressed by existing research. Some other authors also claimed that Business Intelligence cloud is not suited for many organizations, especially in BI (Cloud Security Alliance)[6]. Olszak concluded by saying that the advantages of BI Cloud overshadow the disadvantages due to security threat of most data centers faced. It states clearly the efficiency and productivity of BI and increases the performance of BI Software. Olszak also emphasized that cloud shortens BI implementations and reduces the cost of BI applications.

Lacity et al [7] suggested that before moving to Cloud Computing, it is recommended to consult with a large and mature body of knowledge on IT outsourcing. In this project, I would be implementing Microsoft SQL Server 2012 and Microsoft Data Tool to integrate and analyze my data. With this tool, I will be building a data mining predictive model with Decision Tree, Clustering, Naïve Bayes and Neural Network to predict factors affecting the rate of accident in US.

Baars and Kemper's[9] paper discussed extensively on the three layers of Business Intelligence, which are: (1) Data Layer – This layer discusses the way structured and unstructured data are stored in the Data Warehouse, (2) Logic Layer – emphasizes how the data is being analyzed using different tools available in the marketplace,

and (3) Access Layer – users roles and privileges are created to meet the standard of SaaS. These layers are interconnected for the effectual development of a BI systems. They went further to explained Cloud BI Framework for delivering different cloud scenarios. Based on their discovery, users can implement BI according to their business needs.

The conclusion of the recent Market Study of 859 respondents [10], shows that there is a surge of investment in cloud based BI and developed interest in cloud's benefits among many organizations. While Gartner survey throws the fact that almost one-third of the BI platform users surveyed (27 percent, to be exact) are using or planning to use the cloud / SaaS model to expand their business intelligence capabilities in the next 12 months [11]. These facts simply imply that cloud based BI implementation is on the rise among organizations.

3. ETL PROCESS AND DATA ANALYSIS

3.1 Extract, Transform and Load Process

The dataset for this project is gotten from National Highway Traffic Safety Administration (NHTSA)⁸. The data is stored in its raw format (in CSV). Before this data can be moved into the relational database, I have to perform Extract, Transform and Load (ETL) operation on the data. Therefore, data cleaning is necessary.

3.1.1.1. Data Cleaning: Data cleaning in data mining is the technique of detecting and removing unwanted records from the data. The original data from Fatality Analysis Reporting System (FARS) is in (.CSV file extension) which are: **ACCIDENT, VEHICLE, DRUGS, PERSON, ACC_AUX, DAMAGE** and **VINCODE**. The data contains over 51,873 recorded fatal accidents in the United States in 2018 from different states and counties. I have to remove all unnecessary database from the CSV files.

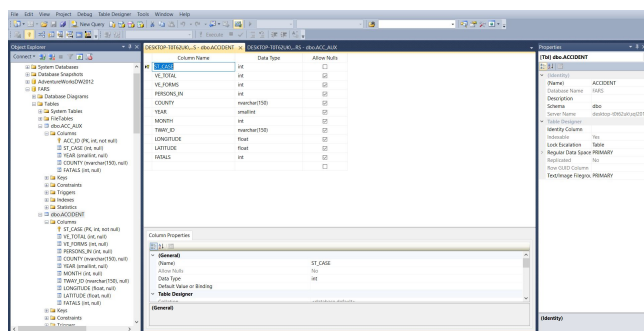


Figure 6: Microsoft SQL Server 2012 Management Studio

3.1.2. Data Integration: In this stage, after proper cleaning of the data, integration is carried out using SSIS tool. Data Integration is a process in which heterogenous (structured or unstructured) data is retrieved and combined as an incorporated form and structure. See more diagram below where I performed Integration on the data from CSV flat files to MS SQL Server Database as my destination.

3.1.3. Transformation: Converting the data format from that of the source system to that of the Data Warehouse.

3.1.4. Loading:

Loading the transformed data into the destination Data Warehouse and creating all the needed index for this data.

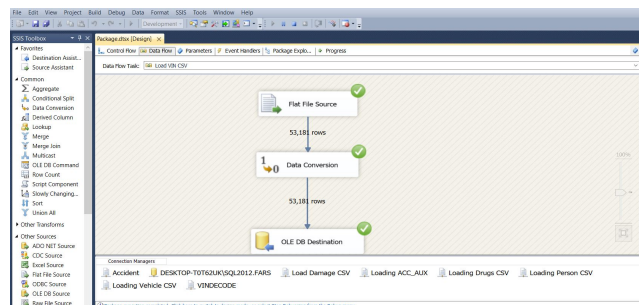


Figure 7: SQL Server Integration Service using MS SQL Data Tools

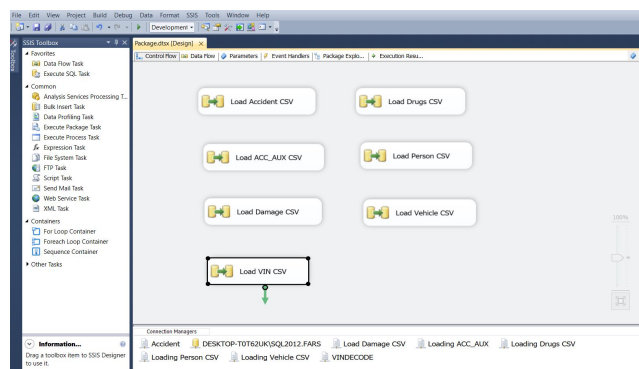


Figure 8: Showing all the Data Integration to load data into FARS database in SOL Server

3.2 Data Analysis using MS SQL Data Tool

Here, I designed the data mart by building cube, measures and dimensions to analyze the data further. In the previous section I created a Fact table called FatalityFact.

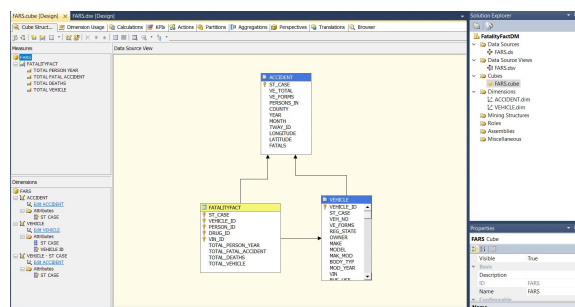


Figure 9: FARS Cube

3.2.1 Cube: According to the diagram above, the cube wizard helps to design my measure groups and dimensions for the cube. The steps are so simple to follow. Let us look at the basic definition. A cube is a multidimensional structure that form the information needed for analysis. The cube contains the following components:

Dimensions, Measures and Measure Groups, Partitions, Perspectives, Hierarchies, Actions, Key Performance Indicators (KPI), Calculations and Translations. For my analysis, I only made use of Dimensions, Measures and Measure Groups in this project.

3.2.2 Measures: It is an aggregated value of a numerical data such as sum, minimum, maximum, average, count or custom expression using MDX.

3.2.3 Measure Group: A measure group is a collection of one or more measures.

3.2.4 Dimensions: A dimension is the putting together of related objects, called attributes. The dimension provides information about fact data in a cube.

4.0 DATA MINING PREDICTIVE MODEL

In this section, I performed prediction by building classification model which made use of these algorithms (Decision Trees, Naïve Bayes, Clustering and Neural Network).

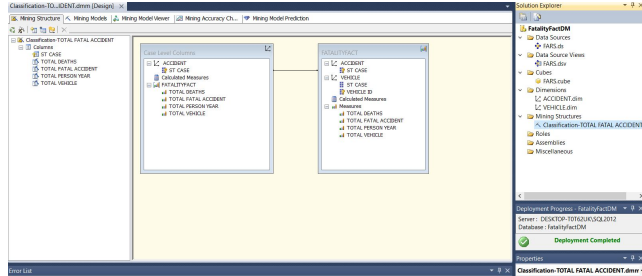


Figure 10: Showing Classification Model

I used the following variables in my prediction: TOTAL DEATHS, TOTAL PERSON YEAR, TOTAL FATAL ACCIDENT AND TOTAL VEHICLE. TOTAL FATAL ACCIDENT is my target prediction, that is what I am trying to predict.

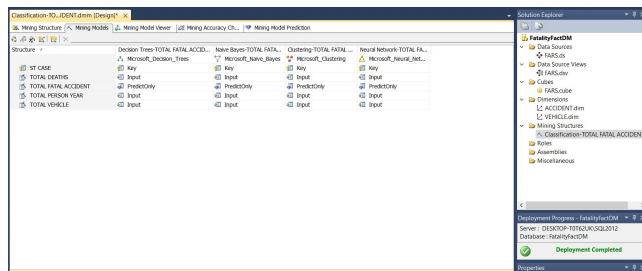


Figure 11: Showing the list of the Algorithms in building the model

4.1 Decision Trees: Before I begin to build the model based on different algorithms, I set the testing data to 30% and the training data to 70%.

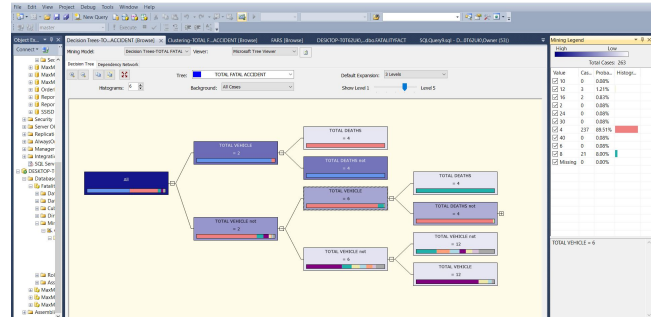


Figure 12: Decision Trees algorithm given 89.5% accuracy

4.2 Naïve Bayes: Naive Bayes algorithm is a classification algorithm based on Bayes' theorems, and can be used for both exploratory and predictive modeling. Naïve Bayes is not heavily computational algorithm.

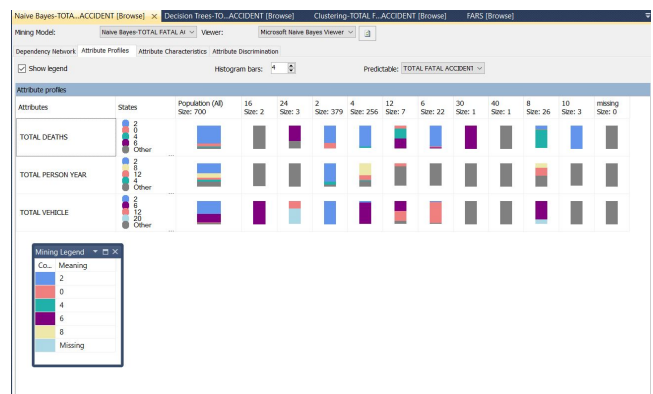


Figure 13: Showing the Attributes Profiles of Naïve Bayes

In the diagram above, Bayes algorithm is showing how different state of each variables are distributed and the value of each predictions.

4.3. Clustering: Clustering model looks at patterns in the dataset we might not see through normal observation. Relationship between these variables will be determined though the cluster as showed in the diagram below:

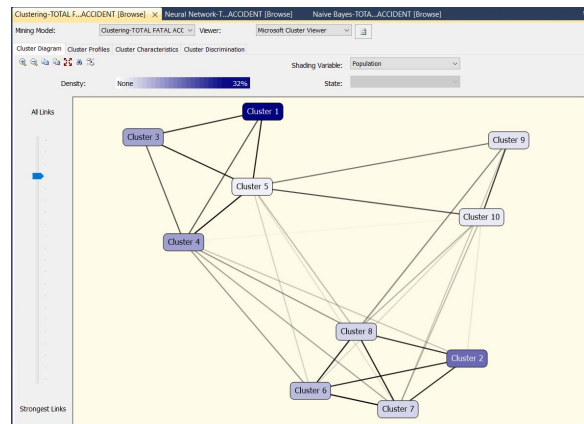


Figure 14: Showing the Attributes in a Cluster diagram

4.4. Neural Network: Works by testing each different state of the input attributes (TOTAL DEATHS, TOTAL PERSON AND TOTAL VEHICLE) against the predictable variable and calculating probabilities for each combination based on the training data. These probabilities are then used to predict the TOTAL FATAL ACCIDENT.

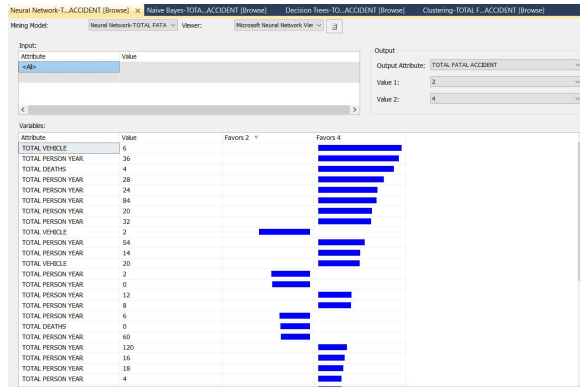


Figure 15: Neural Network model generate 87.4% accuracy prediction

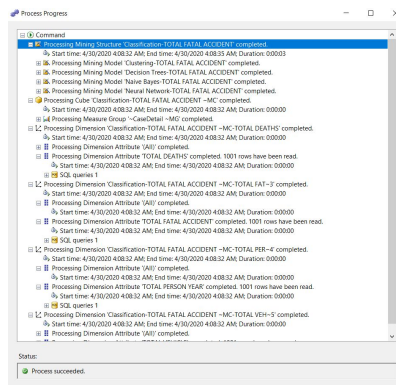


Figure 16: Showing the completed model generation

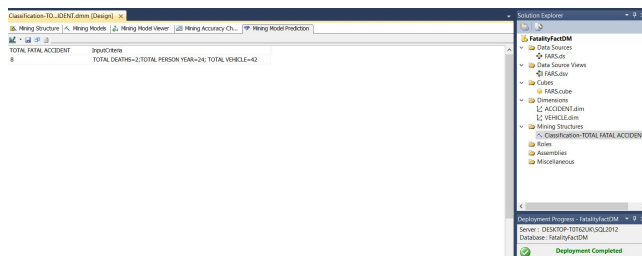


Figure 17: Showing results of prediction of the target variable (TOTAL FATAL ACCIDENT).

5.0 RESULTS AND CONCLUSION

5.1 Results

According to the data mining prediction model, I divided the dataset into two sets, 30% for testing and 70% for training the data. After a successfully deployment according to the diagram below:

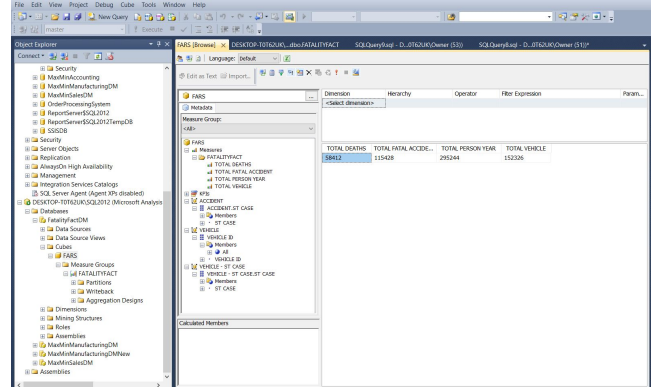


Figure 18: Showing the deployed model.

Decision Trees algorithm gave me the best accuracy (89.5%). Other algorithms performed well too, but not as Decision Tree in this project which resulted in predicting the TOTAL FATAL ACCIDENT in 2018 from different states in the United States. I will recommend for future works on this project, that a large dataset be used so that the data mining model can have enough data to train the model and give more attributes to predict.

5.2 Conclusion

In summary, I deployed the model in SQL Server Analysis Service and built a classification model with four algorithms (Decision Trees, Clustering, Naïve Bayes and Neural Network). I also discovered from one of my KPI (percentage of fatal accident affected by drugs). This indicated that higher percentages of accidents in the United States are caused as a result of one drug or the other.

Finally, the deploying of this project was done on Windows Server 2012 R2 Operating System (MS SQL Server 2012 was installed) on AWS EC2 instance.

The implication of the project shows that fatality in the United States are caused majorly by the influence of drug or alcohol on the driver.

I will recommend for future works on this project, that a large dataset should be used so that the data mining model can have enough data to train the model and give more attributes to predict.

REFERENCES

- [1] Anil K. Waheshwari, PhD (2014). Business Intelligence and Data Mining Published by BEP Business Expert Press, Pg 3 - 9.
- [2] Kern, T., Lacity, M.C. and Willcocks, M.P. (2002). Netsourcing: renting business applications and services over a network. Prentice Hall, Upper Saddle River (NJ, USA).
- [3] Benlian, A., Hess, T. and Buxmann, P. (2009). Drivers of SaaS-Adoption – An Empirical Study of Different Application Types. Business & Information Systems Engineering, 1 (5), 357-369.

[4] Amazon Documentation, EC2, Instance and Benefits.
<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html>

[5] Olszak C.M (2014). Business Intelligence in Cloud

[6] Gurjar and Rathore (2013), Hayes (2008), Ouf and Nasr (2011); Thomason and Van Der Walt (2010). Cloud Security Alliance (2010)

[7] Lacity, M.C., Khan, S.A. and Willcocks, L.P. (2009). A review of the IT outsourcing literature: Insights for practice. Journal of Strategic Information Systems, 18, 130-146.

[8] Dataset (2018), National Highway Safety Administration (Fatality Reporting System). <https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars>

[9] Baars, H. and Kemper, H.G (2008). Management Support with Structured and Unstructured Data – An Integrated Business Intelligence Framework. Information Systems Management, 25(2), 132 – 148.

[10] Wisdom of Crowds Cloud Business Intelligence (2012). Article: <http://www.articlesbase.com/software-articles/key-drivers-for-business-intelligence-making-the-move-to-the-cloud-6354376.html>

[11] <http://www.gartner.com/newsroom/id/1903814>