Faculty of Informatics Engineering

Department of Software & Information Systems Engineering

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Building a machine learning model for the WHO GATS data and analyzing it

A senior project report submitted to obtain a Bachelor's degree in Software and Information System Engineering for the undergraduate students

Prepared by:

Enas Al-Baghdadi

Lujain Shebli

Supervised by:

Dr. Mouhib Al-noukari

Eng. Majd Al-ssabbagh

Academic year

Second semester 2022/2023

Supervisor certification:

I certify that the preparation of this project entitled “Building a machine learning model for the WHO GATS data and analyzing it - Phase III”, prepared by Enas AL Baghdadi and Lujain Shblie, was made under my supervision at Department of Software and Information System Engineering / Faculty of Computer and Informatics Engineering in fulfillment of the requirements of the Degree of Bachelors of Software and Information System Engineering

Name……..…. Signature……... Date:

Abstract:

Decision-making and supporting the knowledge process in decision-making is one of the most important needs of the data age, given that the World Health Organization is at the forefront of improving global health and equity in the field of health, and achieving the organization's goal of ensuring that all people enjoy the highest possible standard of health, and given that tobacco use is among the most worrying health issues that pose ongoing risks to health, and the importance of health-supporting policies and practices based on the best scientific knowledge.

We have issued a set of reports that help understand the relationship between environmental, social and behavioral characteries for smokers, and we classified smokers into three categories (a current smoker who tried to quit smoking and failed to quit - a current smoker who did not try to quit smoking - a smoker who tried to quit smoking and succeeded in quitting), and we applied many data mining and machine learning techniques to identify the factors that most affect smokers on the data of adult smokers within the surveys issued by the World Health Organization for a country India , and predicting the result of quitting smoking on the data of adult smokers within the swabs issued by the World Health Organization for all countries of the world.

The aim is to assist WHO in decision-making and design evidence-based prevention and awareness programs to improve public health and promote health sustainability in the community.

Our motivation for choosing to work on big data comes from our awareness of the availability of a large volume of data that certainly hides useful knowledge, and tobacco control tools can benefit greatly from applying machine learning techniques on it.

ملخص:

إن صناعة القرار ودعم عملية المعرفة في اتخاذ القرار تعد من أهم احتياجات عصر البيانات، و نظراً لاحتلال منظمة الصحة العالمية مكان الصدارة في سبيل تحسين الصحة في العالم والإنصاف في مجال الصحة وبلوغ هدف المنظمة وهو أن تتمتع جميع الشعوب بأعلى مستوى ممكن من الصحة ونظراً لأن استخدام التبغ من بين أكثر القضايا الصحية إثارة للقلق والتي تشكل مخاطر مستمرة على الصحة و لأهمية ان تستند السياسات والممارسات الداعمة للصحة الى افضل المعارف العلمية.

قمنا بإصدار مجموعة من التقارير التي تساعد على فهم العلاقة بين العوامل البيئية والاجتماعية والسلوكية للمدخنين, وقمنا بتصنيف المدخنين الى ثلاث فئات (مدخن حالي حاول الاقلاع عن التدخين وفشل في الإقلاع – مدخن حالي ولم يحاول الإقلاع عن التدخين - مدخن حاول الإقلاع عن التدخين ونجح في الإقلاع)، وطبقنا العديد من تقنيات التنقيب في البيانات والتعلم الآلي للتعرف على العوامل الأكثر تأثيراً على المدخنين وذلك على بيانات المدخنين البالغين ضمن المسحات الصادرة عن منظمة الصحة العالمية لدولة الهند ، والتنبؤ بنتيجة الإقلاع عن التدخين وذلك على بيانات المدخنين البالغين ضمن المسحات الصادرة عن منظمة الصحة العالمية لجميع دول العالم.

والهدف هو مساعدة منظمة الصحة العالمية في اتخاذ القرارات وتصميم برامج الوقاية والتوعية المستندة إلى الأدلة العلمية لتحسين الصحة العامة وتعزيز الاستدامة الصحية في المجتمع.

يأتي دافعنا لاختيار العمل على مثل هذه المجموعة الكبيرة من البيانات من إدراكنا لتوفر حجم كبير من البيانات التي تخفي بالتأكيد المعرفة المفيدة ، ويمكن أن تستفيد أدوات مكافحة التدخين بشكل كبير من تطبيق تقنيات تعلم الآلة.

Dedication

I dedicate this graduation project to all the people who have been part of my journey. To my professors and mentors, for their guidance and support throughout my years in college; to my family, my friends, and my classmates, for their encouragement, friendship, and understanding; and to myself, for persisting through the toughest of times and never giving up.

This project is a culmination of years of hard work, dedication, and determination to reach the finish line. It is a testament to the importance of setting lofty goals and working hard to achieve them, regardless of the obstacles. It is an acknowledgement of the journey I have taken to get to where I am today, and a reminder that hard work pays off in the end.

This project is dedicated to those who have believed in me and guided me along the way. It is dedicated to the times I have pushed

myself to the limit and gone beyond what I thought was possible. It is dedicated to the moments of success and the moments of failure, for they have taught me more than any victory ever could. It is a celebration of my accomplishments, and a reminder to never give up on my dreams. I am so grateful for the support and encouragement I have received throughout my journey, and I thank you all from the bottom of my heart.

Acknowledgments

I would like to express my sincere gratitude to my academic mentor *Prof. Mouhib Alnoukari* and *Eng. majd Alssabbagh* for their invaluable guidance, encouragement and support throughout this project.

I would also like to thank all the faculty members of *Faculty of Informatics Engineering* in *Syrian Private University*, who have provided me with timely advice and support in the development of this project.

I am also thankful to my friends and family for offering assistance and encouragement whenever I needed it most.

I would like to thank my classmates for their unwavering support throughout the project.

Last but not least, I would like to thank all those anonymous people who have offered valuable feedback and insight into the project. Their assistance has made this project possible.

Once again, I would like to express my deepest appreciation and gratitude towards all the people mentioned above for their support and assistance throughout the project.

Contents

Supervisor certification: 2

Abstract: 3

ملخص: 4

Dedication: 5

Aknowledgment: 7

Table of figures: 11

List of Tables: 13

Chapter 1 Theoretical study 14

1.1 Introduction 15

1.2 The Importance of this Project 16

1.3 The Goal of this Project 17

1.4 Brief History of the Used Tools 18

1.4.1 Data warehouse (DW) 18

1.4.2 Architecture and End-to-End Process 19

1.4.3 Extract, transform, and load 21

1.4.4 Data mining 25

1.4.4.1 Introduction 25

1.4.4.2 Methods and Techniques 27

Chapter 2 Analysis and Design Study 30

2.1 Introduction 31

2.2 The Analysis Phase 32

2.2.1 Collecting GATS datasets and understanding it 32

2.2.2 Cleansing datasets using ETL 33

2.2.3 Eliciting the characteristics of smokers 34

2.3 The Design Phase 41

2.3.1 First phase of designing 41

2.3.2 Second phase of designing 43

Chapter 3 Project Implementation 45

3.1 Introduction 46

3.2 The Dashboards 48

3.3 Implementation of Data mining technique… 59

4. Conclusion & Future Expectations 70

4.1 Conclusion 71

4.2 What is coming next? 72

5. References 72

Table of figures:

Figure 1: Data warehouse architecture 19

Figure 2: Data warehouse attributes. 20

Figure 3: ETL tool 21

Figure 4: Load Process in PowerBI……..……25

Figure 5: Data mining methods 29

Figure 6: Workflow Process Architecture……..……31

Figure 7:Basic Datasets Format……..……33

Figure 8:Format for ready data 40

Figure 9: GATS Data warehouse architecture -1- ……..……..……42

Figure 10: Dax Studio export data wizard……..……..……..43

Figure 11: GATS Data Warehouse architecture -2-……..……..……44

Figure 12: percentage of cases by smoker categories.…...……...……..48

Figure 13: %smoke per day group by age first smoke group…….……49

Figure 14: % in Urban and % in Rural by purchase type……...……...50

Figure 15: Comparison of % in Urban and % in Rural by education……...……..51

Figure 16: %Successful quitters by Smoke Family yesno 52

Figure 17: %smoke per day group by education and employment……53

Figure 18: %smoke categories by education and employment…..……54

Figure 19: %male and female per 2009 and 2016…….……..……55

Figure 20: Top 10 smoking states.……..……56

Figure 21: Top 10 smokeless states.…….……..…57

Figure 22: %smoke categories by age first smoke group.…….……58

Figure 23: Classification using data mining algorithms in Orange Data Mining. 60

Figure 24: Result of testing classification algorithms.. 61

Figure 25: Plots a true positive rate against a false positive rate of a test. 62

Figure 26: Result of features importance for logistic regression algorithm. 63

Figure 27: Result of features importance for logistic regression algorithm.. 64

Figure 28: Prediction smoker type. 65

Figure 29: Result of prediction for each algorithm. 66

Figure 30: ARIMA model.. 67

Figure 31: forecasting the percentage of successful quitters and compare with the actual value.. 68

Figure 32: Comparison of percentage of successful quitting between the countries of world…...69

List of Tables:

Table 1: Description of the dependent and independent variables 34

Chapter 1

Theoretical study

1.1 Introduction

Cigarette smoking is an important global health issue and is a well-known modifiable risk factor for many diseases including cancer, cardiovascular diseases, respiratory diseases, malfunction of the reproductive system, and many other organ systems.

It is estimated that smoking and exposure to secondhand smoke reduced life expectancy by 15 years. Smoking cessation has been proven to be beneficial in many aspects of human health. Assisting smokers in quitting smoking is thus an important task. The path to doing so begins with the quality and volume of data they are able to collect from them.

But data alone is not the answer—without a means to interact with the data and extract meaningful insight, it's essentially useless.

Data mining is contributing to your ability to use that data to make accurate and dependable predictions that can allow you to operate at a higher level than simply relying on the smokers data that you have available to you, and guessing at future results.

1.2 The Importance of this Project

The importance of the research lies in the following point Work on important and relatively recent concepts, which are business intelligence systems in decision support.

1.1 The need to implement such systems to improve the level of public health and promote health sustainability in the community.

1.2 The need to develop analytical systems in order to take advantage of the large amount of data present in this organization to reach sound management decisions.

1.3 It is important to better understand why some individuals do not attempt quitting at all and why others despite attempting to quit, fail and relapse. Exploring the enablers and barriers of quitting smoked as well as smokeless tobacco will help to frame policies and strategies for implementing more successful cessation programs.

1.3 The Goal of this Project

This project aims Applying data mining techniques to the available data about the smokers (such as social, behavioral and environmental characteristics) to identify the factors that most influencing the success of trying to quit smoking.

So, goals can be summarized as:

1- Show the impact of prospecting the data mining application at the Global Adult Tobacco Survey (GATS) data from World Health Organization in supporting decision-making.

2- Building a business intelligence system that results in reports useful in knowing the correlation between social, environmental and behavioral characteristics.

3- Using machine learning algorithms to identify the factors that most affect smokers to quit and predict the type of smoker based on his data (a smoker who tried to quit smoking and failed, a smoker who did not try to quit smoking, a smoker who tried to quit smoking and succeeded).

4- Applying time series techniques and exponential smoothing algorithms to data from all countries issued by WHO surveys to predict the rate of smoking cessation in the world in the future.

1.4 A Brief History of the Used Tools

1.4.1 Data warehouse (DW)

Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. Many commercial products and services are now available, and all of the principal database management system vendors now have offerings in these areas. Decision support places some rather different requirements on database technology compared to traditional on-line transaction processing applications. This paper provides an overview of data warehousing and OLAP technologies, with an emphasis on their new requirements. We describe back-end tools for extracting, cleaning and loading data into a data warehouse; multidimensional data models typical of OLAP; front end client tools for querying and data analysis; server extensions for efficient query processing; and tools for metadata management and for managing the warehouse. In addition to surveying the state of the art, this paper also identifies some promising research issues, some of which are related to problems that the database research community has worked on for years, but others are only just beginning to be addressed. This overview is based on a tutorial that the authors presented at the VLDB Conference, 1996.

A data warehouse is a “subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making.” 1 Typically, the data warehouse is maintained separately from the organization’s operational databases. There are many reasons for doing this. The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases.

Data warehouses, in contrast, are targeted for decision support. Historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be orders

of magnitude larger than operational databases; enterprise data warehouses are projected to be hundreds of gigabytes to terabytes in size. The workloads are query intensive with mostly ad hoc, complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Query throughput and response times are more important than transaction throughput.

1.4.2 Architecture and End-to-End Process

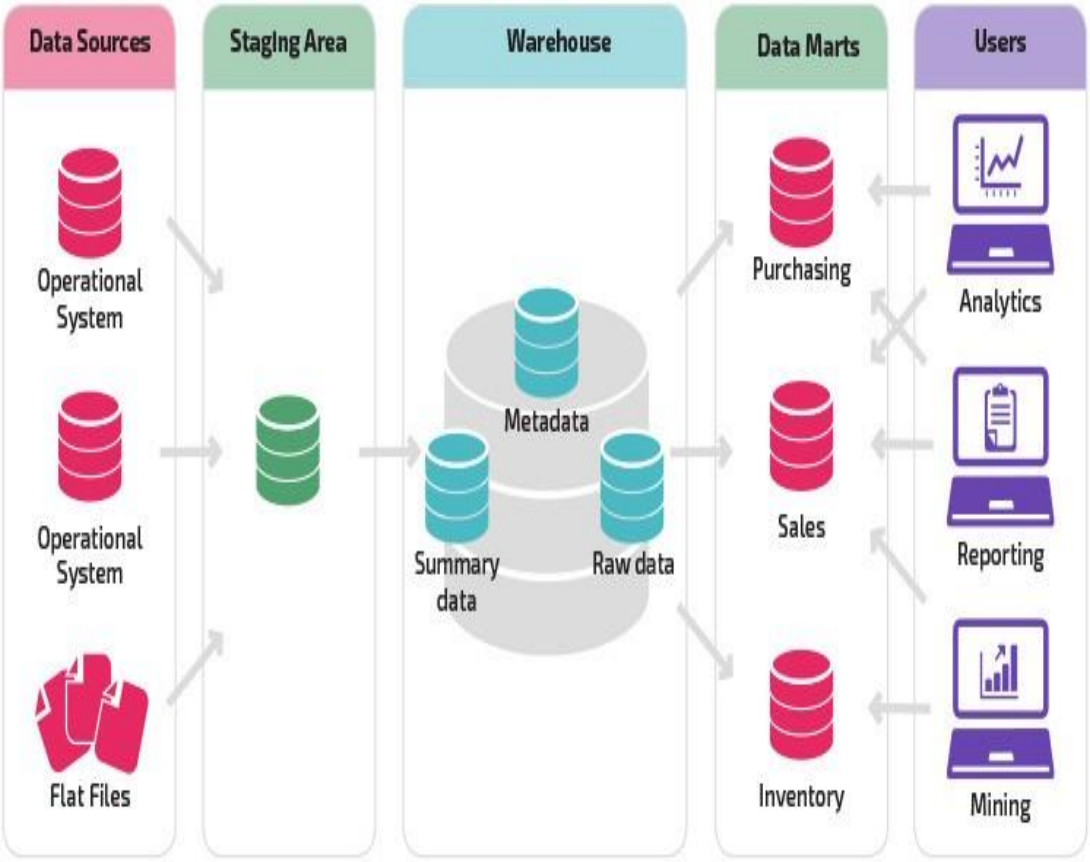


Figure 1: Data warehouse architecture (https://www.astera.com/)

It includes tools for extracting data from multiple operational databases and external sources; for cleaning, transforming and integrating this data; for loading data into the data warehouse; and for periodically refreshing the warehouse to reflect updates at the sources and to purge data from the warehouse, perhaps onto slower archival storage. In addition to the main warehouse, there may be several

departmental data marts. Data in the warehouse and data marts is stored and. managed by one or more warehouse servers, which present multidimensional views of data to a variety of front-end tools: query tools, report writers, analysis tools, and data mining tools. Finally, there is a repository for storing and managing metadata, and tools for monitoring and administering the warehousing system.

Designing and rolling out a data warehouse is a complex process, consisting of the following activities.

• Define the architecture, do capacity planning, and select the storage servers, database and OLAP servers, and tools.

• Integrate the servers, storage, and client tools.

• Design the warehouse schema and views.

• Define the physical warehouse organization, data placement, partitioning, and access methods.

• Connect the sources using gateways, ODBC drivers, or other wrappers.

• Design and implement scripts for data extraction, cleaning, transformation, load, and refresh.

• Populate the repository with the schema and view definitions, scripts, and other metadata.

• Design and implement end-user applications.

• Roll out the warehouse and applications.

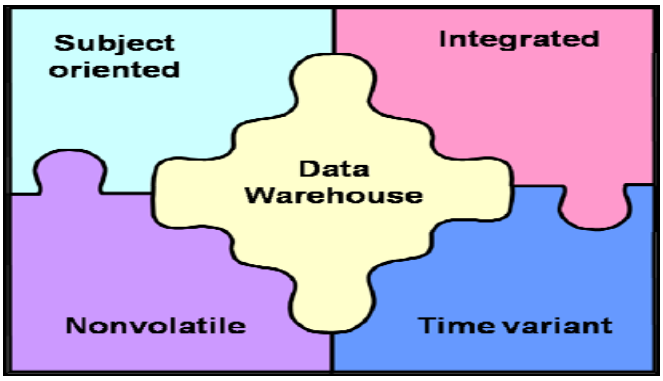


Figure 2: Datawarehouse attributes(https://www.slideserve.com).

1.4.3 Extract, transform, and load

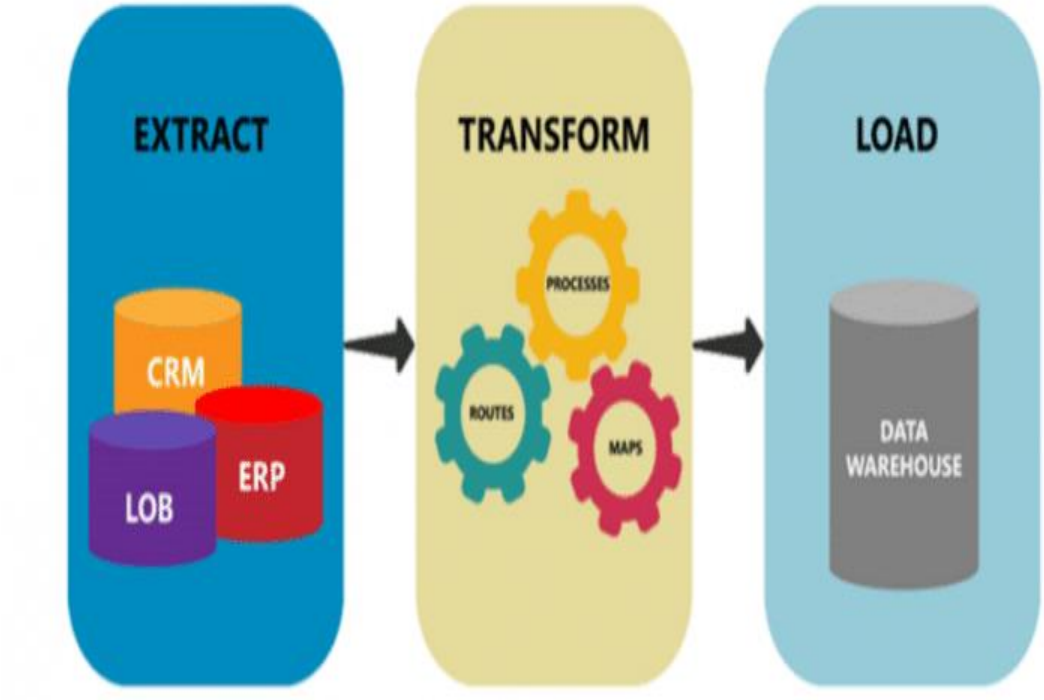


Figure 3: ETL tool (https://www.astera.com/)

An ETL tool is used to execute an integration project, and it includes three steps: extraction, transformation, and loading**.**

Thus, an ETL tool extracts data from disparate sources, transforms it to make it compatible with the destination system, and then loads it into the destination system. This destination could be a data warehouse, data lake, database, or any other application system.

A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

Since the data extraction takes time, it is common to execute the three phases in pipeline. While the data is being extracted, another transformation process executes while processing the data already received and prepares it for loading while the data loading begins without waiting for the completion of the previous phases

1.4.3.1 Extract

In this step of ETL architecture, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data warehouse.

Data warehouse needs to integrate systems that have different DBMS, Hardware, Operating Systems and Communication Protocols. Sources could include legacy applications like Mainframes, customized applications, point of contact devices like ATM, Call switches, text files, spreadsheets, ERP, data from vendors, partners amongst others.

Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship between sources and target data.

Three Data Extraction methods:

1. Full Extraction.
2. Partial Extraction- without update notification.
3. Partial Extraction- with update notification.

Irrespective of the method used, extraction should not affect performance and response time of the source systems. These source systems are live production databases. Any slow down or locking could affect company's bottom line.

1.4.3.2 Transform

Data extracted from source server is raw and not usable in its original form. Therefore, it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.

It is one of the important ETL concepts where you apply a set of functions on extracted data. Data that does not require any transformation is called as direct move or pass-through data.

In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading.

One of the used tools is R Studio:

R Studio is an open-source development environment for statistical and graphical programming. R Studio is used in the data cleaning process because it includes a set of tools and packages that help to analyze and clean data effectively.

If you have a large set of data, it can be difficult to manually identify and correct problems. With R Studio, you can use filter and sort tools to remove missing or duplicate values, change column formulas, and remove unwanted columns.

Overall, R Studio is a powerful data cleaning tool. It has a set of tools and packages that make data cleaning easy and efficient.

R is free programming language used in data analysis and statistics. They are widely used in the data cleaning process, where they are used to identify missing or invalid values in data sets.

R helps a lot with data cleaning, including removing missing values, handling outliers, validating data, and removing duplicates. Also, using the R language helps to create graphs and statistics to show the results of the data cleaning process.

There are a number of offices and packages that users use in the data cleaning process using R language, some of these offices are:

1- dplyr: This library is used for data analysis and cleaning.

2- tidyr: This library is used to convert data from a long format to a wide format.

3- ggplot2: This library is used to generate graphs.

4- stringr: This library is used to work with text strings.

5- lubridate: This library is used to work with dates and times.

6- plyr: This library is used to perform group operations on data.

7- reshape2: This library is used to reshape data.

1.4.3.3 Load

Loading data into the target Datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.

In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.

Types of Loading:

- **Initial Load** — populating all the Data Warehouse tables
- **Incremental Load** — applying ongoing changes as when needed periodically.
- **Full Refresh** —erasing the contents of one or more tables and reloading with fresh data.

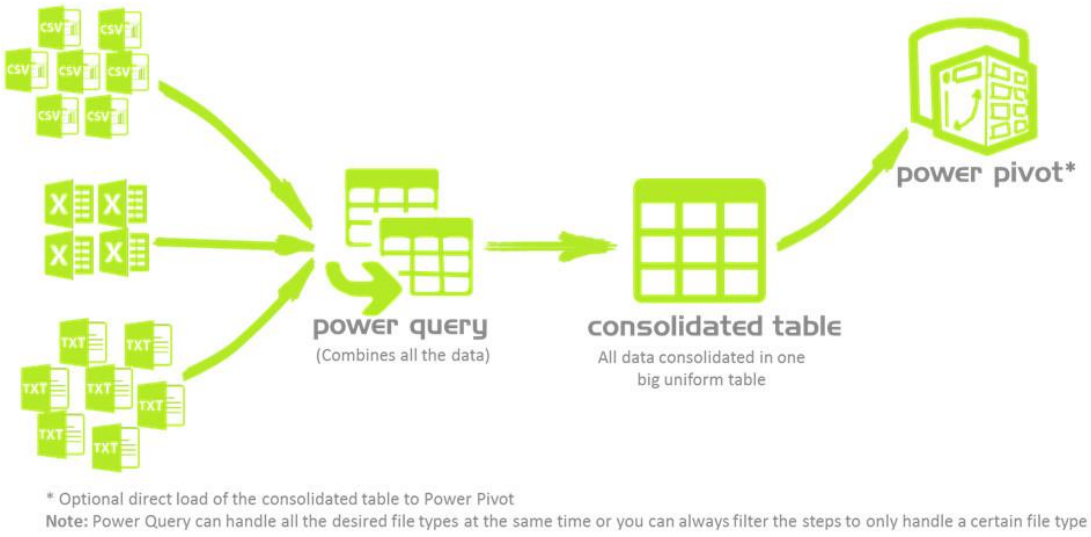


Figure 4: Load Process in Power BI (https://www.excelcampus.com/power-tools/power-query-overview/).

1.4.4 Data mining

1.4.4.1 Introduction

Data mining plays an important role in tobacco control and improving public health. For example, data mining can be used to analyze data on smoking and tobacco, and to identify patterns of smoking and related health effects.

The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases," the term "data mining" wasn't coined until the 1990s. But its

foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions).

What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power.

Over the last decade, advances in processing power and speed have enabled us to move beyond manual, tedious and time-consuming practices to quick, easy and automated data analysis. The more complex the data sets collected, the more potential there is to uncover relevant insights. Data mining can be used to analyze data on the health effects of smoking, identify the factors affecting smoking cessation, identify effective strategies to reduce smoking and improve public health, and improve prevention and health education programs related to smoking control.

Data mining techniques are a very valuable tool in the field of tobacco control, as they enable the analysis of large and diverse data related to tobacco and smoking, identifying factors affecting tobacco addiction and smoking, and analyzing associated patterns. Data mining techniques can be used in the field of tobacco control in many methods and applications, the most important of which are:

1- Analyzing smoking-related data: Data mining techniques can be used to analyze smoking-related data, and to identify smoking patterns and associated health effects. This data can be used to improve preventive and educational programs related to smoking control, and to identify effective strategies to reduce smoking and improve public health.

2- Analyzing data on other tobacco products: Data mining techniques can also be used to analyze data on other tobacco products, such as e-cigarettes and heated cigarettes, identify associated health effects and improve policies on tobacco products.

3- Analyzing data related to tobacco control policies: Data mining techniques can be used to analyze data related to various tobacco control policies, such as tobacco

taxes and smoking-related legislation, and to identify the associated effects and improve tobacco control policies.

Using data mining techniques, big data can be analyzed more efficiently and accurately than manual analysis, thus improving health governance and improving community health outcomes, as well as improving administrative and operational processes in tobacco control. Therefore, data mining is a very valuable tool for improving public health and tobacco control.

There are also disadvantages of data mining, namely in user privacy and security. It has to be clear how and with whom the information will be used and shared. Data mining tools and techniques work with very big amounts of data, so there is great cost at the implementation stage. It requires great IT experts for preprocessing data and finding the best model and technique for analysis. The techniques of data mining are not 100% accurate, so it may cause serious consequences and expenses.

1.4.4.2 Methods and Techniques

Various methods, algorithms and techniques are used for educational data mining. The most often use is for classification, clustering, prediction and association. The most frequently used data mining techniques are neural networks, decision trees, regression analysis and cluster analysis.

Classification is a data mining technique that segments data in a collection to target categories or classes. It helps in analyzing data and predicting outcomes. The goal of classification is to accurately predict the target class for each case in the data. The classifier training algorithm uses pre-classified examples for determining the set of parameters required for classification (Oracle, 2019). In the commercial sector, this technique is often used for classifying customer based on some characteristics such as on their purchase history, the type of food they eat, etc.

Clustering analysis is used to segment similar data into clusters that were not previously defined. It is useful in the data-preprocessing phase to identify homogeneous groups used as input for other models. Similar to classification,

cluster analysis can be used to investigate similarities and differences between customers, products, etc.

Prediction refers to calculated assumptions for certain events made based on available processed data. Regression technique can be used for prediction, to model the relationship between one or more independent variables and dependent variables. Independent variables are attributes already known and response variables are what we want to predict (Bhatnagar, 2013). It has many applications in business planning, trend analysis, financial forecasting, time series prediction, trend analysis, etc.

Association is a data mining technique used to discover the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are called association rules. Association rules are frequently used to analyze sales transactions. This type of finding is valuable for sales promotions, direct marketing, catalog design, cross-sell marketing and for discovering business trends. Based on certain rules, this technique can be used for the introduction of new courses or to open new colleges.

Neural networks are a set of computational algorithms, inspired by the human central nervous system, that is designed to recognize complex patterns and prediction problems by considering learning material, without being programmed (Graupe, 2013). They automatically identify special characteristics from the examples that they process. Neural networks consist of connected nodes called artificial neurons. Every connection can transmit a signal from one to another artificial neuron. A signal is a real number. Artificial neurons and connections have weights that adjust during the learning process. Neurons are divided into three layers, input, output and hidden layer. Signals travel from the input layer through hidden layers to the output layer, performing different kinds of transformations on their inputs. Its most important ability is to learn and model non-linear and complex relationships. The most common use of artificial neural networks is for speech and image recognition, computer vision, machine translation, for playing video games, financial analysis, social network filtering, control, optimization and medical diagnosis.

A decision tree is a decision support tool that uses a tree-shaped graph or model for classification. It is a supervised learning method. Each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf

node represents a class which is a decision after computing all attributes. The paths from the root to leaf are classification rules (Vidal et al, 2014). Their greatest advantage is stability and easy interpretation. Because of their simplicity, they are suitable for solving a different kind of problems in a broad range of industries such as financial, business, healthcare, education, energy, engineering, pharmaceutical, law, etc.

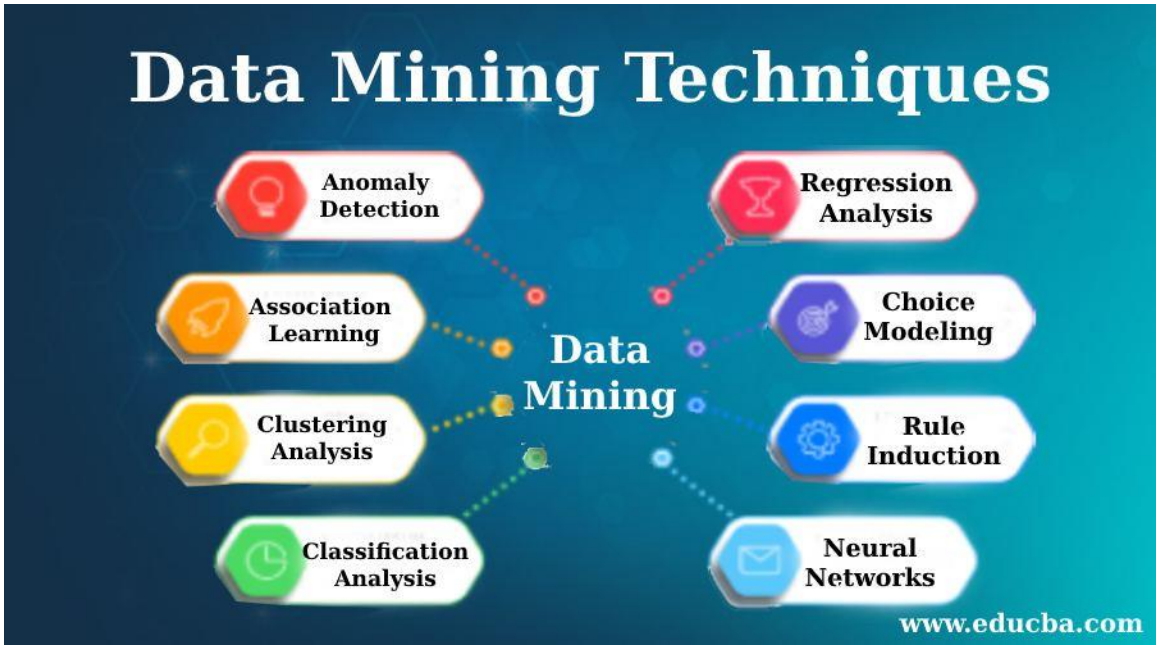


Figure 5: Data mining methods (https://cdn.educba.com)

Chapter 2
Analysis and Design Study

2.1 Introduction

Data Mining comprises the strategies and technologies used by WHO for the data analysis of smokers' information. Data Mining systems leverage software and services beginning with collecting data, cleansing data, integration of data, and then transform data into actionable insights that might be helpful while decision making process.

Implementing a Data Mining Algorithms requires careful planning to assure that it meets expectations. These are the basic steps:

Step 1: Collecting Global Adult Tobacco Survey (GATS) dataset and understanding it.

Step 2: Identify data source with all details.

Step 3: Cleansing data by (ETL).

Step 3: Building the Data Warehouse architecture (DW).

Step 4: Building Business Intelligence solution.

Step 5: Applying data mining algorithms.

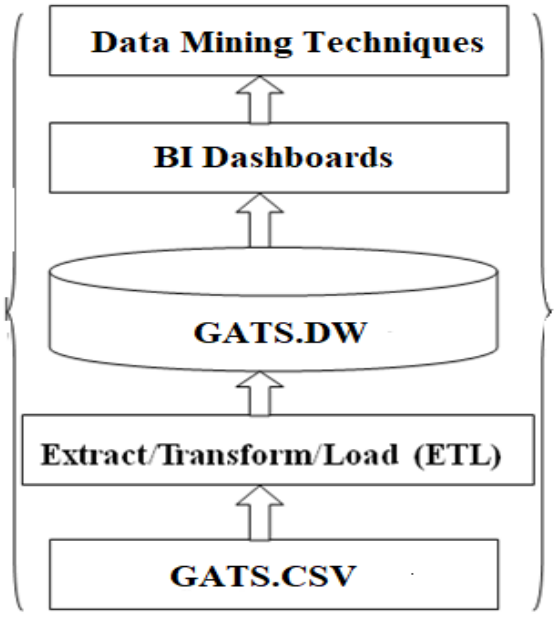


Figure 6: Workflow Process Architecture

2.2 The Analysis Phase

At first, we got the datasets from CDC (Centers for Disease Control and Prevention) website that approved by the World Health Organization.

The basic dataset is about surveys in India for smokers for the year 2009 that contains about (371 columns with 69,500 record) and year 2016 that contains about (465 columns with 74,050 record), and the additional datasets contains about 32 countries with 46 surveys for different years.

The basic step and means understanding the dataset, identifying the columns of importance in the analysis.

We split the data for two categories (smoke and smokeless).

Then we doing extract transform load data on our data. It's the most important and time-consuming step.

After that we design the data model, and finally we build Business Intelligence solution and machine learning model.

2.2.1 Collecting GATS datasets and understanding it:

The datasets were SPSS/STATA files that we converted into CSV files, and they were questions and answers (questionnaire).

The basic step and means understanding the dataset, identifying the columns of importance in the analysis, so we read the questions and answers in full.

The questionnaire contains approximately 450 questions, and it turned out that the datasets contain almost the same questions with a slight difference.

CASEID	AGE	RESIDENCE	NREGION	STATE	HH1	A01	A04	A05	A06a
3065	27	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3067	27	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3069	32	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3095	37	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3119	30	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3129	30	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3199	60	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3223	30	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3225	28	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3253	52	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3255	64	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3265	36	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3331	40	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3351	55	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3379	40	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3401	32	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3439	36	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3531	30	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3541	35	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3587	43	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES
3615	45	Rural	North	Himachal Pradesh	4	FEMALE	NO FORMAL SCHOOLING	HOMEMAKER	YES

Figure 7: Basic Dataset format

2.2.2 Cleansing datasets using ETL:

After gathering the datasets in CSV files, we needed to clean the datasets which the important stage of building a data warehouse and as an example of this process we have the following operation:

- We found many null values in the important column of the analysis process for the question (How old were you when you first started using smoke/smokeless tobacco daily?):

-since we have the current age column and the column for the question (How many years ago did you first start using smoke/smokeless tobacco daily?) we have the following cases:

1. In the case that the current age is greater than the number of years since smoking began, we replaced the null values by subtracting the value of the current age column from the value of the last question column.

2. In the case that the current age is less than the number of years since smoking began, we replaced the null values with a symbol indicating that these values are irrational.

3. In the case that the current age is equal to the number of years that have passed since smoking began, we replaced the null values with the column value of the last question, and this case means that the person has recently started smoking.

To clean up the data, we used R language and Power BI program.

2.2.3 Eliciting the characteristics of smokers:

Since the data are questions and answers, which are not useful in the analysis process, we decided to convert them into meaningful and useful variables and elicit the characteristics of smokers (Environmental, behavioral and social characteristics) from these questions so that each characteristic contains several factors.

We divided the variables into two parts, dependent and independent variables.

Table1: Description of the dependent and independent variables

Dependent variables		
Name of the variable	**Categories**	**Definitions**
Quit attempt among current smokers / smokeless tobacco users	Yes No	Current smokers / smokeless tobacco users responded 'daily' or 'less than daily' to question: 'Do you currently smoke/ use smokeless tobacco? Making a quit attempt: Current smokers/smokeless users who tried to quit during the past 12 months for more than one day. Those attempted to quit responded 'Yes' and those did not attempt to quit responded 'No' to the question: 'During the past 12 months, have you tried to stop smoking/using smokeless tobacco', and the following question was used to calculate if they quit for more than one day: 'Thinking about the last time you tried to quit, how long did you stop smoking/using smokeless tobacco?' (if a current smoker/smokeless user responded

		'Yes' to the question: 'During the past 12 months, have you tried to stop smoking/using smokeless tobacco' and 'less than one day' to the question 'Thinking about the last time you tried to quit, how long did you stop smoking/using smokeless tobacco?', it was categorized as current smokers/smokeless user who did not attempt to quit)
Successful quitting	Yes No	Former smokers / smokeless tobacco users responded: a) 'not at all' to question 'Do you currently smoke/ use smokeless tobacco? b) 'daily' or 'less than daily' to question 'In the past, have you smoked/used smokeless tobacco on a daily basis, less than daily, or not at all?' Being successful in quitting: Successful quitters responded 'more than 6 months and up to 24 months' to the question "How long has it been since you stopped smoking/using smokeless tobacco?"
Independent variables		
Name of the variable	**Categories**	**Remarks if any**
Socio-demographic characteristics		
Residence	• Urban • Rural	-
Age	• 15-24 • 25-44 • 45-64 • >65	-
Sex	• Male • Female	-
Education	• No formal education • < Primary to Primary • < Secondary to higher secondary	Individuals with missing information were excluded from the analyses

	• College	
Employment	• Employed • Self & ad-hoc employed • Student • Not employed	
Environmental characteristics		
Smokers/smokeless users at home	• No • Yes	If there are one or more smokers/smokeless users at home then it is 'Yes', and any other case is 'No'.
Permissibility of smoking at home	• Not allowed • Allowed • Don't know/ Refused/ Missing/Not Applicable	"Allowed" include: Allowed, Not allowed but exceptions and no rules.
Magnitude of second-hand exposure at home	• Never • Daily • Occasionally • Don't know/ Refused/ Missing/Not Applicable	"Occasionally" include: weekly, monthly and less than monthly.
Permissibility of smoking at work	• Not allowed • Allowed • No/don't work • Don't know/ Refused/ Missing/Not Applicable	"Allowed" include: Allowed anywhere, allowed only in some indoor areas and there is no policy. "No/don't work" is one possible answer to the question: Do you currently work outside of your home?
Advise to quit by health care provider	• No • Yes • Does not apply • Refused/ Missing/Not Applicable	"Does not apply" are those who either did not visit health care provider or those who were not asked by health care provider if they used tobacco.

Behavioral characteristics		
Age at initiation	• 1-24 • 25-44 • >45 • Don't know/ Refused/ Missing/Not Applicable	Data on age at initiation of first smoke/smokeless use was available for the same participant in two different variables. In this case, the variable with less missing values namely the question, "How old were you when you first started smoking/using smokeless tobacco daily?" was used as the primary variable; for those missing, data was then imputed from the second variable, "How many years ago did you first start smoking/using smokeless tobacco daily?", the data was computed by subtracting this value from the age.
Time of first smoke (minutes)/ Time of first smokeless use (minutes)	• Up to 30 • >30 • Not asked • Refused/ Missing/Not Applicable	"Not asked" category includes less than daily current smokers/smokeless tobacco users from whom this data was not collected. Data was not collected for the successful quitters.
Typical daily consumption	• <1 • 1 to10 • >10 • Don't know/ Refused/ Missing/Not Applicable	This variable was created by summing up the number of times a participant used any of the smoked/smokeless products per day. For less than daily users, the numbers smoked/smokeless used per week was divided by 7. Data was not available for the successful quitters.
Purchase type	For smoked tobacco: • Loose • Pack • Both • Others For smokeless tobacco: • Pouch • Can • Loose and others	For smoked tobacco: "Loose" if both cigarettes and bidis were purchased as loose "Pack" if both cigarettes and bidis were purchased as pack or carton "Both" if one product was purchased as loose and the other as pack Everything else was considered as "Others". For smokeless tobacco:

	• Never bought • Refused/ Missing/Not Applicable	"Pouch" refers to Pouches (GATS 1) or Single use pouch (GATS 2). "Can" refers to cans (GATS 1) or large pouch or can (GATS 2). "Loose and others" refer to purchase as loose or other forms Data was not collected for successful quitters.
Periodicity	• Daily • Less than daily	
Knowledge on the ill effects of smoking/smokeless tobacco	• No • Yes • Don't know/ Refused/ Missing/Not Applicable	Number of questions was different in GATS 1 and 2. One common question on the serious ill effects of smoking/smokeless tobacco has been used as a surrogate to assess the knowledge.
Noticing anti-smoking/anti-smokeless advertisement	-	This variable was created by counting the number of "Yes" as an answer to questions related to noticing anti-smoking/anti-smokeless advertisement in the last 30 days in the two surveys and dividing it by number of questions. This was done for each survey separately to account for the different number of questions between the two. For smoking, questions used from GATS 1: G01a1, G01a2, G01b1, G01b2, G01c1, G01c2, G01d1, G01d2, G01e1, G01e2, g02, and from GATS 2: G01A, G01B, G01C, G01D, G01E, G01F, G01G, G01H, G01I, G02, G02B. For smokeless, questions used from GATS 1: G01a3, G01b3, G01c3, G01d3, G01e3, G02a, and from GATS 2:

		G201A, G201B, G201C, G201D, G201E, G201F, G201G, G201H, G201I, G02A.
Noticing smoking/smokeless tobacco promotions	-	This variable was created by counting the number of "Yes" as an answer to questions related to noticing smoking/smokeless tobacco promotions in the last 30 days in the two surveys and dividing it by number of questions. This was done for each survey separately to count for the different number of questions between the two. For smoking, questions used from GATS 1: G04a1, G04a2, G04b1, G04b2, G04c1, G04c2, G04d1, G04d2, G04e1, G04e2, G04f1, G04f2, G04g1, G04g2, G04h1, G04h2, G04i1, G04i2, G04j1, G04j2, G04k1, G04k2, G06a1, G06a2, G06b1, G06b2, G06c1, G06c2, G06d1, G06d2, G06e1, G06e2, G06f1, G06f2, G06g1, G06g2, and from GATS 2: G04A, G04B, G04C, G04D, G04E, G04F, G04G, G04H, G04I, G04J, G04K, G06A, G06B, G06C, G06D, G06E, G06F, G06G, G206A, G206B, G206C, G206D, G206E, G206F, G206G. For smokeless, questions used from GATS 1: G04a3, G04b3, G04c3, G04d3, G04e3, G04f3, G04g3, G04h3, G04i3, G04j3, G04k3, G06a3, G06b3, G06c3, G06d3, G06e3, G06f3, G06g3, and from GATS 2: G204A, G204B, G204C, G204D, G204E, G204F, G204G, G204H, G204I, G204J, G204K, G306A, G306B, G306C, G306D, G306E, G306F, G306G.
Users of the other tobacco form	• Mono • Dual • Former	Mono: Have not used the other form. Dual: currently using the two forms of tobacco. Former: used the other form of tobacco in the past.

After extracting the previous characteristics, the data became subject to a normalization approach because it became in the form of categories, that is, within a specific range, and this approach is the most important to be very accurate machine learning and prediction models.

Since the economic income status of smokers is the most important factor for the analysis, but it is not available in the data, so we found that the employment and education factor are the most reflective of the economic situation, we filtered the data on their basis, that is, we ignored all the people whose educational and employment data was missing.

Since we classified smokers into three categories (a current smoker who tried to quit smoking and failed to quit - a current smoker who did not try to quit smoking - a smoker who tried to quit smoking and succeeded in quitting) within the dependent variables, which are the basis in the prediction process and the process of determining the factors that most affect the cessation of smokers, we also filtered this variable and ignored all the people whose type was missing.

age_group	gender	education	employment	age_first_smk_group	time_first_smk
2	1	3	2	1	1
3	1	1	2	1	1
2	2	1	4	2	2
2	1	3	2	1	1
2	1	3	1	1	2
3	1	3	2	1	1
1	2	1	4	1	1
1	1	2	2	1	3
2	2	2	2	1	1
2	1	3	1	4	4
2	1	3	2	1	4
2	2	4	3	4	4
3	1	3	1	4	4
2	1	3	2	4	4
4	1	2	4	4	4
3	2	2	4	2	4
2	1	2	2	4	4

Figure 8: Format for ready data

Now the data is ready, which is 46 files, and we have merged all the files into one file and added a column for the year and country.

2.3 The Design Phase

2.3.1 First phase of designing

The Data warehouse was built using SQL 2008 Enterprise products, and we used the Star schema to model the DW.

Star Schema in data warehouse, in which the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star. The Star Schema data model is the simplest type of Data Warehouse schema. It is also known as Star Join Schema and is optimized for querying large data sets. The fact table contains all the facts, while the dimension table contains the objects and properties, where the dimension tables are connected to the fact table using the foreign key from each dimension table.

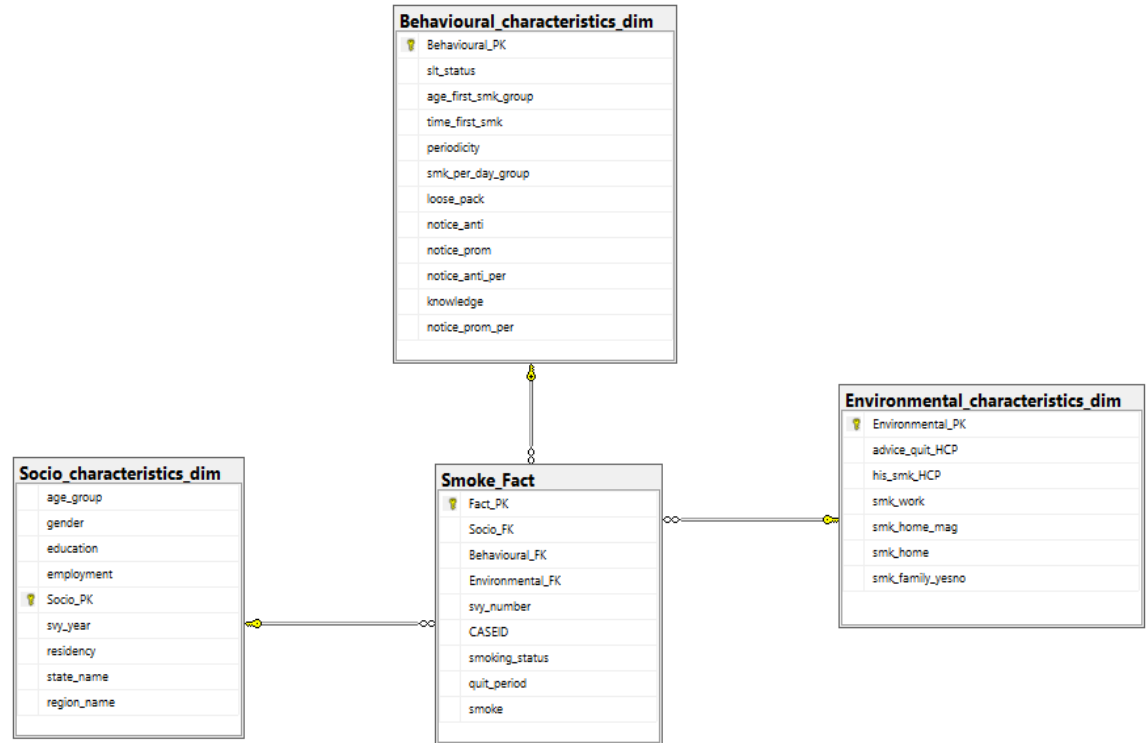


Figure 9: GATS Data warehouse architecture -1-

To load the data from the power pi file into the data warehouse we used the DAX Studio export data from Power bi to SQL Server 2008 enterprise.

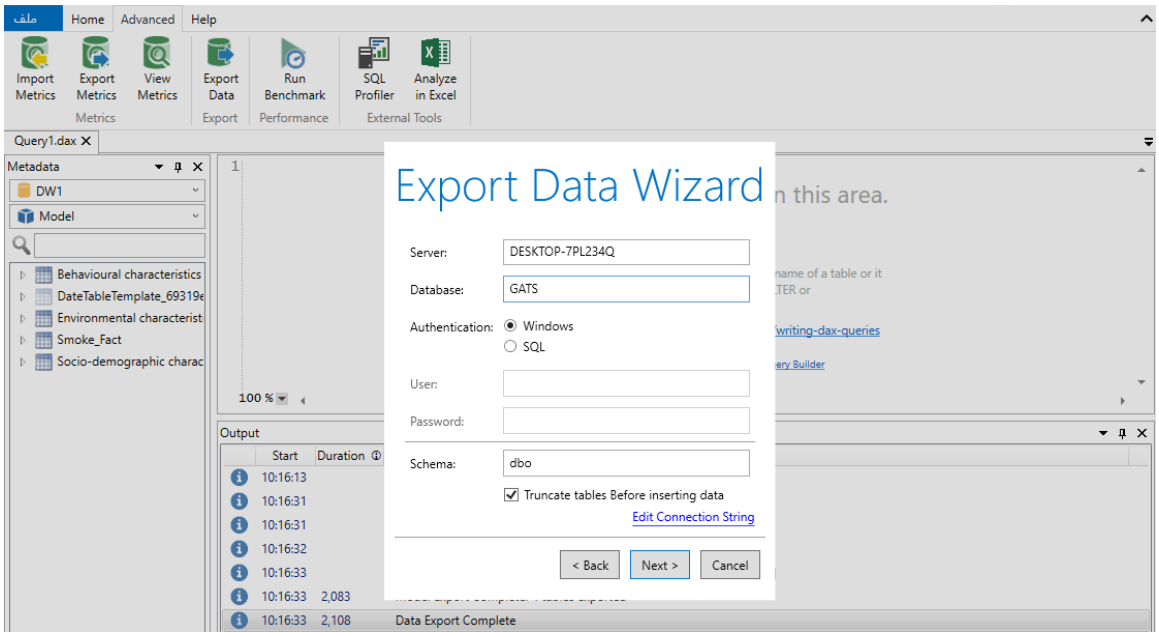


Figure 10: Dax Studio export data wizard

After the data was loaded into the data warehouse, we needed to connect the DW with the power BI environment to build the dashboards.

2.3.2 Second phase of Designing

After calculating the measures and adding them to the fact table, the Data Warehouse where this scheme is designed flexible and efficient to help in the next stages:

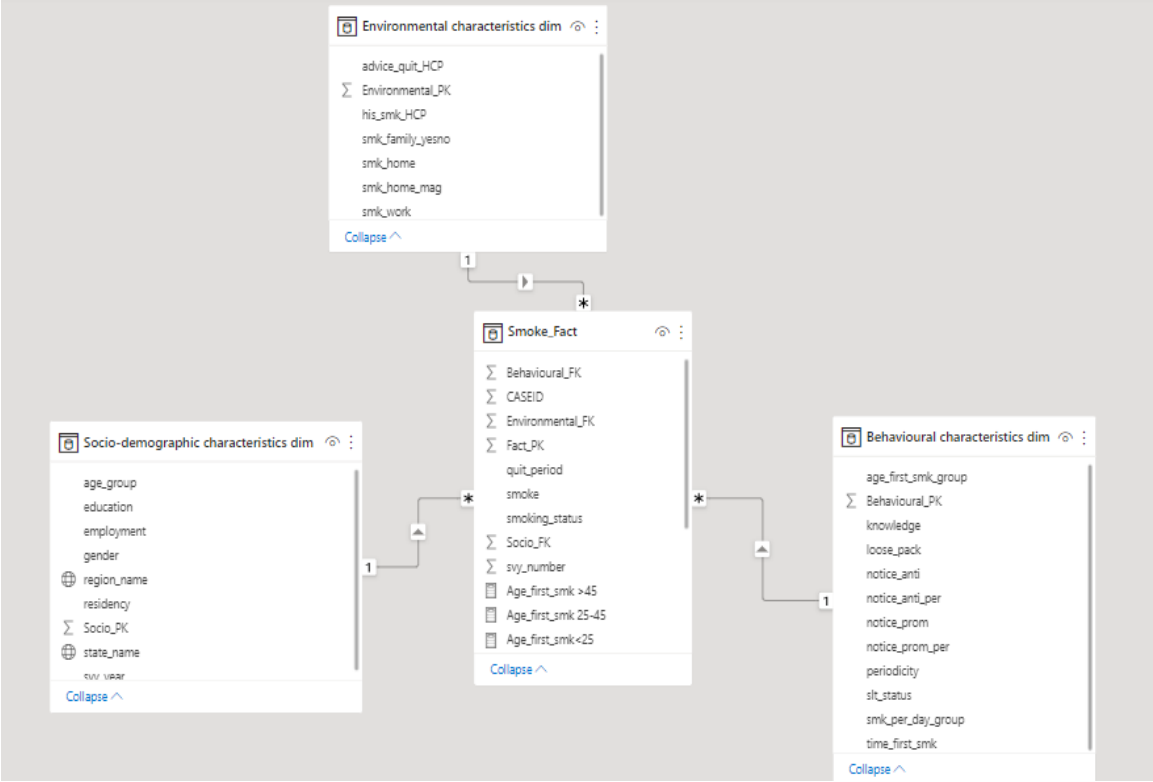


Figure 11: GATS Data Warehouse architecture -2-

Chapter 3
Project Implementation

3.1 Introduction

While building the data warehouse we build a data warehouse that consists of one Fact table in addition three Dimension tables.

While building the data warehouse we consider the approach of de-normalization of the DW which means more redundancy of data with faster performance while calculating the measures.

For India data that containing both smoke/smokeless this model allows us to ask the following analysis questions:

1. What is the count of smokers in each country?
2. What is the %GT purchase type by residency?
3. What is the %GT smoke/smokeless per day group by age first smoke/smokeless group?
4. What is the education ratio by residency?
5. What is the %Successful quitters by Smoke/smokeless family statues?
6. What is the % smoke/smokeless per day group by education and employment?
7. What is the number of cases by gender and survey year?

8. What is the %GT age first smoke/smokeless group by smoke category?

And then we implement some of the most famous data mining techniques to provide some powerful predictions and classification and determine the most affect factors, and apply time series techniques and exponential smoothing algorithms for forecasting.

3.2 The Dashboards

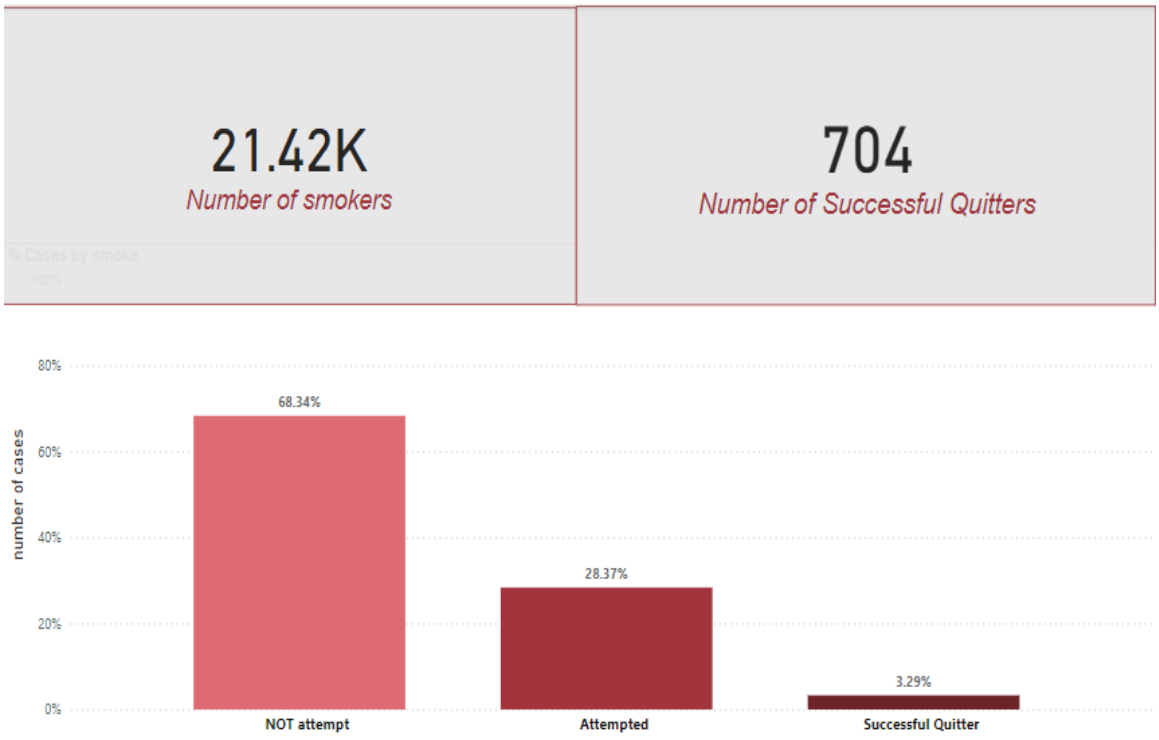


Figure 12: percentage of cases by smoker categories

The previous report shows us the percentage of cases per each smoker category.

Using this report, we find a disparity in the percentage of cases per smoker categories, The percentage of cases that did not attempt to quit smoking is the maximum, in smokeless data the disparity between the categories was also similar to smoke data and the category of not trying is the greatest as well.

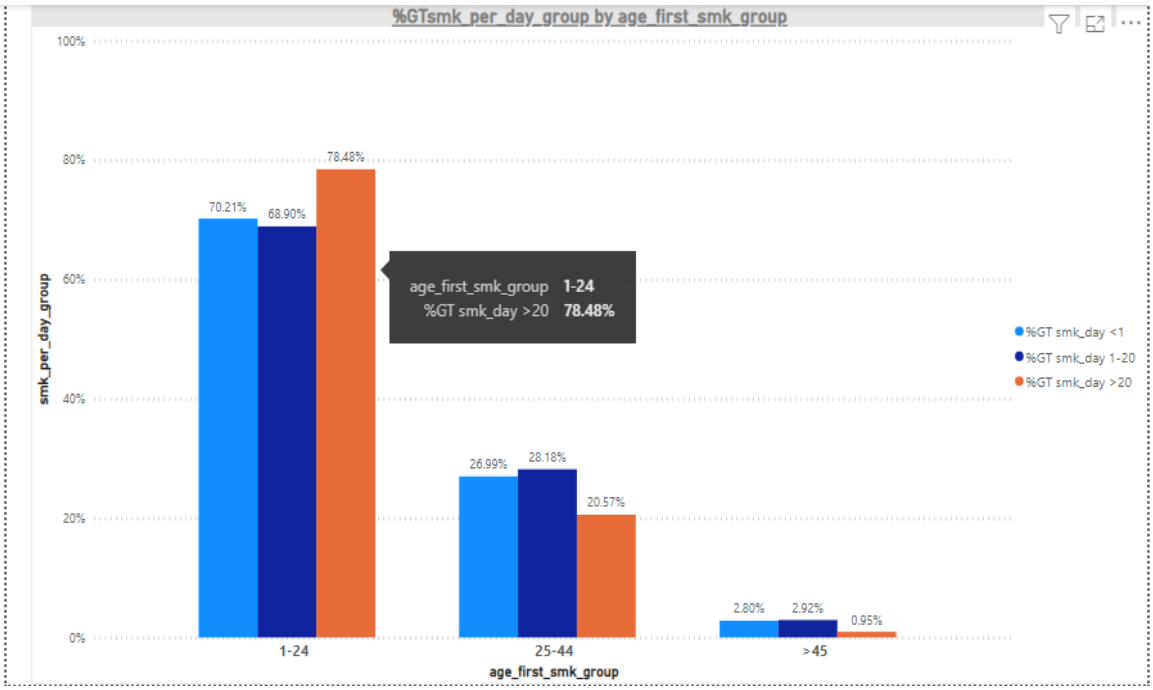


Figure 13: %smoke per day group by age first smoke group

The previous report shows us the percentage of daily smoking in each of them for each age first smoke group.

Using the previous report, we find that if smoking starts from a young age, the percentage of daily smoking is greater than other cases, in smokeless data the results were similar to smoke data.

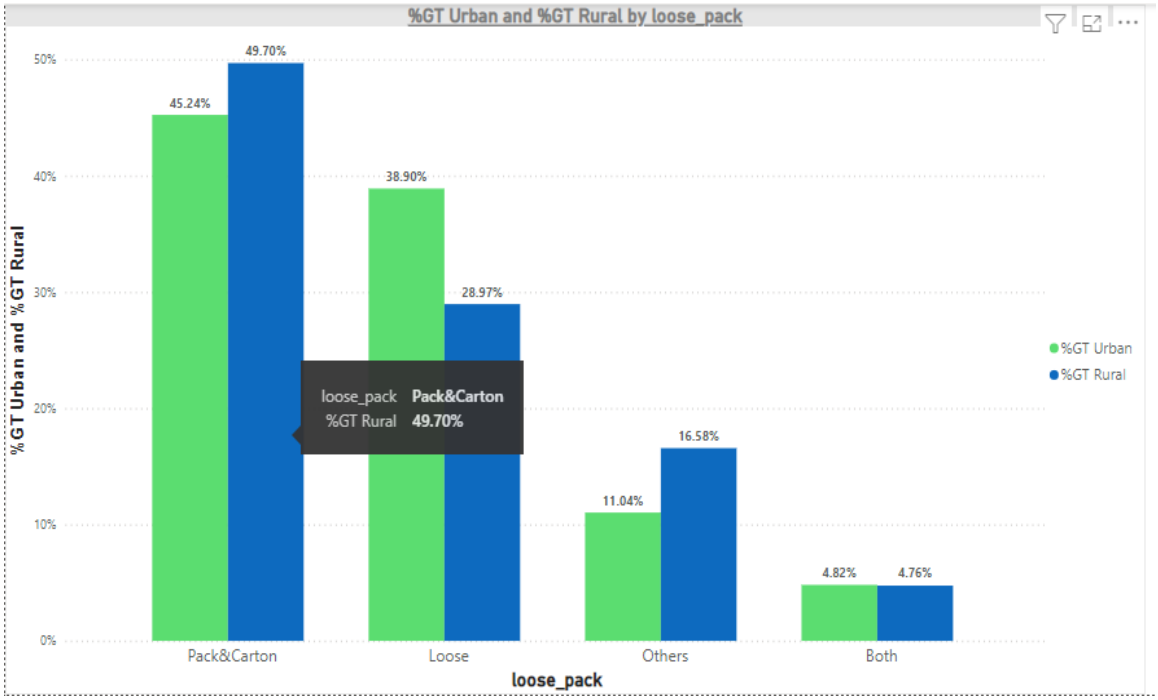


Figure 14: % in Urban and % in Rural by purchase type

The previous report shows us the percentage of purchase of each type by residency.

Using the previous report, the percentage of purchase a cartons&packets in the rural is the greatest, while in the urban large quantities of the loose type are bought, meaning that large quantities of smoke are bought in the rural.

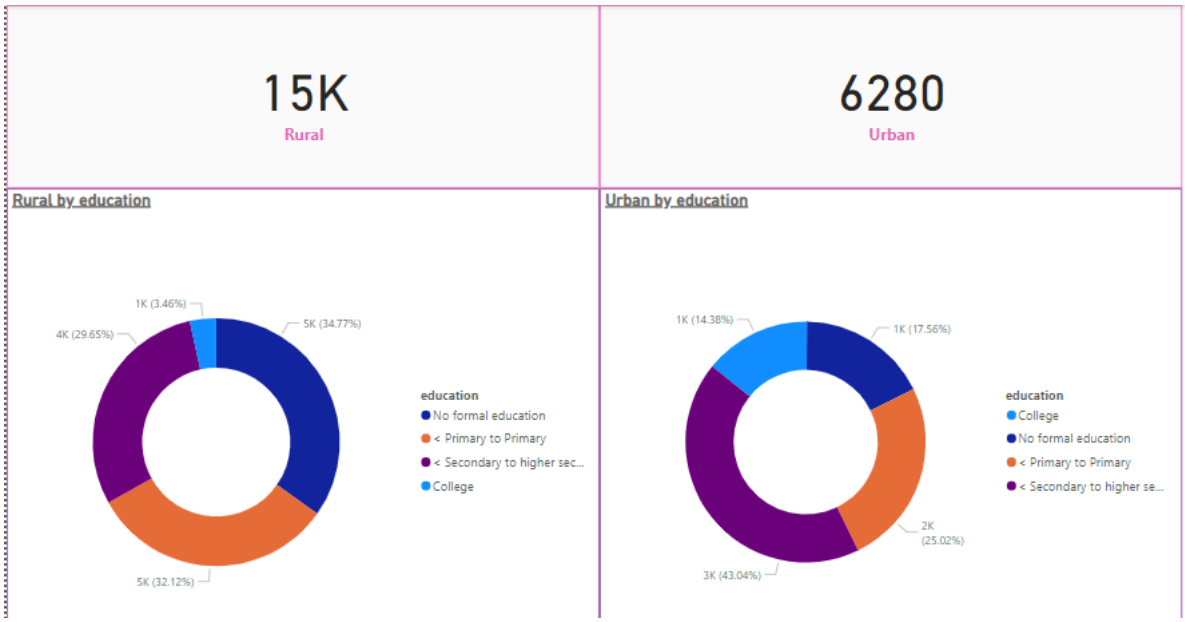


Figure 15: Comparison of % in Urban and % in Rural by education

The previous report shows the Comparison between the percentage of education in urban and rural areas.

Using the previous report, we find the level of education in rural is very low, so using smoke/smokeless tobacco rate was very high compared to urban.

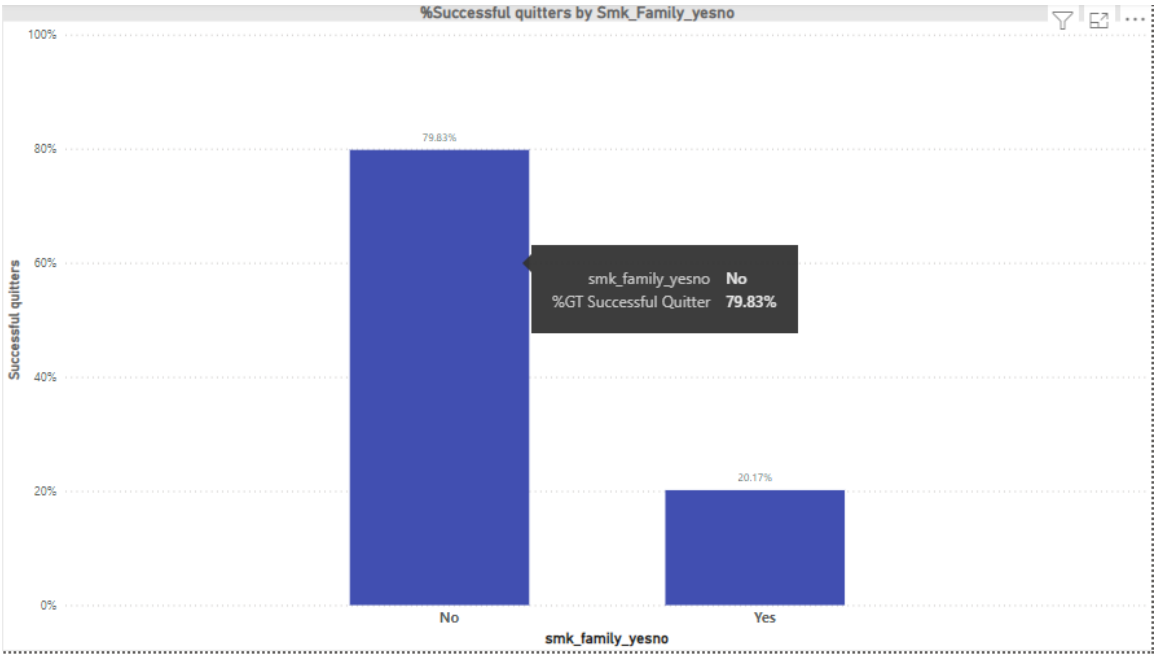


Figure 16: %Successful quitters by Smoke Family yesno

The previous report shows us the percentage of Successful quitters by smoke family yes-no.

Using the previous report, we find a disparity in the percentage of Successful quitters, that is, if a person is exposed to secondhand smoke/smokeless from the family, his success rate is lower in quitting smoking.

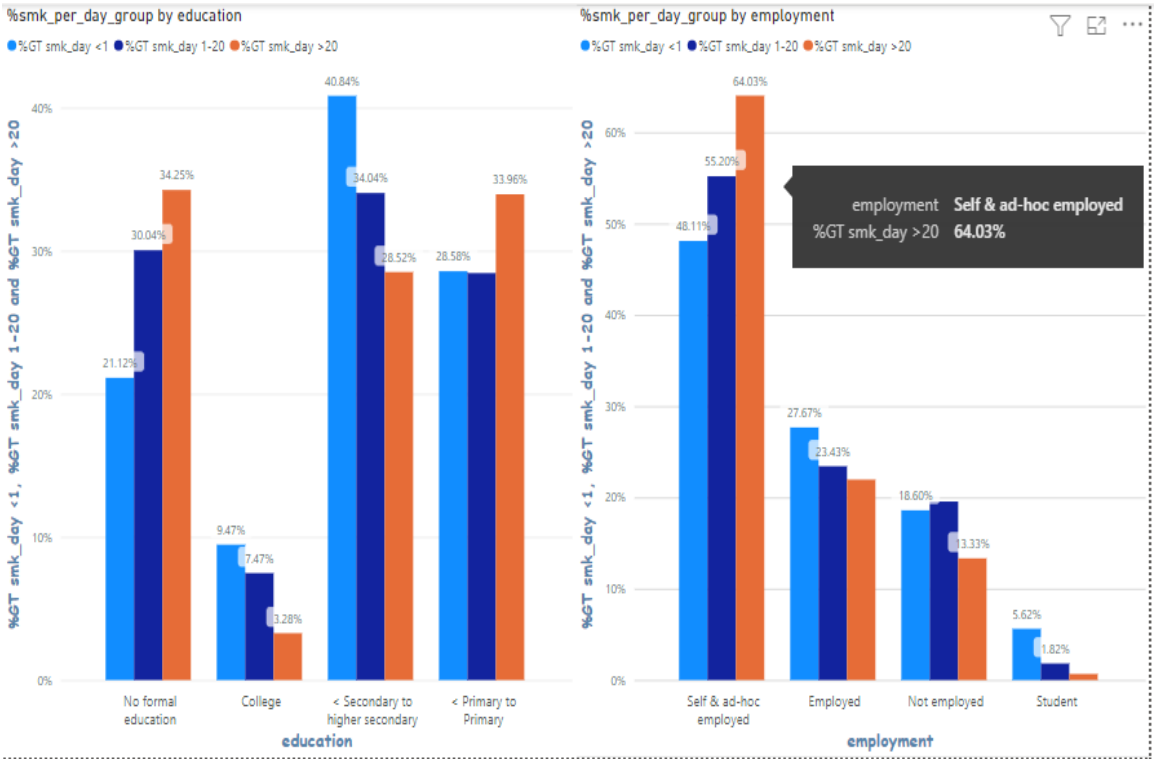


Figure 17: %smoke per day group by education and employment.

The previous report shows us the percentage of daily smoking in each of them for each education and employment level.

Using the previous report, we find if the person is illiterate, his daily smoking rate is very large for educated people, and if the person is a self & ad-hoc employed, his daily smoking rate is also high compared to the student or employee.

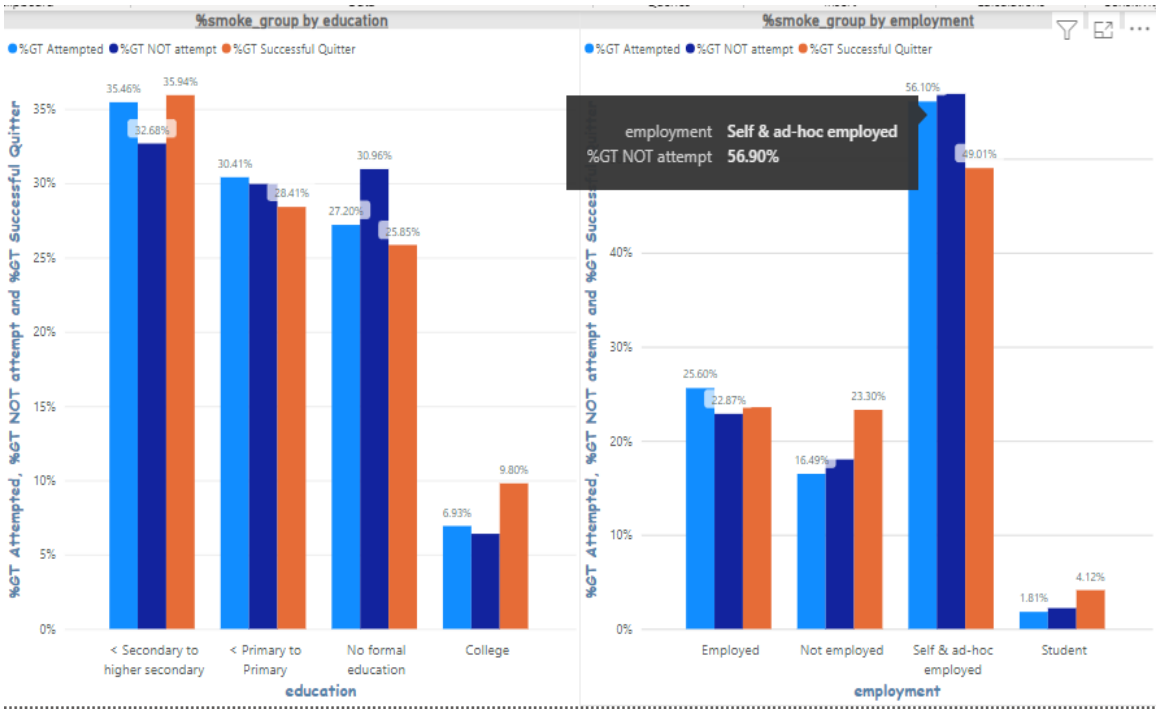


Figure 18: %smoke categories by education and employment.

The previous report shows us the percentage of smoke categories in each of them for each education and employment level.

Using the previous report, we find if the person is illiterate, his smoke category is not attempt compared to other educated people, and if the person is a self & ad-hoc employed, his smoke category is also not attempt compared to the student or employee.

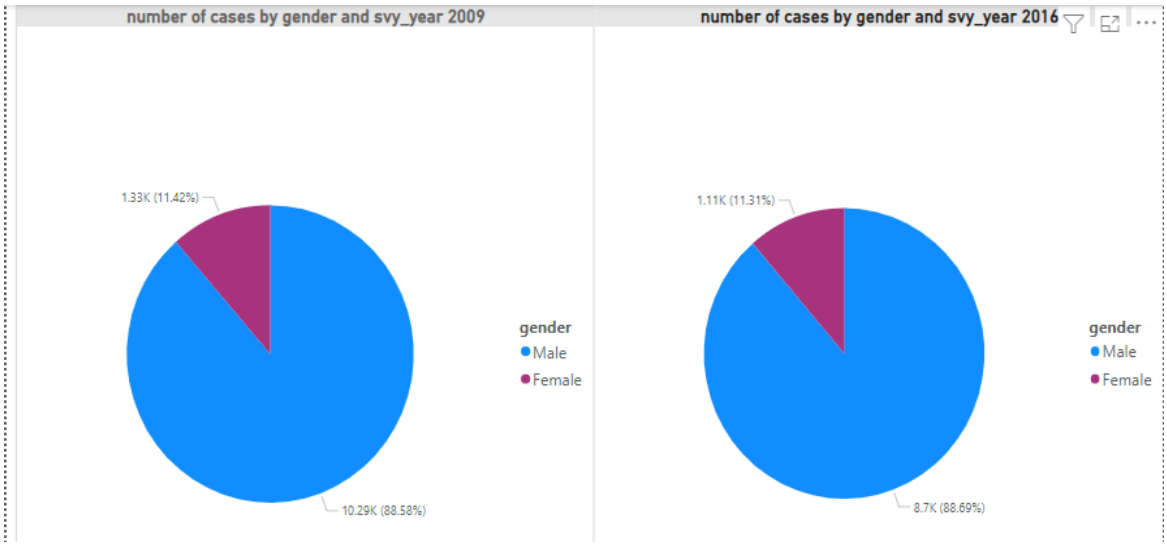


Figure 19: %male and female per 2009 and 2016.

The previous report shows us the percentage of male and female per year 2009 and 2016.

Using the previous report, we find the percentage of male smokers is much higher than the percentage of female smokers in both years.

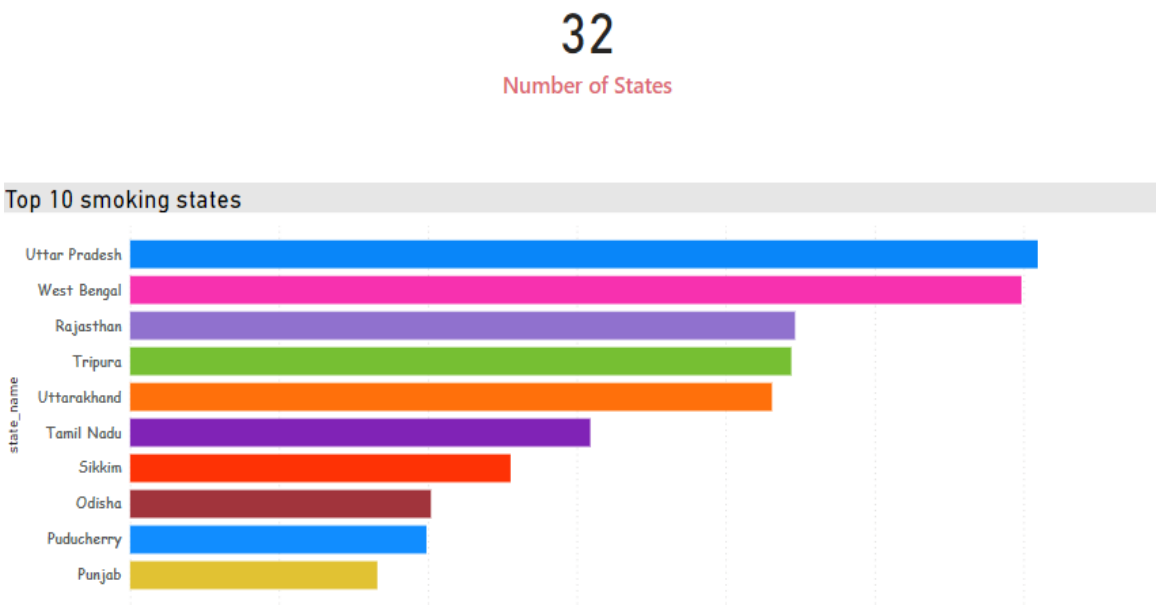


Figure 20: Top 10 smoking states.

The previous report shows us the top 10 smoking states.

Using the previous report, we find that Uttar Pradesh state is contain the most percentage of smokers.

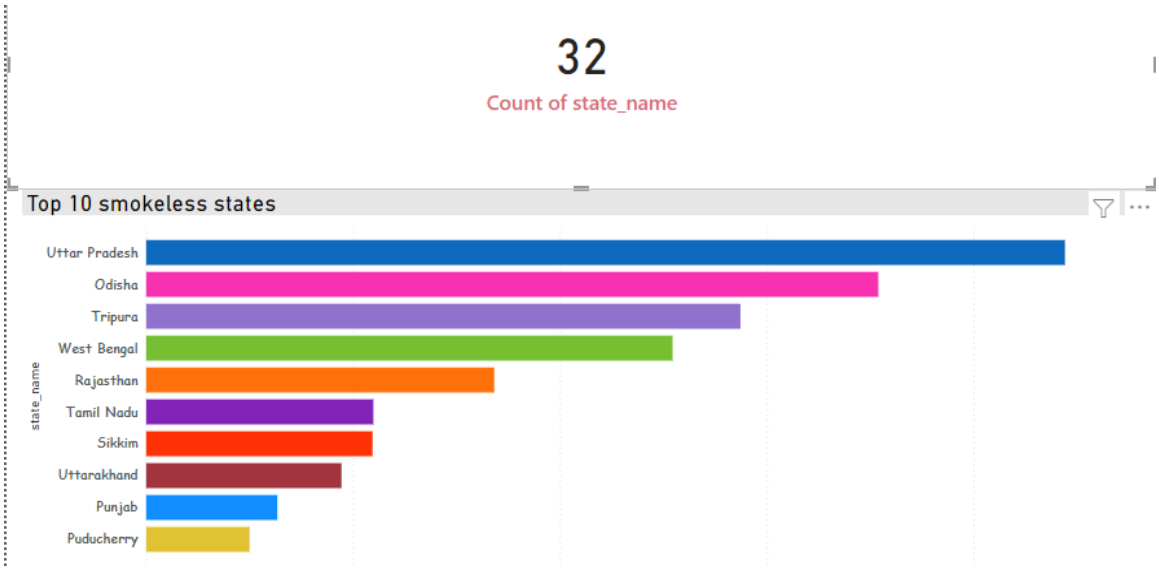


Figure 21: Top 10 smokeless states.

The previous report shows us the top 10 smokeless states.

Using the previous report, we find that Uttar Pradesh state is also contain the most percentage of smokeless.

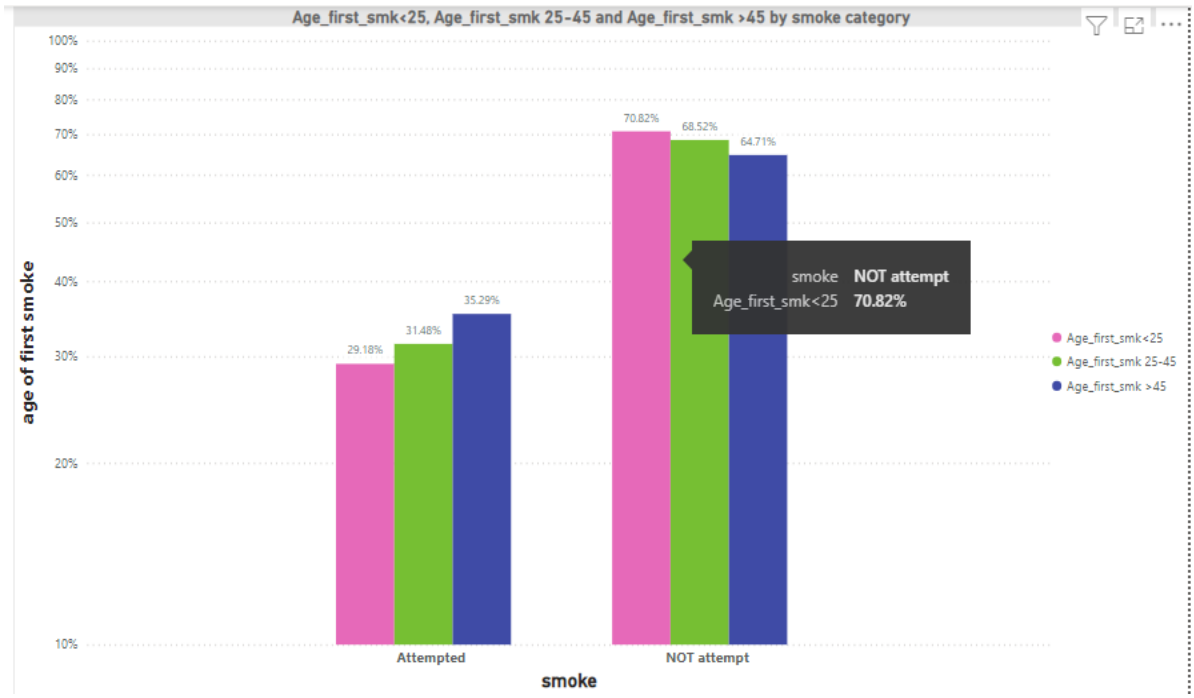


Figure 22: %smoke categories by age first smoke group.

The previous report shows us the percentage of attempt and not attempt by the age of first smoke group.

Using the previous report, we find that people who start smoking at a young age have a large percentage of not attempt to quit smoking, while those who start smoking at an old age have a large percentage of attempt to quit smoking.

3.3 Implementation of Data mining techniques:

We used two main method of data mining:

1- Classification: allow us to show relation between the qualitative variables to other variables.

2- Predictions: allow us to predict the future.

Classification algorithms:

We used KNN and SVM and Neural network and Random Forest and Decision Tree and Logistic regression to classify the smokers and identify the most factors effect on successful quitting for smokers.

Prediction algorithms:

The main predictions were made using the exponential smoothing algorithms to predict the success rate in smoking cessation and smoking rate in the world over the years and compare it to the actual rate.

Other predictions were made using time series techniques such as ARIMA model to predict the percentage of successful quitting and smoking rate for several years of the shop.

And there's a prediction based on Classification algorithms to predict smoker category.

.

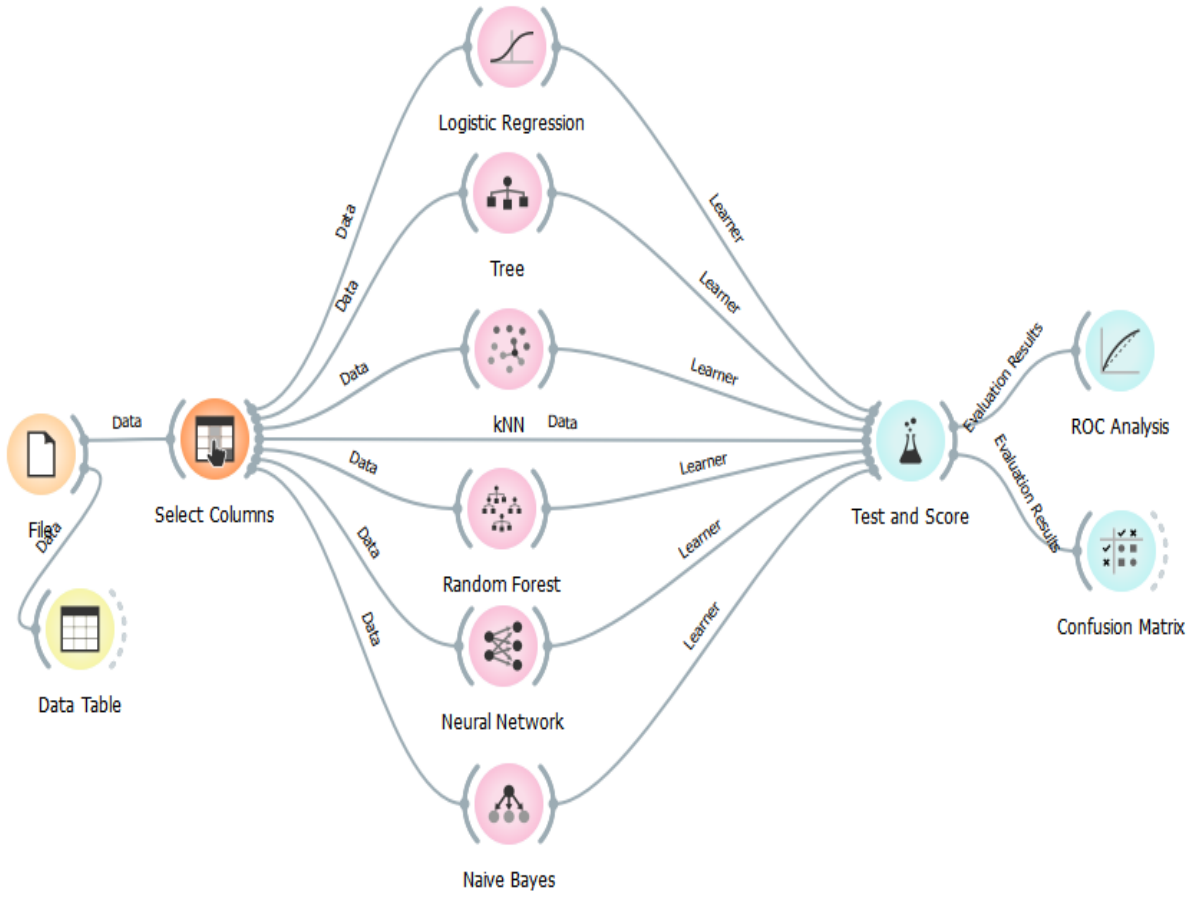


Figure 23: Classification using data mining algorithms in Orange Data Mining.

Comparison of classification algorithms (Naive Bayes and KNN and SVM and Neural network and Random Forest and Decision Tree and Logistic regression) in Orange.
The performance of a classifier can be evaluated by calculating the accuracy, precision, recall, specificity.

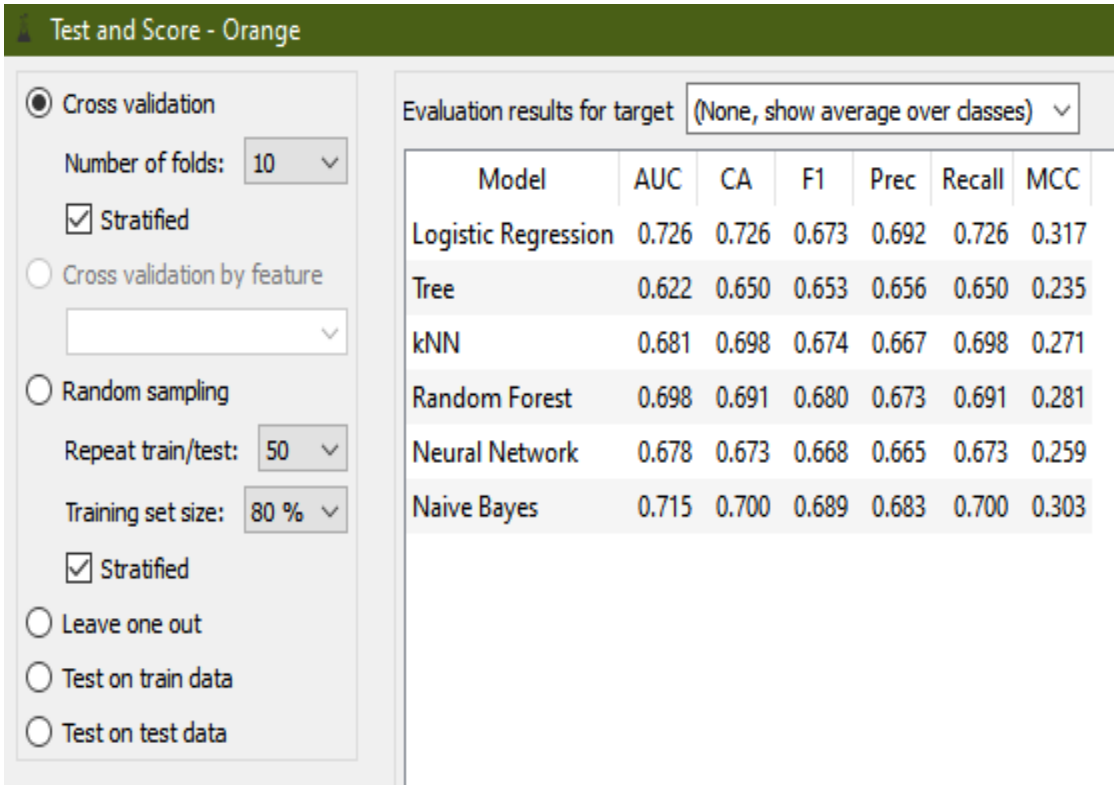


Figure 24: Result of testing classification algorithms.

The widget does two things:

First, it shows a table with different classifier performance measures, such as classification accuracy and area under the curve.

Second, it outputs evaluation results, which can be used by other widgets for analyzing the performance of classifiers, such as ROC Analysis or Confusion Matrix.

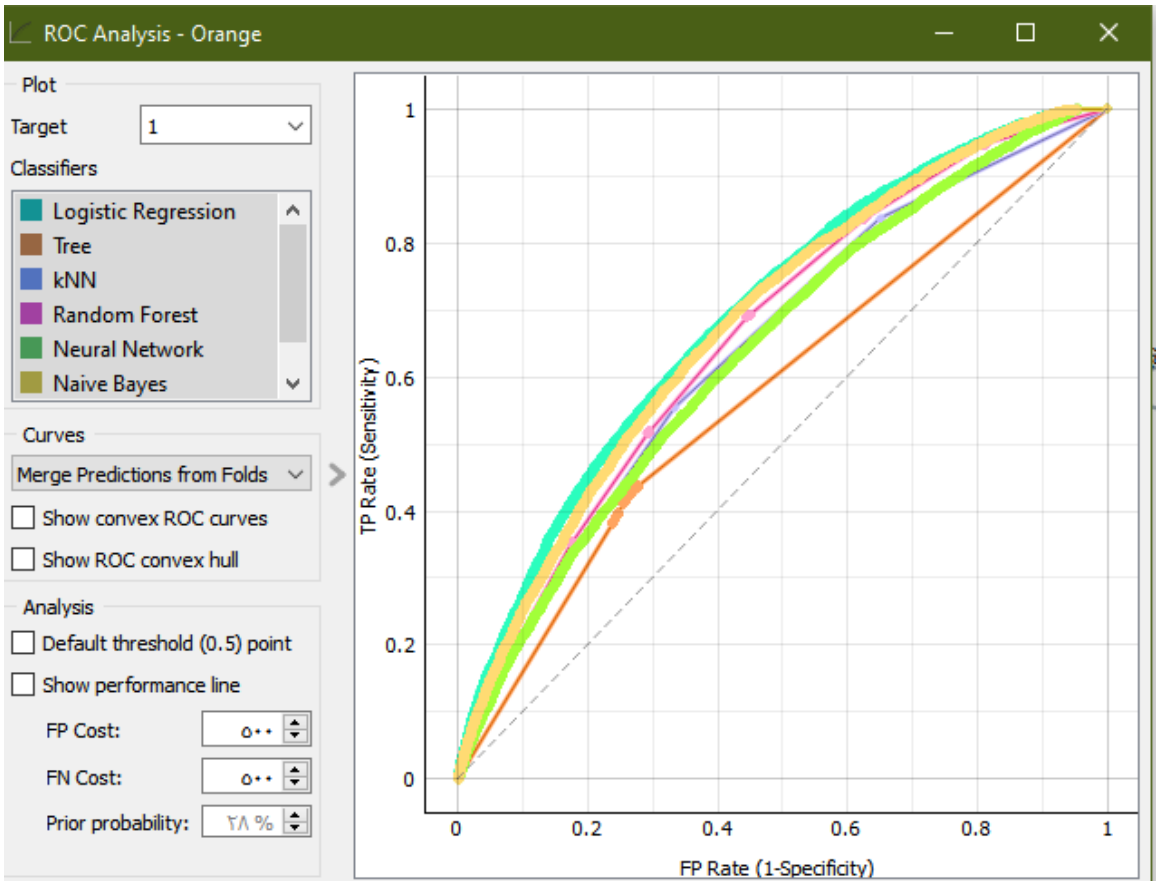


Figure 25: Plots a true positive rate against a false positive rate of a test.

The widget shows ROC curves for the tested models and the corresponding convex hull. It serves as a mean of comparison between classification models. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the classifier.

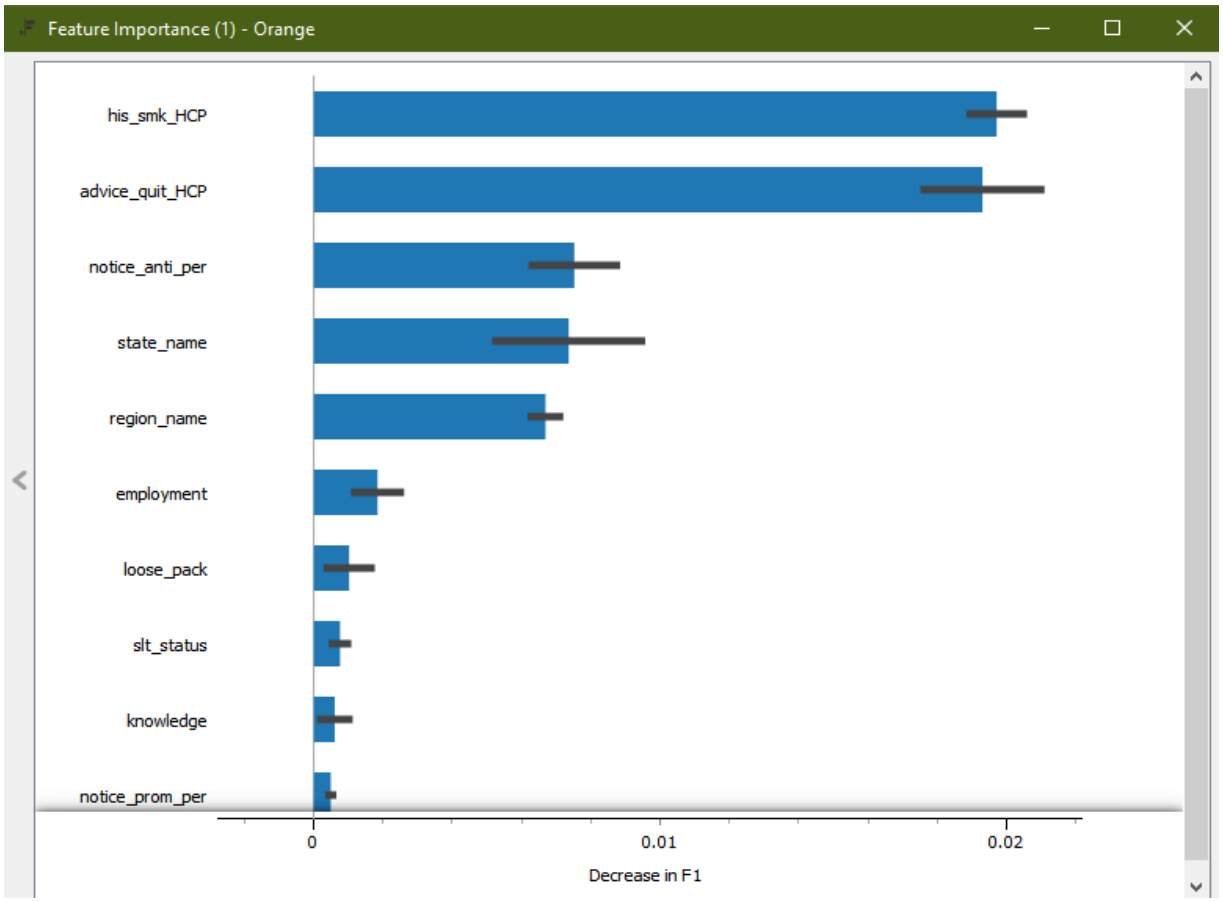


Figure 26: Result of features importance for logistic regression algorithm.

The previous widget shows ranking the importance and impact of factors on the category of smoker using the most accurate algorithm to predict.

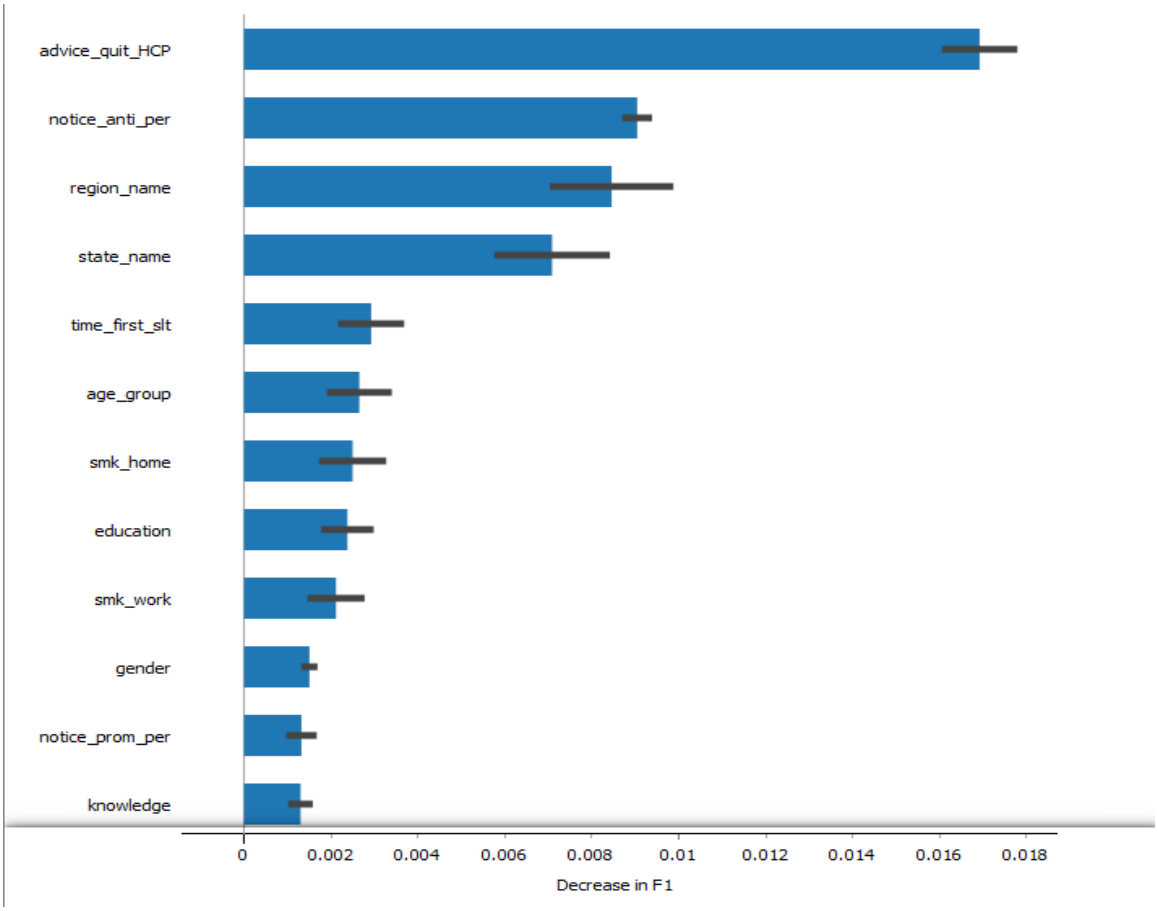


Figure 27: Result of features importance for logistic regression algorithm.

The previous widget shows ranking the importance and impact of factors on the category of smokeless users using the most accurate algorithm to predict.

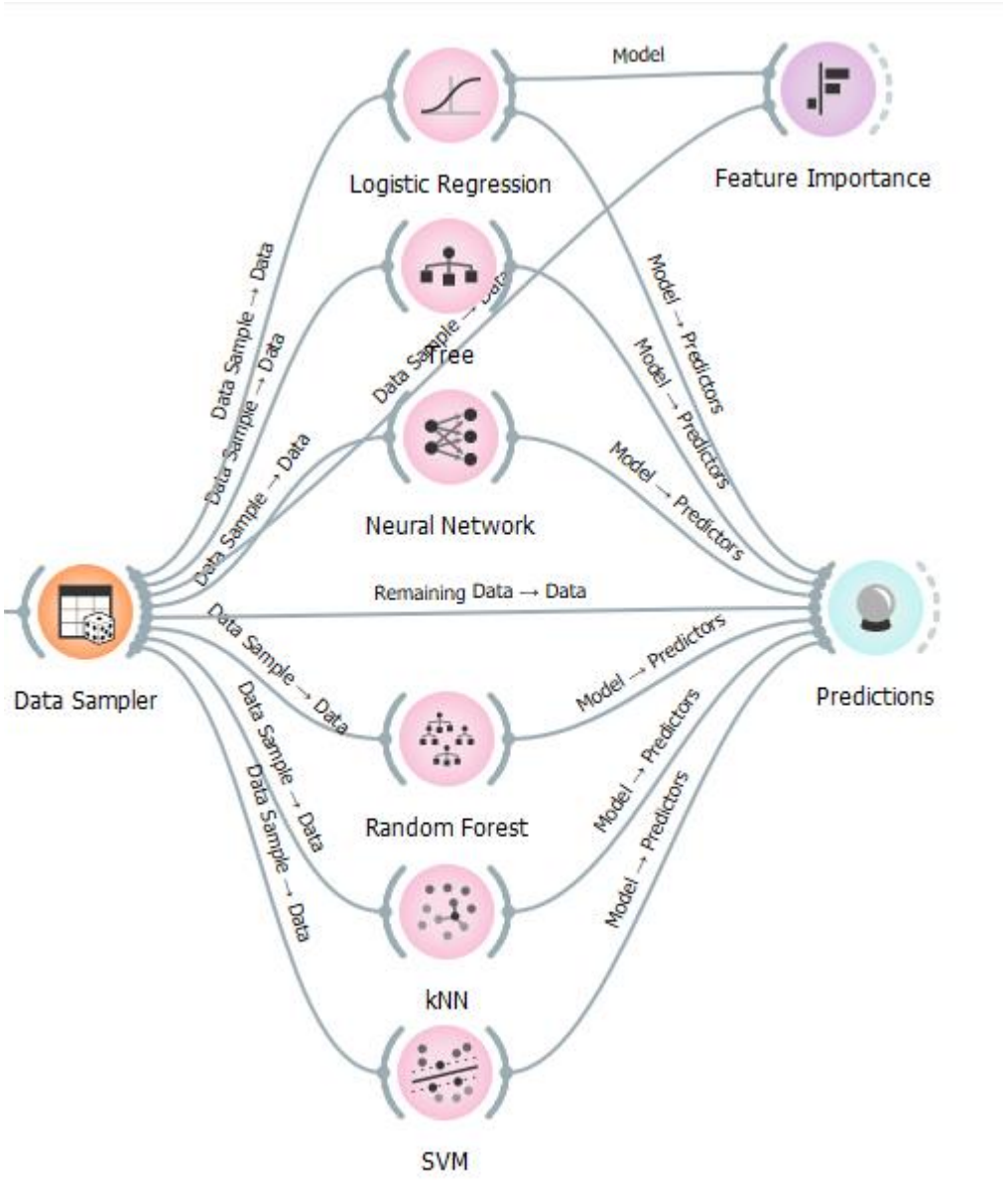


Figure 28: Prediction smoker type.

The previous figure shows a prediction the smoker type based on his characteristics used classification algorithms and in confusion matrix widget there are smoker whose type is wrongly predicted in each algorithm.

Predictions - Orange

Show probabilities for (None) Show classification errors

	Tree	Logistic Regression	Neural Network	Random Forest	kNN	SVM	smoke	smoking_status	quit_period	slt_status	residency
1	1	1	1	1	1	1	1	1	2	3	2
2	1	1	1	2	1	1	1	1	2	3	1
3	2	1	1	1	1	1	1	1	2	3	2
4	1	1	2	1	2	1	1	1	2	3	2
5	2	1	1	2	2	1	1	1	2	3	1
6	1	1	1	1	1	1	2	1	2	3	1
7	1	1	1	1	1	1	1	1	2	3	1
8	2	1	1	2	1	1	1	1	2	1	2
9	1	1	1	1	2	1	1	1	2	1	1
10	1	1	2	2	2	2	1	1	2	1	2
11	1	1	1	1	1	2	1	1	2	3	2
12	1	1	1	1	1	1	2	1	2	3	1
13	2	1	2	2	2	1	1	1	2	3	1
14	1	1	2	1	1	1	1	1	2	1	1
15	1	1	2	1	2	1	2	1	2	3	1
16	2	1	2	1	1	1	2	1	2	3	2
17	1	1	1	1	1	1	1	1	2	3	1
18	1	1	1	1	1	2	1	1	2	3	2
19	2	1	1	1	1	1	1	1	2	3	1

Figure 29: Result of prediction for each algorithm.

The previous figure shows us the result of the smoker category prediction for all classification algorithms and to be able to compare it with the original smoker category.

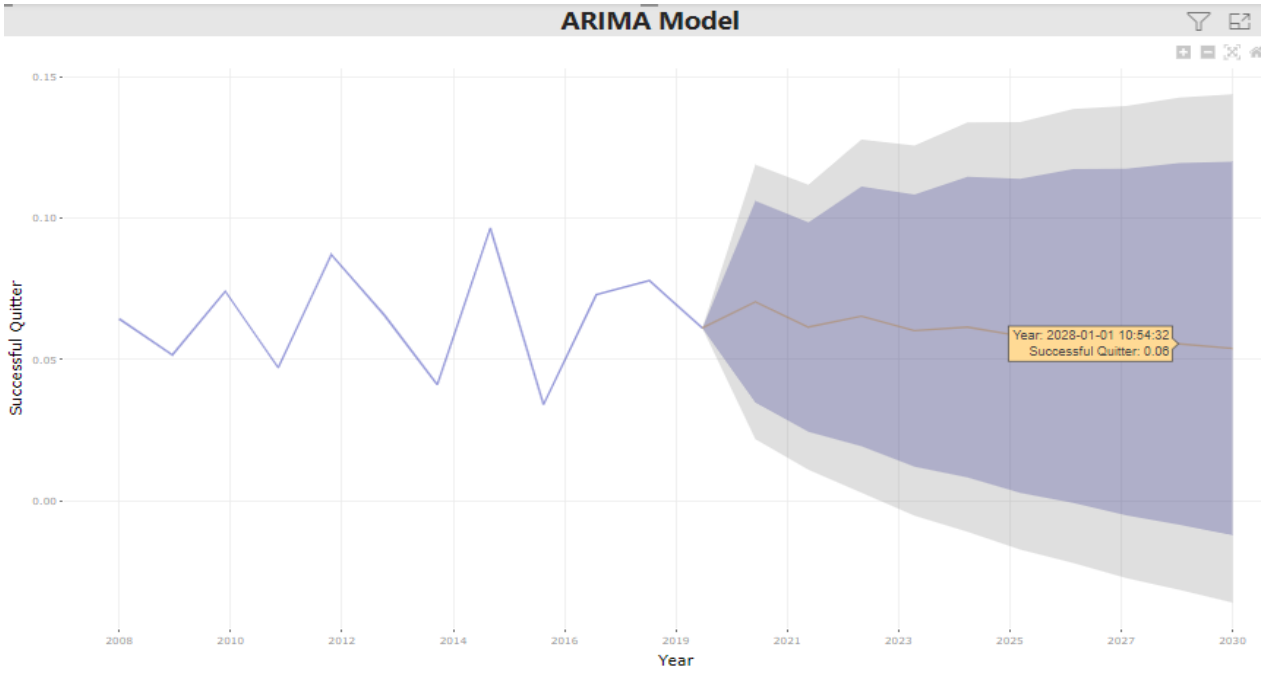


Figure 30: ARIMA Model.

The ARIMA model is a method for predicting a time series that exhibits more than one seasonal effect, it's forecast the percentage of successful quitters for smokers in the world.

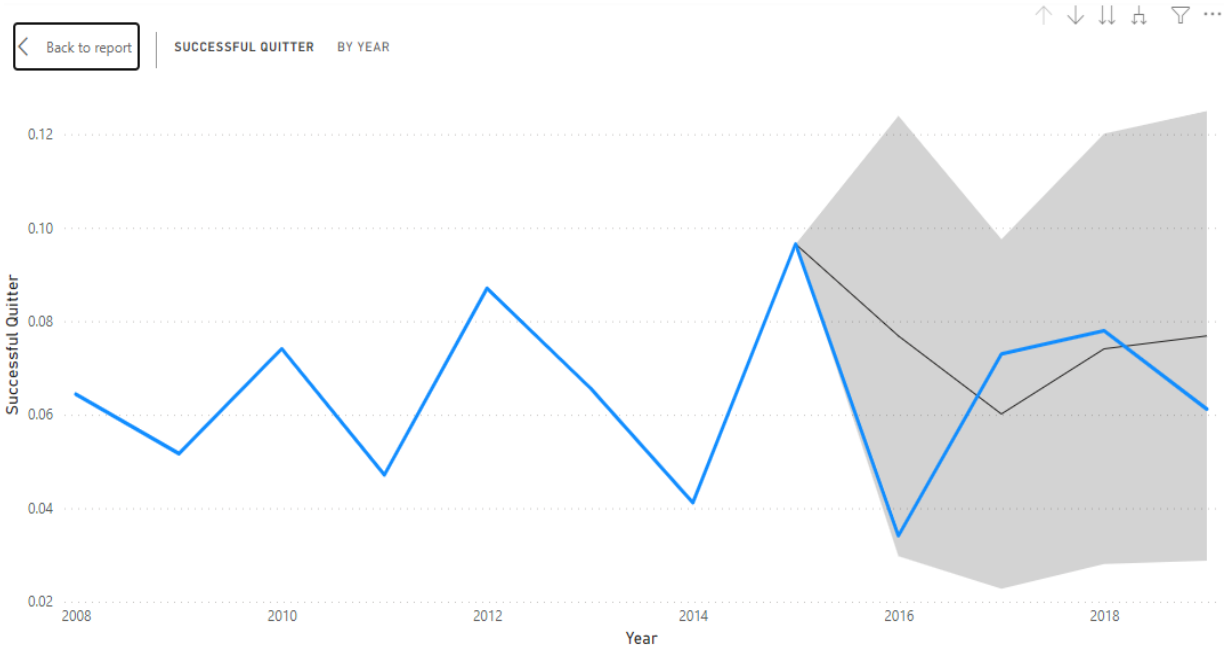


Figure 31: forecasting the percentage of successful quitters and compare with the actual value.

The previous figure shows an implementation of the exponential smoothing algorithm to forecast the percentage of successful quitters for smokers in the world and compare with the actual value for 2016 to 2019.

The highlighted area shows the upper and lower bound of the prediction.

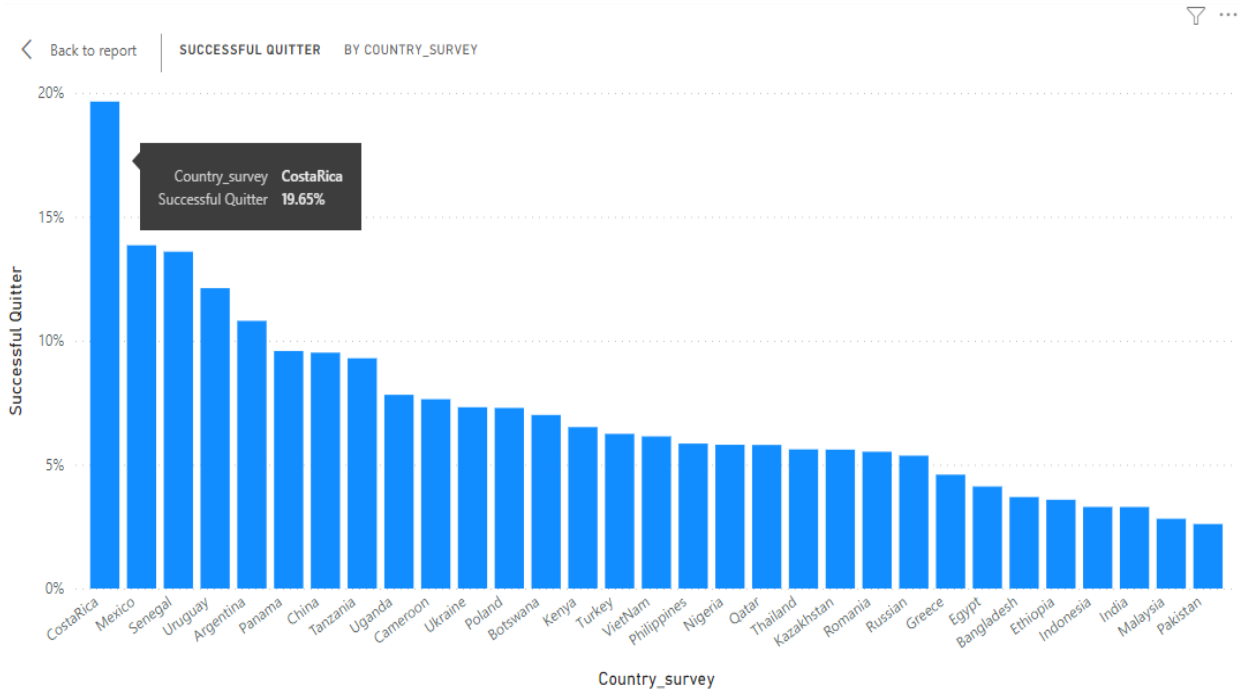


Figure 32: Comparison of the percentage of successful quitting between the countries of the world.

The previous report shows us Comparison of the percentage of successful quitting between the countries of the world.

Using the previous report, we find that Costarica country is contain the most percentage of successful quitting.

4. Conclusion & Future Expectations

4.1 Conclusion

After analyzing smokers' data, it can be said that smoking poses a great danger to human health and leads to many chronic diseases. Thus, raising awareness of the harmful effects of smoking and warning against it must be a priority for WHO, Smokers should also change their behavior and stop smoking to preserve their health and the health of their community.

It can be said that this project represents an important step in understanding the challenges smokers face in trying to quit smoking. Thus, this analysis helps to develop more effective strategies to help smokers quit smoking that not only target the consumer himself, but must also support him in all aspects of his life.

The classification results also showed that there are people who need special support to quit smoking, and this indicates the need to develop support programs dedicated to these people.

Through the National Tobacco Control Program, pan India, current cessation policies need to be revisited and strengthened, in particular, strategies to promote smoke-free home and work-place; fostering comprehensible anti-smoke/smokeless advertisements and label warnings; school and mass-media based education campaigns both to sensitive and educate on the health hazards of tobacco and benefits of quitting. Empowering the health care workers in India for tobacco cessation advocacy seems critical. The Indian government needs to strengthen the enforcement of the law, including surveillance of violations and strategies for the roll out of cessation services like counselling, quit-line and pharmacotherapy.

Finally, it can be said that this project represents an important contribution to understanding the challenges smokers face and developing effective strategies to quit smoking. Thus, this project helps to improve the health of smokers and improve their quality of life.

4.2 What is coming next?

Create a research paper with this project and develop it by taking youth data (GYTS).

5. References :

[1] Majd Al-ssabbagh, Varalakshmi Elango, Volker Winkler, " What makes people quit tobacco and succeed at it? An exploratory analysis of smoked and smokeless tobacco from India" Heidelberg Institute of Global Health, Heidelberg University, Germany,2022.

[2] K. Tsiptsis and A. Chorianopoulos, “Data Mining Techniques in CRM: Inside Customer Segmentation”, John Wiley and Sons, Ltd., Publication, 2009.

[3] The data for INDIA GATS 2009 and 2016 is available from CDC (https://nccd.cdc.gov/GTSSDataSurveyResources/Ancillary/DataReports.aspx?CAID=2).

[4] Ling Luo, Bin Li et. al., Tracking the Evolution of Customer Purchase Behavior Segmentation via a Fragmentation-Coagulation Process, Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

[5] Building the Data Warehouse (4th ed)2005Wiley Publishing, Inc.

[6] Sarkar, B.K., Arora, M., Gupta, V.K., et al., 2013. Determinants of tobacco cessation behavior among smokers and smokeless tobacco users in the states of Gujarat and Andhra Pradesh, India. Asian Pac. J. Cancer Prev. 14 (3), 1931–1935. (https://doi.org/10.7314/APJCP.2013.14.3.1931).

[7] Business intelligence1989.

[8] A Data Mining Approach to New Library Book Recommendations2002*E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555*229-240Kaohsiung 80424, Taiwan Springer