UNIVERSIDADE FEDERAL
DO RIO DE JANEIRO

CIÊNCIA DA
COMPUTAÇÃO
UFRJ

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GODSGOOD CHRIS CHINEDOZIE

EMERGING PATTERNS OF TUBERCULOSIS DRUG-RESISTANT BY DATA MINING WITH ASSOCIATION RULES (Apriori)

Using the R-library tbdr19prediction

RIO DE JANEIRO
2023

GODSGOOD CHRIS CHINEDOZIE

EMERGING PATTERNS OF TUBERCULOSIS DRUG-RESISTANT BY DATA MINING WITH ASSOCIATION RULES (Apriori)

Using the R-library tbdr19prediction

Graduate Final project presented to the Institute of computer science at the Federal University of Rio de Janeiro as obligated for obtaining the bacherol of science degree in Computer Science.

Supervisor: Rejane Sobrino Pinheiro
Co-supervisor: Valeria Menezes Bastos

RIO DE JANEIRO
2023

CIP - Catalogação na Publicação

```
        Chinedozie, Godsgood Chris
C539e      Emerging patterns of tuberculosis drug-resistant
        by data mining with association rules (Apriori):
        using the R-library tbdr19prediction / Godsgood
        Chris Chinedozie. -- Rio de Janeiro, 2023.
           74 f.

           Orientadora: Rejane Sobrino Pinheiro.
           Coorientadora: Valeria Menezes Bastos.
           Trabalho de conclusão de curso (graduação) -
        Universidade Federal do Rio de Janeiro, Instituto
        de Computação, Bacharel em Ciência da Computação,
        2023.

           1. Mineração de dados. 2. Tuberculose. 3. R
        (Linguagem de programação de computador). I.
        Pinheiro, Rejane Sobrino, orient. II. Bastos,
        Valeria Menezes, coorient. III. Título.
```

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos
pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

GODSGOOD CHRIS CHINEDOZIE

EMERGING PATTERNS OF TUBERCULOSIS DRUG-RESISTANT BY DATA MINING WITH ASSOCIATION RULES (Apriori)
Using the R-library tbdr19prediction

Graduate Final project presented to the Institute of computer science at the Federal University of Rio de Janeiro as obligated for obtaining the bacherol of science degree in Computer Science.

Approved on 22 of August of 2023

EXAMINATION BOARD:

gov.br Documento assinado digitalmente
REJANE SOBRINO PINHEIRO
Data: 20/09/2023 16:45:56-0300
Verifique em https://validar.iti.gov.br

Rejane Sobrino Pinheiro
D.Sc. (UFRJ)

gov.br Documento assinado digitalmente
VALERIA MENEZES BASTOS
Data: 21/09/2023 09:26:52-0300
Verifique em https://validar.iti.gov.br

Valeria Menezes Bastos
D.Sc. (UFRJ)

gov.br Documento assinado digitalmente
CLAUDIA MEDINA COELI
Data: 22/09/2023 14:07:01-0300
Verifique em https://validar.iti.gov.br

Claudia Medina Coeli
D.Sc. (UFRJ)

gov.br Documento assinado digitalmente
JULIANA BAPTISTA DOS SANTOS FRANCA
Data: 19/09/2023 17:55:39-0300
Verifique em https://validar.iti.gov.br

Juliana Baptista dos Santos França
D.Sc. (UFRJ)

ACKNOWLEDGEMENTS

I thank God almighty and the *Chinedozie's* family for their steadfast love and support, I thank Rejane Sobrino Pinheiro and Valeria Menezes Bastos for supervising and guarding me throughout the whole project, the project was developed at the laboratory Linkage and Data Analysis Population at the Institute of Studies in Collective Health at UFRJ (LinkDataPop) with funds by the Bill & Melinda Gates Foundation.

ABSTRACT

Tuberculosis (TB) is one of the top 10 causes of death in Brazil, approximately 70,000 new cases are reported each year and there are approximately 4,500 deaths due to tuberculosis. Drug resistance in Mycobacterium tuberculosis arises from spontaneous chromosomal mutations at low frequency, clinical tuberculosis drug-resistant (TBDR) largely occurs as a result of man-made selection during disease treatment of these genetic alterations through erratic drug supply, suboptimal physician prescription and poor patient adherence. In light of this, it became necessary to develop a data mining algorithm to produce predictive models, aiming to analise emerging patterns.

To achieve this purpose, the implementation of an association rule algorithm with apriori was the most suitable choice, allowing the identification of patterns linked to unfavourable outcomes like abandonment, failed treatment and death. Furthermore, the valid rules obtained from the predictive model were explored and analised, providing new knowledge to identify significant items of vulnerable patients for each unfavourable outcome.

Keywords: tuberculosis; drug resistant; apriori; data mining; r-language; models.

RESUMO

A tuberculose (TB) é uma das 10 principais causas de morte no Brasil, aproximadamente 70.000 novos casos são notificados a cada ano e há aproximadamente 4.500 mortes por tuberculose. A resistência a medicamentos em Mycobacterium tuberculosis surge de mutações cromossômicas espontâneas em baixa frequência, a tuberculose droga resistente (TBDR) clínica ocorre em grande parte como resultado da seleção feita pelo individuo durante o tratamento da doença dessas alterações genéticas por meio de fornecimento errático de medicamentos, prescrição médica abaixo do ideal e paciente pobre aderência. Diante disso, tornou-se necessário desenvolver um algoritmo de mineração de dados para produzir modelos preditivos, visando analisar padrões emergentes.

Para atingir esse objetivo, a implementação de um algoritmo de regra de associação com a apriori foi a escolha mais adequada, permitindo a identificação de padrões ligados a desfechos desfavoráveis como abandono, falência e morte. Além disso, as regras válidas obtidas do modelo preditivo foram exploradas e analisadas, fornecendo novos conhecimentos para identificar itens significativos de pacientes vulneráveis para cada desfecho desfavorável.

Palavras-chave: tuberculose; droga resistante; apriori; mineração de dados; r-linguagem; modelos.

LIST OF CÓDIGOS

código 1 Apriori ((AGRAWAL; SRIKANT et al., 1994)) 32

código 2 Apriori Function Call with Parameters. 52

LIST OF TABLES

Table 1 – The list of items purchased per transaction 31

Table 2 – Frequent Itemsets 31

Table 3 – Dictionary for the selected variables 42

Table 4 – Compilation of Various configurations of Support and Confidence 53

Table 5 – Main Compilation: SUPP[0.05] CONF[0.05] 54

Table 6 – Nº Of Items x Nº of rules 55

Table 7 – Nº Of Items x Nº of rules (Abandon) 56

Table 8 – Comparing Abandonment to Failed & Death 58

Table 9 – 4 Unique Profiles with $LIFT \geq 4.8$ 58

Table 10 – Nº Of Items x Nº of rules 59

Table 11 – Comparing Failed to Abandonment & Death 61

Table 12 – 3 Unique Profiles with $LIFT \geq 3.9\&Confidenceof33\%$ 61

Table 13 – Nº Of Items x Nº of rules 62

Table 14 – Comparing Death to Abandonment & Failed 63

Table 15 – 3 Unique Profiles with $LIFT \geq 4.0$ 64

LIST OF FIGURES

Figure 1 – A Snapshot Of The R Studio 34
Figure 2 – Run R scripts from the command line 34
Figure 3 – R-Studio IDE Cheat Sheet for Keyboard Shortcut 36
Figure 4 – Analised Variables: Characterization Model 42
Figure 5 – The KDD process (Fayyad et al., 1996) 45
Figure 6 – Divide the DB into 34% Test and 66% Train 46
Figure 7 – Random division into Test/Train 47
Figure 8 – Mining for Coincidents from 3 sets of Valid Rules. 47
Figure 9 – Optimizing with K-number of samples. 48
Figure 10 – Random division into Test/Train 49
Figure 11 – Extracting Outcomes from Coincidents "Cured - Abandonment - Failed - Death". 49
Figure 12 – Intersection of Common Rules 53
Figure 13 – Plot: N^{o} of Rules in relations to Items leading to Cure 55
Figure 14 – Significant Items: Total x weight (Cure) 55
Figure 15 – Plot: N^{o} of Rules in relations to Items leading to Abandonment 56
Figure 16 – Significant Items: Total x weight (Abandonment) 57
Figure 17 – Plot: N^{o} of Rules in relations to Items leading to Failed 59
Figure 18 – Significant Items: Total x weight (Failed) 60
Figure 19 – Plot: N^{o} of Rules in relations to Items leading to Death 62
Figure 20 – Significant Items: Total x weight (Death) 63

LIST OF ABBREVIATIONS AND ACRONYMS

TB	Tuberculosis
TBDR	Tuberculosis Drug-resistant
MDR	Multidrug-resistant
RR	Rifampicin resistant
SUS	Unified Public Health System
SITE-TB	Special Tuberculosis Treatment Information System
CRAN	Comprehensive R Archive Network
JICTAC	Jornada De Iniciação Científica, Tecnológica, Artística e Cultural
KDD	Knowledge Discovery in Databases
WHO	World Health Organization
SINAN	Information system on notifiable diseases
SIM	Mortality Information System
Supp	Support
Conf	Confidence
MoH	Ministry of Health
LHS	Left Hand Side
RHS	Right Hand Side
Lfx	Levofloxacin
Mfx	Moxifloxacin
Ofx	Ofloxacino

CONTENTS

1 PRESENTATION . . . 13

1.1 BACKGROUND . . . **13**

1.2 PURPOSE . . . **14**

1.3 OBJECTIVES . . . **14**

1.4 METHODOLOGY . . . **14**

1.5 TCC ORGANIZATION . . . **15**

2 INTRODUCTION . . . 16

2.1 TUBERCULOSIS DRUG-RESISTANT (TBDR) . . . **16**

2.1.1 Impact and significance of tuberculosis . . . 17

2.1.1.1 Global Impact: . . . 17

2.1.1.2 Socio-economic Impact: . . . 17

2.1.1.3 Co-infection with HIV: . . . 17

2.1.1.4 Drug Resistance: . . . 17

2.1.1.5 Impact on Health Systems: . . . 18

2.1.1.6 Stigma and Discrimination: . . . 18

2.1.2 Definition and types of Tuberculosis Drug-Resistance (TBDR) 18

2.1.3 Causes of Tuberculosis Drug Resistance . . . 19

2.1.4 Mechanisms of tuberculosis Drug Resistance . . . 20

2.1.5 Epidemiology and Global Burden . . . 21

2.1.6 Diagnostic Methods for Drug-Resistant Tuberculosis . . . 22

2.1.7 Treatment Options for Drug-Resistant Tuberculosis . . . 23

2.1.8 Challenges in the Control and Management of Drug-Resistant Tuberculosis . . . 25

2.1.9 Prevention Strategies and Control Measures for Drug-Resistant Tuberculosis . . . 27

2.2 ASSOCIATION RULES AND APRIORI . . . **29**

2.2.1 Support . . . 29

2.2.2 confidence . . . 30

2.2.3 The Apriori Algorithm . . . 32

2.3 UNDERSTANDING R AND R-STUDIO BASIC INSTRUCTIONS . . . **33**

2.3.1 Why R ? . . . 33

2.3.2 The R Studio . . . 34

2.3.2.1 The Console . . . 34

2.3.2.2 The Scripts . . . 35

2.3.2.3 The Panes . . . 35
2.3.2.4 The Key Bindings . . . 35
2.3.2.5 Running commands while editing script . . . 35
2.3.2.6 Changing global options . . . 35
2.3.2.7 Installing R packages . . . 36
2.4 BUILDING A R-PACKAGE USING DEVELOPMENT TOOLS . . . **37**
2.4.1 Setting Up Your Development Environment: . . . 37
2.4.2 Creating a New Package: . . . 37
2.4.3 Defining the Package Structure: . . . 38
2.4.4 Documenting Your Package: . . . 38
2.4.5 Building Your Package . . . 38
2.4.6 Checking Your Package: . . . 39
2.4.7 Adding Functions: . . . 39
2.4.8 Testing Your Package . . . 39

3 METHODOLOGY . . . 41
3.1 DATABASE SOURCE, SELECTION AND DICTIONARY: . . . **41**
3.1.1 Database Source and Study Population: . . . 41
3.1.2 Dictionary for the Selected Variables . . . 42
3.2 THE MINING TECHNIQUE . . . **44**
3.2.1 Algorithm for Valid Rules using KDD . . . 45
3.2.2 The Mining Technique: Bootstrap and Bagging . . . 46
3.2.3 The Mining Techniques: Test/Train and Apriori . . . 46
3.3 THE "TBDR19PREDICTION" R-LIBRARY . . . **49**
3.3.1 Installation . . . 49
3.3.1.1 Features and functions . . . 50

4 RESULTS . . . 52
4.1 CONFIGURATION OF METRICS . . . **52**
4.1.1 Preliminary Results . . . 53
4.2 CONFIGURATION: SUPP[0,05], CONF[0,05] . . . **54**
4.2.1 Favourable Results: Cure . . . 54
4.2.2 Unfavourable Results: Abandonment . . . 56
4.2.3 Unfavourable Results: Failed . . . 58
4.2.4 Unfavourable Results: Death . . . 61
4.3 CONFIGURATION: SUPP[0,01], CONF[0,1] . . . **65**
4.3.1 Failed . . . 65
4.3.2 Death . . . 66
4.3.3 Abandonment . . . 66

5 CONCLUSION 69

. . . . **69**

BIBLIOGRAPHY 70

GLOSSARY 73

1 PRESENTATION

This course completion work is related to a subproject ("Data Science in the Control of Drug-Resistant Tuberculosis") belonging to a broader project ("Data Science in Public Health: linkage, data mining and machine learning in a perspective of Academic Health Secretariat"), which received funding from the National Council for Scientific and Technological Development - CNPQ and the Bill & Melinda Gates Foundation.

Despite the treatment of tuberculosis (TB) being free, and its cure being highly effective, the number of cases whose treatment outcomes is unfavorable is still high. Drug-Resistance (DR) is an additional barrier to disease control. It is important to know the factors that lead to abandonment, Failed and Death in order to develop effective actions. Machine learning techniques can help to understand the special profiles of patients at higher risk. The aim of the study was to identify, using association rules, profiles of patients with TBDR in Brazil and each unfavorable treatment outcomes.

The Focus of this project is to develop an exploratoy mechanism to predict results and outcomes of patients suffering from Tuberculosis Drug Resistant through data mining models. This project was presented to 2 editions of the Jornal of Scientific iniciation (JICTAC, 2019, 2021) and winning honourable mentions in 2021.

The results and outcomes are used to analise new patients suffering from Tuberculosis Drug Resistant, certain characteristics of a patient can highlight the possible outcome of the patient and with guarded diagnostics or treatment alter the result to a favourable outcome.

1.1 BACKGROUND

In Brazil, during the year 2013 and 2014, there were reports of different treatment outcomes caused by a new pulmonary tuberculosis (TB) cases of multidrug-resistant or extensively resistant, rifampicin-resistant. The database used in this project was obtained for this year from the Special Tuberculosis Treatment Information System (SITE-TB).

There is a research study nation wide and internationally showcasing that death bt TB occurs prematurely to patients, even during the two months referred to as intensive treatment phase, when compared to other causes of death (JONNALAGADA; HARRIES; ZACHARIAH, 2011; RODRIGO et al., 2016)

1.2 PURPOSE

Understanding whether different patterns of people who died, abandoned treatment or treatment failure from tuberculosis (TB) can help enable a better effectiveness of health care for specific group of people.

1.3 OBJECTIVES

To develop an algorithm for generating valid rules from the Apriori algorithm and apply it to a dataset of patients with drug-resistant tuberculosis in Brazil.
Identifying the patterns connected to various outcomes, using data mining tools and techniques, we gathered new informations and knowledges that is better and different from methods obtained from traditional statistical approach.

The objectives of this project was to build an analysis that is an evidence that there are factors related to unfavourable outcomes *(death, failed treatment and abandonment)* (BARTHOLOMAY, 2019). For death, elaborating issues related to individual, clinical and treatment characteristics, and follow-up treatment. For failed treatment or failure, we explore variables of clinical characteristics, proposed treatment and follow-up treatment. For abandonment, we elaborated variables from in individual, clinical and treatment characteristics, and follow-up treatment.

Exploring all four outcomes (favourable and un-favourable) separately, through the association rule data mining algorithm, determine unique profiles and items that stoodout the most and deterministic to that outcome.

1.4 METHODOLOGY

The SITE-TB was the main source for the database (BARTHOLOMAY et al., 2019), The history of previous TB treatment before the notification of SITE-TB was retrieved from the Information System on Notifiable Diseases (Sinan) and the resulting treatment outcomes of patients profiles were qualified with TB death records from Mortality Information System (SIM), through the probabilistic relationship of the database (ROCHA.MS et al., 2019).

We used a compiled database of 1000 registers and selected 15 variables through individual, clinical and treatment characteristics as shown in Figure 4. From the register, we have 64% for Cure, 16,5% Abandonment, 7,7% Failed and 12% Death. Using R programming language (R-PROGRAMMING-LANGUAGE, 2023), we developed a data mining model by association rules (Apriori) and applied the bootstrap and Bagging technique in chapter 3.2.2, creating this model using apriori we generated a refined dataset we called coincident as shown in Figure 8, having antecedents and consequents values in our coincident dataset, we explored different results by filtering-out by outcomes as shown in Figure 11.

1.5 TCC ORGANIZATION

We started working on this project in 2019, since then different results has been discovered, submitted and presented in 2 editions of the Journal of Scientific Iniciation (JICTAC, 2020), and also winning honourable mentions in 1 occasions.

A broad study was done in Brazil focusing mainly by the effect of TBDR in major cities like Rio de Janeiro and São Paulo. Study materials from the following articles (BARTHOLOMAY, 2019; RS, 2018 Sep 7; BARTHOLOMAY et al., 2019).

This work is organized into items/chapters. In item/chapter 2, a review is made of tuberculosis and drug-resistant tuberculosis in a section, with emphasis on epidemiology, causes, types, treatments, tests, outcomes and other issues related to the diseases. In section 2.2 of chapter 2, the Apriori algorithm is presented. In section 2.3 of chapter 2, the R language is presented, used in the development of the algorithm for generating valid rules. In item/chapter 3, the methodology for developing the valid rules algorithm is presented, the description of the database and the variables is presented, the TBDR library is described. In item/chapter 4, the result of the model is presented with exploratory analysis.

2 INTRODUCTION

In this introduction, we discussed 4 important topics that helps us understand the problem faced with Tuberculosis illness, how we used association rules to create transactions and relationships between each register and its outcome, apriori algorithm was important and promising, we generated rules using apriori and studying this rules results to founding new knowledge and information about this study.
In this chapter, we discussed about the R programing language and the R-studio, these softwares are one of the most used technologies in data science and machine learning.

2.1 TUBERCULOSIS DRUG-RESISTANT (TBDR)

Tuberculosis (TB) is an infectious disease caused by the bacteria Mycobacterium tuberculosis. As stated in (W.H.O, 2022) despite significant advancements in TB treatment, the emergence of drug-resistant strains has become a pressing global health issue. This work aims to explore the various aspects of tuberculosis drug resistance treatment outputs, including the relation to sociodemographic characteristics of the patient, the severity of the disease, presence of comorbidities, soem health behaviour, type of resistant, treatment location and the challenges it poses to global TB control efforts. Additionally, we will discuss the importance of early detection strategies and treatment focus on patient in high risk of negative treatment output, ongoing research efforts in combating drug-resistant tuberculosis.

Tuberculosis Drug-Resistant (TBDR) is a serious threat to the control of tuberculosis in Brazil and worldwide.(GLYNN et al., 1995) Tuberculosis is a disease that has existed for many years and its transformation to drug-resistant is due to several factors that occur during the treatment phase, with abandonment being one of the major causes of the change from TB to TB-DR. And, despite the fact that tuberculosis treatment is free and highly effective, the number of cases whose outcome is death is still very high.

The factors that lead to healing and, mainly, abandonment, failed treatment and death can be explored in different ways, so that health agencies can know their causes and effectively apply combat methods, thus reducing the occurrence of cases. Some of them may suggest the cause of alarms based on the patient's profile, at the time of notification, both for the case of TB and for its evolution to drug-resistant. The application of machine learning methods can help in understanding the behavior of patients and identifying patients at higher risk, in order to improve the control of TBDR in the country. We identified the profiles of patients with TBDR at greater risk of unfavorable outcomes in Brazil in order to support the development of actions and the construction of classifiers and alarms for special follow-up.

2.1.1 Impact and significance of tuberculosis

Given the significant impact of tuberculosis on public health systems, economies, and societies in Brazil and the whole world (BARTHOLOMAY, 2019), global efforts have been undertaken to combat the disease. International organizations, governments, and stakeholders collaborate to improve diagnostics, develop new drugs and vaccines, enhance healthcare infrastructure, promote treatment adherence, and raise awareness about TB prevention. Eradicating tuberculosis requires sustained commitment, resources, and a comprehensive approach to address its social, economic, and medical dimensions.

2.1.1.1 Global Impact:

TB is one of the top 10 causes of death worldwide and the leading cause of death from a single infectious agent, surpassing HIV/AIDS. According to the World Health Organization (WHO) (W.H.O, 2022), an estimated 10 million people fell ill with TB in 2020, and approximately 1.5 million people died from the disease.

2.1.1.2 Socio-economic Impact:

TB disproportionately affects low- and middle-income countries, where poverty, overcrowding, malnutrition, and limited access to healthcare contribute to its spread (CONTROL; (CDC), 2023). The disease exacerbates existing social and economic disparities, as it primarily affects marginalized populations, including those living in poverty, migrants, prisoners, and people with weakened immune systems.

2.1.1.3 Co-infection with HIV:

TB and HIV/AIDS form a deadly combination, as individuals with weakened immune systems due to HIV are at a higher risk of developing active TB. The synergistic relationship between TB and HIV/AIDS has contributed to the global TB epidemic, particularly in sub-Saharan Africa, where both diseases are highly prevalent.

2.1.1.4 Drug Resistance:

The emergence of drug-resistant TB poses a significant challenge to TB control efforts. Multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB) strains are resistant to the most potent first-line anti-TB drugs, making treatment more complex, costly, and less effective. This further complicates the management of TB cases and increases the risk of transmission.

2.1.1.5 Impact on Health Systems:

TB places a substantial burden on healthcare systems, including diagnostic resources, treatment facilities, and skilled healthcare professionals. The costs associated with TB prevention, diagnosis, and treatment, along with the loss of productivity due to illness and death, strain healthcare budgets and hinder socio-economic development.

2.1.1.6 Stigma and Discrimination:

TB is often associated with social stigma and discrimination, leading to delayed diagnosis, treatment non-adherence, and isolation of affected individuals. The stigma can prevent individuals from seeking healthcare services, resulting in increased transmission and poor treatment outcomes.

2.1.2 Definition and types of Tuberculosis Drug-Resistance (TBDR)

This project has its focus on the impact of drug-resistant in Tuberculosis. This resistance can occur due to various factors, such as improper or incomplete treatment, inadequate drug supply, or improper use of antibiotics. Tuberculosis drug resistance refers to the ability of the bacteria that cause tuberculosis (Mycobacterium tuberculosis) to withstand the effects of one or more anti-tuberculosis drugs. There are two primary types of tuberculosis drug resistance (KJ; S; ML, 2015):

1. **Multidrug-Resistant Tuberculosis (MDR-TB):** MDR-TB is a form of tuberculosis that is resistant to at least two of the most potent first-line drugs used for TB treatment, namely isoniazid and rifampicin. These drugs are considered the backbone of tuberculosis treatment, and resistance to them significantly complicates the management of the disease.

2. **Extensively Drug-Resistant Tuberculosis (XDR-TB):** XDR-TB is an even more severe form of drug resistance. It is resistant to both isoniazid and rifampicin, as well as to any fluoroquinolone and at least one of the three injectable second-line drugs (amikacin, kanamycin, or capreomycin). XDR-TB is challenging to treat, as it limits the availability of effective drugs and requires more prolonged and complex treatment regimens.

Apart from MDR-TB and XDR-TB, there is also a category called Pre-Extensively Drug-Resistant Tuberculosis (Pre-XDR-TB) (S. et al., 2018). Pre-XDR-TB refers to strains that are resistant to isoniazid, rifampicin, and any fluoroquinolone or one of the three injectable second-line drugs, but not both. Pre-XDR-TB is an intermediate level of resistance between MDR-TB and XDR-TB.

It's worth noting that extensively drug-resistant and multidrug-resistant tuberculosis pose significant challenges to tuberculosis control programs due to the limited treatment options available, increased risk of treatment failure, and higher mortality rates compared to drug-sensitive tuberculosis.

2.1.3 Causes of Tuberculosis Drug Resistance

Tuberculosis (TB) drug resistance occurs when the bacteria that cause TB develop resistance to one or more of the drugs used to treat the disease (KA et al., 2017). There are two primary types of TB drug resistance: multidrug-resistant tuberculosis (MDR-TB) and extensively drug-resistant tuberculosis (XDR-TB). (A; AJ; CA, 2006) The causes of TB drug resistance can be attributed to several factors, including:

1. **Inadequate or Inappropriate Treatment**: One of the primary causes of drug resistance is the improper use of anti-TB drugs. This includes incorrect dosages, incomplete treatment courses, or irregular medication adherence. When the drugs are not taken as prescribed, the bacteria can survive and develop resistance to the drugs.

2. **Mismanagement of TB Programs:** Weak healthcare infrastructure, inadequate diagnostic capabilities, and poor management of TB treatment programs can contribute to drug resistance. Limited resources, lack of trained healthcare personnel, and inefficient monitoring systems can all lead to suboptimal treatment practices and increase the risk of drug resistance.

3. **Improper Diagnosis:** Delayed or incorrect diagnosis of TB can contribute to the development of drug resistance. When TB is not diagnosed promptly or accurately, the bacteria can continue to multiply, and the infection can progress, increasing the likelihood of drug resistance.

4. **Transmission of Drug-Resistant Strains:** When individuals with drug-resistant TB transmit the bacteria to others, drug-resistant strains can spread within communities. Factors that facilitate the transmission of drug-resistant strains include crowded living conditions, inadequate infection control measures in healthcare facilities, and HIV co-infection, which weakens the immune system and makes individuals more susceptible to drug-resistant TB.

5. **Poor Quality of Medications:** Low-quality or counterfeit anti-TB drugs may not contain the necessary active ingredients or may have inadequate potency. Sub-

standard drugs can contribute to treatment failure and the development of drug resistance.

6. **HIV Co-infection:** HIV weakens the immune system, making individuals more susceptible to TB infection and increasing the risk of developing drug resistance. People living with HIV who are not on antiretroviral therapy (ART) or have suboptimal HIV control are more likely to develop drug-resistant TB.

7. **Overuse and Misuse of Antibiotics:** Inappropriate use of antibiotics in the general population and in healthcare settings can contribute to the emergence of drug-resistant strains of TB. Overuse of antibiotics for non-TB respiratory infections or misuse of broad-spectrum antibiotics can create selective pressure that promotes the growth of drug-resistant bacteria.

Preventing and managing drug resistance requires a comprehensive approach that includes appropriate diagnosis, standardized treatment regimens, adherence support, infection control measures, and ongoing surveillance of drug resistance patterns (D; AL, 2019).

2.1.4 Mechanisms of tuberculosis Drug Resistance

The development of drug resistance in tuberculosis (TB) is primarily driven by genetic mutations within the bacteria responsible for the disease, Mycobacterium tuberculosis (ZHANG; WW, 2009). These mutations can occur spontaneously or can be acquired through genetic exchange with other drug-resistant strains. Here are some of the key mechanisms of TB drug resistance:

1. **Target Gene Mutations:** Anti-TB drugs work by targeting specific proteins or enzymes involved in the replication or metabolism of M. tuberculosis (ZUMLA; P, 2013). Mutations in the genes encoding these targets can alter the structure or function of the target proteins, reducing the binding affinity of the drugs and rendering them ineffective. For example, mutations in the genes rpoB, katG, and inhA are associated with resistance to the first-line drug rifampicin, isoniazid, and ethionamide respectively.

2. **Efflux Pump Overexpression:** Efflux pumps are transport proteins that actively pump drugs out of the bacterial cell, reducing their intracellular concentration. Some drug-resistant strains of M. tuberculosis can overexpress efflux pumps, leading to decreased drug accumulation within the bacterial cells and reduced drug efficacy.

3. **Enzyme Inactivation:** Certain drugs require activation by enzymes produced by M. tuberculosis in order to become active against the bacteria. Resistance can occur

if mutations inactivate these enzymes, rendering the drugs ineffective. For instance, mutations in the ethA gene can lead to inactivation of ethionamide, and mutations in the pncA gene can result in resistance to pyrazinamide.

4. **Bypass Pathways:** In some cases, drug-resistant strains can develop alternative metabolic or biochemical pathways that bypass the target proteins or enzymes inhibited by the drugs. This allows the bacteria to survive and multiply in the presence of the drugs. This mechanism is seen in resistance to drugs like para-aminosalicylic acid (PAS) and cycloserine.

5. **Genetic Transfer:** Tuberculosis has the ability to acquire resistance genes from other bacteria through horizontal gene transfer. This can occur through processes like conjugation or transduction, where genetic material containing drug resistance mutations is transferred from one bacterium to another. This mechanism can contribute to the spread of drug resistance within populations.

It's important to note that these mechanisms can contribute to resistance against both first-line and second-line anti-TB drugs. Additionally, the development of multidrug-resistant tuberculosis (MDR-TB) and extensively drug-resistant tuberculosis (XDR-TB) usually involves a combination of multiple resistance mechanisms.

Understanding the mechanisms of drug resistance is crucial for the development of new drugs and treatment strategies to combat resistant strains of TB.

2.1.5 Epidemiology and Global Burden

Tuberculosis drug resistance is a global health concern affecting numerous countries and regions worldwide. The World Health Organization (WHO) estimates that in 2020 (W.H.O, 2022), there were approximately 465,000 new cases of MDR-TB, and an additional 248,000 cases of rifampicin-resistant TB. Furthermore, about 6.2% of these cases were classified as XDR-TB.

The burden of drug-resistant TB is particularly high in several countries, including India, China, Russia, and South Africa. Factors contributing to the high prevalence of drug-resistant TB in these regions include inadequate access to quality healthcare, poor infection control measures, and limited resources for diagnostic testing and treatment.

The impact of tuberculosis drug resistance extends beyond individual patients and has severe implications for public health and healthcare systems. First and foremost, drug-resistant TB is more difficult and expensive to treat than drug-susceptible TB, requiring longer treatment durations and more potent and costly medications (ZHANG;

WW, 2009). This places a tremendous financial burden on healthcare systems and individuals, often resulting in limited access to appropriate treatment for those in need.

Furthermore, drug-resistant TB is associated with higher rates of treatment failure, relapse, and mortality compared to drug-susceptible TB. These unfavorable treatment outcomes lead to increased healthcare costs and, more importantly, the loss of valuable human lives. The longer duration of treatment also increases the risk of patient non-adherence, potentially contributing to the further spread of drug-resistant strains.

Moreover, the global impact of tuberculosis drug resistance extends beyond the affected individuals themselves. The continued transmission of drug-resistant strains hampers TB control efforts, making it even more challenging to reduce the overall burden of the disease. As a result, the potential for drug-resistant TB to fuel a cycle of transmission and perpetuate the epidemic is a major concern for public health authorities.

2.1.6 Diagnostic Methods for Drug-Resistant Tuberculosis

Drug-resistant tuberculosis (TB) is a growing concern worldwide, and early detection of drug resistance is crucial for effective management and treatment of the disease. Several diagnostic methods are used to identify drug-resistant strains of Mycobacterium tuberculosis, the bacterium that causes tuberculosis. Here are some of the commonly employed diagnostic methods for drug-resistant TB (ORGANIZATION; W.H.O, 2008):

1. **Drug susceptibility testing (DST):** This is the gold standard for diagnosing drug-resistant TB. DST involves testing the susceptibility of M. tuberculosis isolates to different anti-TB drugs (D.; S.; E., 2007). The bacteria are cultured from patient samples (such as sputum or tissue specimens), and their growth is assessed in the presence of various drugs. The results determine which drugs the bacteria are sensitive or resistant to.

2. **Phenotypic methods:** Phenotypic methods assess the growth and behavior of M. tuberculosis in the presence of specific drugs. The most widely used phenotypic method is the proportion method, which involves culturing the bacteria on drug-containing media and determining the proportion of colonies that grow in the presence of a specific drug. Other methods include the absolute concentration method, the BACTEC™ MGIT™ (Mycobacteria Growth Indicator Tube) system, and the MODS (Microscopic Observation Drug Susceptibility) assay.

3. **Genotypic methods:** Genotypic methods: Genotypic methods detect specific genetic mutations in M. tuberculosis that are associated with drug resistance. These

methods are often faster than phenotypic tests and can provide information on multiple drug resistances simultaneously. Some commonly used genotypic methods include:

- **Line probe assays (e.g., Xpert MTB/RIF):** These tests detect specific mutations in the M. tuberculosis genome associated with resistance to rifampicin, a key anti-TB drug (CC et al., 2011). Some assays can also detect resistance to other drugs.
- **Next-generation sequencing (NGS):** NGS techniques can sequence the entire M. tuberculosis genome and identify mutations associated with drug resistance. Whole-genome sequencing (WGS) is increasingly used for drug resistance profiling.

4. **Molecular tests:** Molecular tests are rapid and sensitive methods that detect drug resistance-associated mutations in M. tuberculosis. They are based on various molecular techniques, such as polymerase chain reaction (PCR) or DNA hybridization (D.; S.; E., 2007). Examples include the GenoType MTBDRplus and GenoType MTBDRsl assays, which detect resistance to both rifampicin and isoniazid (MTBDRplus) or second-line drugs (MTBDRsl).

5. **Culture-based methods:** Culturing M. tuberculosis isolates in the presence of specific drugs can help identify drug resistance. Automated liquid culture systems like the BACTEC™ MGIT™ system are widely used for this purpose. Additionally, culture-based methods can be combined with DST to determine the susceptibility pattern of the bacteria.

It is important to note that different diagnostic methods have varying levels of sensitivity, specificity, cost, and technical requirements. A combination of different approaches, such as using both phenotypic and genotypic methods, can enhance the accuracy of drug-resistant TB diagnosis and help guide appropriate treatment. The choice of diagnostic method depends on factors such as availability, resources, and local guidelines.

2.1.7 Treatment Options for Drug-Resistant Tuberculosis

As the effectiveness of standard TB drugs diminishes due to the emergence of drug-resistant strains of Mycobacterium tuberculosis, the need for alternative treatment options becomes increasingly critical. In this article, we will explore the various treatment options available for drug-resistant TB and highlight the evolving landscape of TB management.

1. **First-line drugs for drug-resistant TB:** In some cases, drug-resistant strains may still respond to certain first-line anti-TB drugs, such as isoniazid, rifampicin, pyrazinamide, ethambutol, and streptomycin. If a strain shows limited resistance

to one or more of these drugs, they may be included in the treatment regimen, combined with second-line drugs.

2. **Second-line drugs:** Second-line drugs play a crucial role in the treatment of drug-resistant TB. These drugs are generally more toxic, less potent, and more expensive than first-line drugs. They include fluoroquinolones (e.g., moxifloxacin, levofloxacin), injectable agents (e.g., kanamycin, amikacin), and other drugs like ethionamide, cycloserine, and linezolid. The selection and combination of second-line drugs depend on the specific drug susceptibility profile of the strain.

3. **Drug susceptibility testing (DST):** Before initiating treatment for drug-resistant TB, it is essential to determine the drug susceptibility profile of the infecting strain. DST helps identify the drugs to which the strain is resistant or susceptible (D.; S.; E., 2007). This information forms the basis for designing an effective treatment regimen.

4. **Individualized treatment regimens:** Unlike drug-susceptible TB, drug-resistant TB requires more complex and individualized treatment approaches. Treatment regimens are tailored based on the drug susceptibility test results and the patient's medical history. The goal is to construct a regimen that includes drugs to which the strain is susceptible, while avoiding those to which it is resistant.

5. **Bedaquiline and Delamanid:** Two newer drugs, bedaquiline and delamanid, have been approved for the treatment of multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB). These drugs have shown promising results and offer additional treatment options. They are often used in combination with other second-line drugs to form a comprehensive regimen.

6. **Treatment duration and monitoring:** Treating drug-resistant TB requires prolonged therapy compared to drug-susceptible TB. The duration of treatment can range from 9 months to 24 months or longer, depending on the severity and extent of resistance. Regular monitoring of treatment response, drug tolerability, and potential side effects is essential to ensure optimal patient care.

7. **Adherence support and comprehensive care:** Managing drug-resistant TB goes beyond prescribing medications. It necessitates a comprehensive approach that includes patient education, adherence support, nutritional support, psychosocial care, and infection control measures. Close collaboration between healthcare providers, patients, and support systems is vital to achieve successful treatment outcomes.

8. **Future directions:** The field of drug-resistant TB treatment is continually evolving. There is ongoing research to identify new drugs and treatment regimens that

can combat resistance more effectively. Additionally, efforts are being made to improve diagnostic tools, strengthen healthcare infrastructure, and enhance access to quality care in resource-limited settings.

Treating drug-resistant TB requires a multidimensional approach that combines accurate drug susceptibility testing, individualized treatment regimens, and comprehensive patient care. The availability of second-line drugs and newer agents like bedaquiline and delamanid has expanded treatment options and provided hope for improved outcomes. However, challenges such as drug toxicity, treatment duration, and limited access to these medications remain. By investing in research, healthcare infrastructure, and collaborative efforts, we can continue to make progress in the fight against drug-resistant tuberculosis and protect the well-being of affected individuals and communities worldwide.

2.1.8 Challenges in the Control and Management of Drug-Resistant Tuberculosis

Tuberculosis (TB) continues to be a major global health concern, and the emergence of drug-resistant strains, particularly multidrug-resistant tuberculosis (MDR-TB) and extensively drug-resistant tuberculosis (XDR-TB), has further complicated its control and management. This chapter explores the challenges faced in the control and management of drug-resistant tuberculosis, highlighting the various factors that contribute to the persistence and spread of these resistant strains.

1. **Limited Access to Diagnostics:** One of the primary challenges in managing drug-resistant tuberculosis is the limited access to accurate and timely diagnostic tools. Traditional methods for diagnosing TB are often slow and may not identify drug resistance accurately. Advanced diagnostic techniques, such as molecular assays and drug susceptibility testing, are more effective but can be expensive and inaccessible in resource-limited settings. The lack of accessible and affordable diagnostics hampers early detection and appropriate treatment initiation for drug-resistant TB cases.

2. **Complex Treatment Regimens:** Treating drug-resistant tuberculosis requires long and complex treatment regimens, which can last up to two years or more. These regimens involve the use of multiple second-line drugs that are often less effective, more toxic, and costly compared to first-line drugs. The need for prolonged treatment increases the risk of non-adherence, leading to treatment failure and the development of further drug resistance. Additionally, the management of drug interactions and potential side effects further complicates the treatment process.

3. **Inadequate Drug Supply and Quality:** Ensuring a stable and uninterrupted supply of quality-assured second-line anti-TB drugs is crucial for effective manage-

ment of drug-resistant tuberculosis. However, in many countries, there are challenges related to drug procurement, distribution, and quality control. Limited availability of second-line drugs, stockouts, and substandard drug quality undermine the effectiveness of treatment programs, compromising patient outcomes and contributing to the development of additional drug resistance.

4. **Inadequate Health Systems:** The control and management of drug-resistant tuberculosis require a robust and well-functioning healthcare system. However, many countries with a high burden of TB struggle with weak health systems characterized by inadequate infrastructure, inadequate human resources, and fragmented healthcare delivery. The lack of trained healthcare personnel, including doctors, nurses, and laboratory staff, contributes to delays in diagnosis, inappropriate treatment, and inadequate patient monitoring.

5. **Patient Adherence and Support:** Achieving high levels of treatment adherence among patients with drug-resistant tuberculosis is critical for successful outcomes. However, several factors make adherence challenging. The long duration of treatment, the side effects of medications, socioeconomic barriers, and the stigma associated with TB can all affect patient adherence. Additionally, the provision of comprehensive patient support, including psychosocial support, nutritional assistance, and social protection, is often inadequate, further hindering treatment adherence and completion.

6. **Cross-Border Transmission and Migration:** Drug-resistant tuberculosis knows no boundaries and can spread across countries through cross-border transmission and population migration. The movement of people with undiagnosed or inadequately treated TB poses a significant challenge to the control and management of drug-resistant strains. Coordinated efforts between countries are essential to strengthen surveillance systems, facilitate information sharing, and ensure continuity of care for mobile populations.

7. **High Cost of Treatment:** Treating drug-resistant tuberculosis is expensive, placing a considerable financial burden on individuals, families, and health systems. The cost of second-line drugs, specialized diagnostic tests, hospitalization, and supportive care can be prohibitive, particularly in low- and middle-income countries where TB is most prevalent. The high cost of treatment contributes to delays in accessing care, inadequate treatment initiation, and the abandonment of treatment, thereby fueling the development and spread of drug-resistant strains.

The control and management of drug-resistant tuberculosis present numerous challenges that require multi-faceted and comprehensive solutions. Strengthening health systems, improving access to accurate diagnostics, ensuring a reliable supply of quality-assured drugs, enhancing patient support, and promoting international collaboration are vital components of a successful response. Addressing these challenges effectively is crucial to curbing the burden of drug-resistant tuberculosis and achieving global TB elimination goals.

2.1.9 Prevention Strategies and Control Measures for Drug-Resistant Tuberculosis

Preventing the emergence and spread of drug-resistant tuberculosis (TB) is of paramount importance in reducing the global burden of this challenging disease (KJ; S; ML, 2015). This chapter focuses on the prevention strategies and control measures that can effectively mitigate the development and transmission of drug-resistant TB, ensuring better outcomes for individuals and communities.

1. **Strengthening Basic TB Control:** A fundamental step in preventing drug-resistant TB is strengthening basic TB control measures. This includes early and accurate diagnosis, prompt initiation of treatment, and adherence to standardized treatment regimens (W.H.O, 2022). By ensuring that drug-susceptible TB cases are diagnosed and treated appropriately, the risk of developing drug resistance can be significantly reduced. Robust surveillance systems, efficient case management, and infection control practices are essential components of this approach.

2. **Universal Drug Susceptibility Testing:** Universal drug susceptibility testing (DST) is crucial in identifying drug-resistant TB cases at an early stage. By conducting DST for all diagnosed TB cases, healthcare providers can promptly identify drug-resistant strains and initiate appropriate treatment. Accessible and affordable DST methods, such as molecular assays and rapid diagnostics, need to be made available in all healthcare settings to ensure accurate and timely detection of drug resistance.

3. **Adoption of Shorter, Simpler Treatment Regimens:** The use of shorter and simpler treatment regimens can improve treatment outcomes and reduce the risk of developing drug resistance. Streamlined treatment approaches, such as all-oral regimens and fixed-dose combinations, have shown promising results in treating drug-resistant TB. These regimens simplify the treatment process, enhance patient adherence, and reduce the burden on healthcare systems. Wider adoption of such regimens can help prevent the development of additional drug resistance.

4. **Infection Control Measures:** Implementing effective infection control measures is crucial for preventing the transmission of drug-resistant TB within healthcare

facilities and the community. This includes improving ventilation systems, implementing respiratory hygiene practices, using personal protective equipment, and ensuring proper cough etiquette. Infection control measures should be tailored to the specific needs of drug-resistant TB cases, considering the higher infectiousness and prolonged treatment duration associated with these cases.

5. **Contact Tracing and Screening:** Rigorous contact tracing and screening of individuals who have been in close contact with drug-resistant TB cases are vital for early detection and prevention of further transmission. Identifying and testing individuals with a higher risk of infection, such as household contacts and healthcare workers, allows for prompt treatment initiation and reduces the chances of developing drug resistance. Contact investigations and targeted screening should be carried out systematically, with efficient communication and coordination among healthcare providers.

6. **Improved Surveillance and Data Management:** Robust surveillance systems and effective data management are crucial for monitoring the prevalence, trends, and patterns of drug-resistant TB. Timely and accurate reporting of drug-resistant TB cases enables public health authorities to respond promptly, allocate resources efficiently, and identify areas where prevention and control measures need strengthening. Standardized data collection, analysis, and reporting systems should be established and integrated into national TB control programs.

7. **Strengthening Healthcare Systems:** A strong and resilient healthcare system is a cornerstone of effective drug-resistant TB prevention and control. This includes ensuring an adequate supply of quality-assured drugs, strengthening laboratory capacity for accurate diagnostics, training healthcare providers in TB management, and improving access to comprehensive care and support services. Investment in healthcare infrastructure, human resources, and health system strengthening initiatives is vital for sustainable prevention and control efforts.

Preventing and controlling drug-resistant tuberculosis requires a multi-pronged approach that encompasses strengthening basic TB control, universal drug susceptibility testing, implementing infection control measures, contact tracing and screening, robust surveillance, and strengthening healthcare systems. By implementing these prevention strategies and control measures, we can effectively reduce the burden of drug-resistant TB and work

2.2 ASSOCIATION RULES AND APRIORI

Association Rules is a Data Mining technique that has been used a lot in recent times, according to (WITTEN, 2002). It's characterizes as the presence of a set of items in the register from the database, that implies the presence of some other distinct set of items in the same registers. Therefore, the purpose of Association Rules is to find relations that can be used to understand and explore patterns of data behavior. For example, observing the sales data of a supermarket, it is known that 80% of the customers who buy product Q also purchase product W at the same time. In other words, it can be said that this rule presents reliability of 80%.

The format of an Association Rule can be represented as an implication $LHS \implies RHS$, where LHS and RHS are, respectively, the left side *(Left Hand Side)* and the right side *(Right Hand Side)* of the rule, defined by disjoint sets of items. Association Rules can be defined as follows (AGRAWAL; SRIKANT et al., 1994):

Let D be a Database composed of a set of items A = *{a1, ..., am}* ordered lexicographically and a set of transactions T = *{t1, ..., tn}*, in which each transaction $t_i \in T$ is composed of a set of items (itemset) such that $t_i \subseteq A$.

The Association Rule is an implication of the form $LHS \implies RHS$, where $LHS \subset A$, $RHS \subset A$, and $LHS \cap RHS = \emptyset$. The rule $LHS \implies RHS$ occurs in the set of transactions T with confidence *conf* if in *conf%* of the transactions of T in which LHS occurs also RHS occurs. The rule $LHS \implies RHS$ has *sup* support if in *sup%* of transactions in D $LHS \cup RHS$ occurs.

The support value measures the power of the association rule between LHS and RHS and does not relate possible dependencies of RHS with LHS. On the other hand, confidence measures the strength of the logical implication described by the (ZHANG; ZHANG, 2002) rule. Seeking to facilitate the understanding of the measures, they are defined below:

2.2.1 Support

Support represents the generality of a rule, it quantifies the incidence of an *itemset* X or a rule in the data set, that is, it indicates the frequency with which X or with which $LHS \cup RHS$ occurs in the data set. As defined, support for an itemset X can be represented by:

$$sup(X \implies Y) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

where *n(X)* is the number of transactions in which X occurs and N is the total number of transactions considered. The support of a rule $LHS \implies RHS$ can be represented by:

$$sup(LHS \implies RHS) = sup(LHS \cup RHS) = \frac{n(LHS \cup RHS)}{N}$$

where $n(LHS \cup RHS$ is the number of transactions in which LHS and RHS occur together and N is the total number of transactions considered.

2.2.2 confidence

Confidence represents the reliability of a rule. It shows how often the item was considered true, that is, the reliability of the association made by the confidence of the rule estimates the conditional probability of the rule.

$$Conf(X \implies Y) = \frac{sup(X \cup Y)}{sup(X)}$$

Confidence is the frequency at which elements of Y appear in the dataset with transactions that have X. Confidence indicates how often LHS and RHS occur together relative to the total number of transactions in which LHS occurs. As defined, the confidence of a rule $LHS \implies RHS$ can be represented by:

$$conf(LHS \implies RHS) = \frac{sup(LHS \cup RHS)}{sup(LHS)} = \frac{n(LHS \cup RHS)}{n(LHS)}$$

where *n(LHS)* is the number of transactions in which *LHS* occurs.

Usually, the minimum support and confidence values are defined by the user before mining the Association Rules. In general, setting high values for these parameters generates only trivial rules; setting low values generates, in general, a large volume of information in the form of rules, making it difficult for the user to analyze in the post-processing stage. One way to overcome the difficulties in analyzing these rules in Post-Processing is to use algorithms that make it possible to use taxonomies already during the Pattern exploring step.

The problem of obtaining association rules is decomposed into two sub-problems (the steps) (AGRAWAL; SRIKANT et al., 1994):

1. Find all k-itemsets (set of k items) that have support greater than or equal to the minimum support specified by the user (sup-min). The *itemsets* with support equal to or greater than sup-min are defined as frequent *itemsets*, the other sets are called infrequent itemsets;

2. Use all frequent *k-itemsets*, with $k \geq 2$, to generate the association rules. For each frequent *itemset* $l \subseteq A$, find all subsets $\tilde{a}$ of non-empty items of l. For each subset $\subseteq l$, generates a rule in the form of $\implies (l -)$ if the ratio of *sup(l)* by *sup(ã)* is

greater than or equal to the minimum confidence specified by the user (conf-min). With a set of frequent *itemsets* a, b, c, d and a subset of frequent *itemsets* {a, b}, for example, you can generate a rule of type $ab \implies cd$, since $conf(ab \implies cd) \geq conf - min$, where $\frac{conf(ab \implies cd)=\{sup(a,b,c,d)}{sup(a,b)\}}$.

The example below shows how to extract association rules using the 2 steps described.

Let D be a database that contains a set of items A = shorts, pants, shirt, sandals, sneakers and a set of transactions T = 1, 2, 3, 4, in which the list of items purchased by each transaction t_i is presented in the Table 1.

Transaction	**Purchased Items**
1	trouser, tshirt, sneaker
2	tshirt, sneaker
3	short, sneaker
4	trouser, sandal

Table 1 – The list of items purchased per transaction

Considering the value of sup-min = 50% (2 transactions) and conf-min = 50%, it is possible to obtain all association rules contained in Table 1 using the two steps described above.

1. To find all *k-itemsets*, from Table 1, which have support greater than or equal to sup-min (frequent itemsets). Table 2 shows all frequent *k-itemsets*.

Frequent Itemsets	**Support**
{Sneaker}	75%
{Trouser}	50%
{Tshirt}	50%
{Tshirt, sneaker}	50%

Table 2 – Frequent Itemsets

2. With the frequent *k-itemsets* obtained in step 1, where $k \geq 2$, generate all the association rules contained in Table 1, which in this example are:

   Rule 1: $sneaker \implies tshirt$

   - support = support({sneakers, tshirt}) = 50%, which equals **sup-min**.
   - confidence = $\frac{suporte(\{tenis,camiseta\})}{suporte(\{tenis\})} = \frac{50}{75} = 66.66\%$, which is greater than **conf-min**.

   Rule 2: $tshirt \implies sneakers$

- support = support({tshirt, sneakers}) = 50%, which equals **sup-min**.
- confidence = $\frac{support(\{tshirt,sneakers\})}{support(\{tshirt\})} = \frac{50}{50}$ = 100%, which is greater than **conf-min**

Obtaining frequent *itemsets* to generate Association Rules can be performed using several algorithms. One of the most important algorithms for generating frequent *itemsets*, the Apriori algorithm, will be discussed in the next section.

2.2.3 The Apriori Algorithm

The Apriori algorithm is one of the most popular algorithms when it comes to mining association rules in large centralized databases. It finds all sets of frequent items, called frequent itemsets (L_k). Developed by (AGRAWAL; SRIKANT et al., 1994)), it is used to find all frequent *k-itemsets* contained in a database. This algorithm generates a set of candidate *k-itemsets* and then runs through the database to determine if they are frequent, thereby identifying all frequent *k-itemsets*.

The main algorithm (Apriori) makes use of two functions: the Apriori_gen function, to generate candidates and eliminate those that are not frequent, and the Genrules function, used to extract the association rules. The Apriori algorithm is presented in Algorithm 1. It uses the notation L_k to represent the set of frequent *k-itemsets* and C_k to represent the set of *k- itemsets* candidates.

Código 1 – Apriori ((AGRAWAL; SRIKANT et al., 1994))

```
L[1] := {frequent 1-itemsets};
    for (k = 2; L[k-1] !=; k++) do
        C[k] := apriori-gen(L[k-1]); // Generates new set of items
        for all (transactions t    T) do
            C[t] := subset(C[k], t); //Set of item in t
            for all (candidates c    C[t]) do
                c.count++
            for-end
        for-end
        L[k] := {c \in C[k] c.count      sup - min}
    for-end
Return := U[k]L[k]
```

Initially the algorithm counts the occurrence of items, determining the frequent *1-itemsets* that are stored in L_1. The next step, called step k, is divided into two steps. In the first (line 3 of Algorithm 1) the set of frequent *1-itemsets* L_{k-1} obtained in step (k-1) is used to generate the set of *k-itemsets* candidates C_k using the **apriori-gen** function, described in the Algorithm. Next (lines 4 to 9 of Algorithm 1), the database is traversed to determine the support value of *k-itemsets* candidates in C_k. Finally, the frequent *k-itemsets* of each step are identified (line 10). The final solution is given by the union of sets L_k of *k-itemsets* frequent (AGRAWAL; SRIKANT et al., 1994). This solution is used as input to some algorithm that generates association rules.

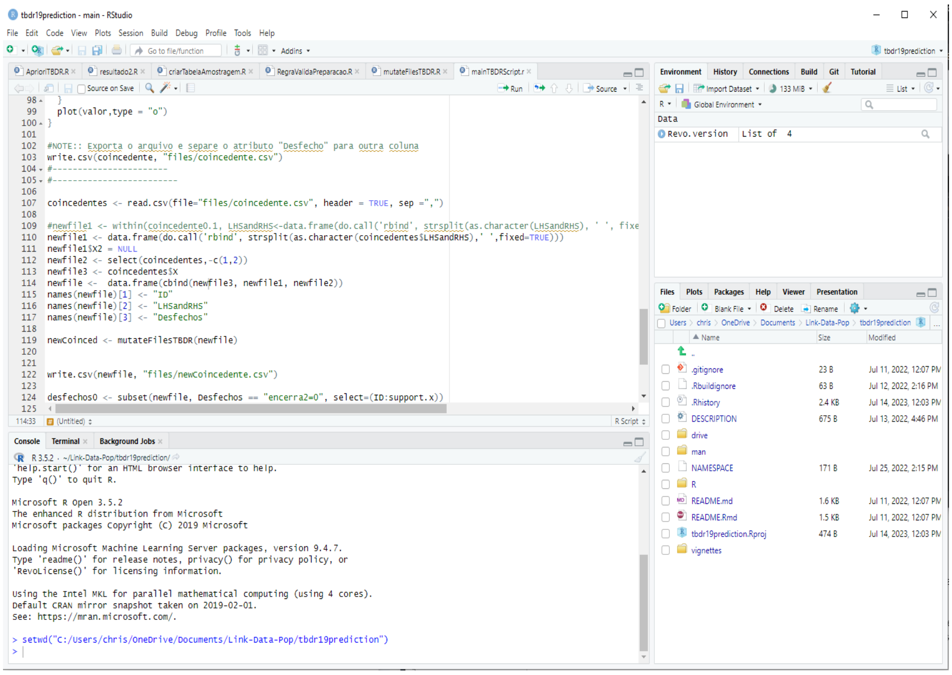
2.3 UNDERSTANDING R AND R-STUDIO BASIC INSTRUCTIONS

In this chapter, we discussed the principle technologies used, why we used them and how we used them. We used R programming language as our original programming language, Rstudio was used as our default workspace environment because it provides features like console, R-libraries, graph visualization and an IDE. When building in R, it´s important to practice clean code and modulizing your functions into packages and libraries, in section 4.3, we talked about how to build a R-package to modulize your functions and deploy them into the CRAN repository (CRAN, 2023) and how to test and you use your package.

2.3.1 Why R ?

In this project, we will be using the R Programming Language software environment for all our analysis. R is a free and open source software developed by statisticians as an interactive environment for data analysis. The full history can be found in the paper "A Brief History of S" (BECKER, 2004). Interactivity is an indispensable feature in data science because the ability to quickly explore data is a necessity for success in this field. R is not a programming language like C and Java. However, like in other programming languages, you can save your work as scripts that can be easily executed at any moment. These scripts serve as a record of the analysis you performed, a key feature that facilitates reproducible work. If you are an expert programmer, you shouldn't expect R to follow the conventions you are used to since you will be disappointed. If you are patient, you will come to appreciate the unequal power of R when it comes to data analysis and data visualization.

The R program runs on all major platforms: UNIX/Linux, Windows, Mac Os. Scripts and data objects can be shared seamlessly across platforms, according to the official website (R-PROGRAMMING-LANGUAGE, 2023). There is a large, growing and active community of R users and, as a result, there are numerous resources for learning and asking questions, This makes it easy for others to contribute add-ons which enables developers to share software implementations of new data science methodologies. This gives R users early access to the latest methods and tools which are developed for a wide variety of disciplines, including ecology, molecular biology, social sciences and geography, just to name a few examples.



```
  }
  plot(valor,type = "o")
}

#NOTE:: Exporta o arquivo e separe o atributo "Desfecho" para outra coluna
write.csv(coincedente, "files/coincedente.csv")
#-----------------------
#-------------------------

coincedentes <- read.csv(file="files/coincedente.csv", header = TRUE, sep =",")

#newfile1 <- within(coincedente0.1, LHSandRHS<-data.frame(do.call('rbind', strsplit(as.character(LHSandRHS), ' ', fixe
newfile1 <- data.frame(do.call('rbind', strsplit(as.character(coincedentes$LHSandRHS),' ',fixed=TRUE)))
newfile1$X2 = NULL
newfile2 <- select(coincedentes,-c(1,2))
newfile3 <- coincedentes$X
newfile <-  data.frame(cbind(newfile3, newfile1, newfile2))
names(newfile)[1] <- "ID"
names(newfile)[2] <- "LHSandRHS"
names(newfile)[3] <- "Desfechos"

newCoinced <- mutateFilesTBDR(newfile)


write.csv(newfile, "files/newCoincedente.csv")

desfechos0 <- subset(newfile, Desfechos == "encerra2=0", select=(ID:support.x))
```

```
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Microsoft R Open 3.5.2
The enhanced R distribution from Microsoft
Microsoft packages Copyright (C) 2019 Microsoft

Loading Microsoft Machine Learning Server packages, version 9.4.7.
Type 'readme()' for release notes, privacy() for privacy policy, or
'RevoLicense()' for licensing information.

Using the Intel MKL for parallel mathematical computing (using 4 cores).
Default CRAN mirror snapshot taken on 2019-02-01.
See: https://mran.microsoft.com/.

> setwd("C:/Users/chris/OneDrive/Documents/Link-Data-Pop/tbdr19prediction")
> 
```

Data	
Revo.version	List of 4

Name	Size	Modified
..		
.gitignore	23 B	Jul 11, 2022, 12:07 PM
.Rbuildignore	63 B	Jul 12, 2022, 2:16 PM
.Rhistory	2.4 KB	Jul 14, 2023, 12:03 PM
DESCRIPTION	675 B	Jul 13, 2022, 4:46 PM
drive		
man		
NAMESPACE	171 B	Jul 25, 2022, 2:15 PM
R		
README.md	1.6 KB	Jul 11, 2022, 12:07 PM
README.Rmd	1.5 KB	Jul 11, 2022, 12:07 PM
tbdr19prediction.Rproj	474 B	Jul 14, 2023, 12:03 PM
vignettes		

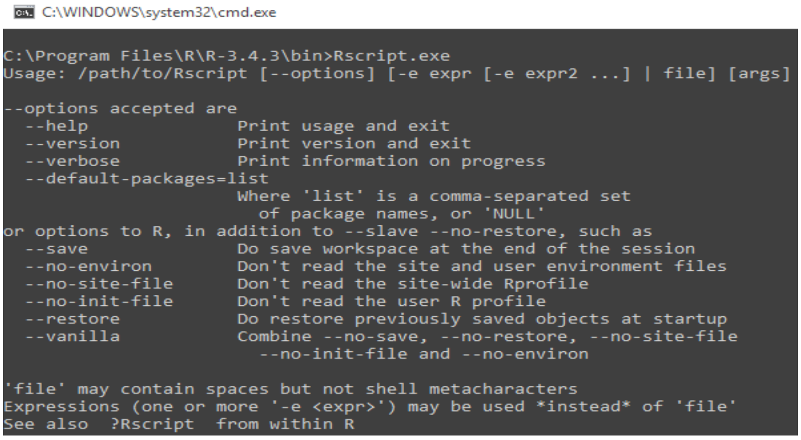
Figure 1 – A Snapshot Of The R Studio

2.3.2 The R Studio

R Studio is our launching pad for this data science projects. It doesn't only provide an editor for us to create and edit our scripts but also provides many useful tools. The R studio looks something like this Figure 1:

2.3.2.1 The Console

Interactive data analysis usually occurs on the R console that executes commands as you type them. There are several ways to gain access to the R console. One way is to simply find the path R.exe or Rscript.exe on your computer. If you try to run R.exe from the command line, you enter into the R terminal as shown in figure 2:



```
C:\Program Files\R\R-3.4.3\bin>Rscript.exe
Usage: /path/to/Rscript [--options] [-e expr [-e expr2 ...] | file] [args]

--options accepted are
  --help              Print usage and exit
  --version           Print version and exit
  --verbose           Print information on progress
  --default-packages=list
                      Where 'list' is a comma-separated set
                        of package names, or 'NULL'
or options to R, in addition to --slave --no-restore, such as
  --save              Do save workspace at the end of the session
  --no-environ        Don't read the site and user environment files
  --no-site-file      Don't read the site-wide Rprofile
  --no-init-file      Don't read the user R profile
  --restore           Do restore previously saved objects at startup
  --vanilla           Combine --no-save, --no-restore, --no-site-file
                        --no-init-file and --no-environ

'file' may contain spaces but not shell metacharacters
Expressions (one or more '-e <expr>') may be used *instead* of 'file'
See also  ?Rscript  from within R
```

Figure 2 – Run R scripts from the command line

2.3.2.2 The Scripts

One of the great advantages of R over point-and-click analysis software is that you can save your work as scripts. You can edit and these scripts using text editor. The material in this project was developed using the interactive integrated development environment (IDE) RStudio (RSTUDIO, 2023). RStudio includes an editor with many R specific features, a console to execute your code, and other useful panes, including one to visualize figures.

2.3.2.3 The Panes

When you start RStudio for the first time, you will see three panes. The left pane shows the R console. On the right, the top pane includes tabs such as ENVIRONMENT and HISTORY, while the bottom pane shows five tabs: File, Plots, Packages, Help, and Viewer (these tabs may change in new versions). You can click on each tab to move across the different features.

2.3.2.4 The Key Bindings

Many tasks we perform with the mouse can be achieved with a combination of keystrokes instead. These keyboard versions for performing tasks are referred to as key binding. For example, we just showed how to use the mouse to start a new script, but you can also use a key binding: Ctrl + Shift + N on Windows and command +Shift + N on the MacOs. Although in this tutorial we often show how to use the mouse, we highly recommend that you memorize key bindings for the operations you use most. RStudio provides a useful cheat sheet with the most widely used commands. You can get it from RStudio directly as shown in Figure 3 (location: rstudio: Help-Cheatsheet- RStudio IDE Cheat sheet)

2.3.2.5 Running commands while editing script

There are many editors specifically made for coding. These are useful because color and indentation are automatically added to make code more readable. RStudio is one of these editors, and it was specifically developed for R. One of the main advantages of RStudio over other editors is that we can test our code easily as we edit our scripts. Another feature you may have noticed is that when you type "library(" the second parenthesis is automatically added. This will help you avoid one of the most common errors in coding: forgetting to close a parenthesis.

2.3.2.6 Changing global options

You can change the look and functionality of RStudio quite a bit, like changing how Rstudio saves your workspace in global options. To change the global options you click on Tools then Global Options. As an example we show how to make a change that we highly recommend. This is to change the Save workspace to .RData on exit to "Never" and uncheck the Restore .RData into workspace at start. By default, when you exit R saves all the objects you have created into a file called .RData. This is done so that when you restart the session in the same folder, it will load these objects. We find that this causes confusion especially when we share the code with colleagues and assume they have this .RData file.

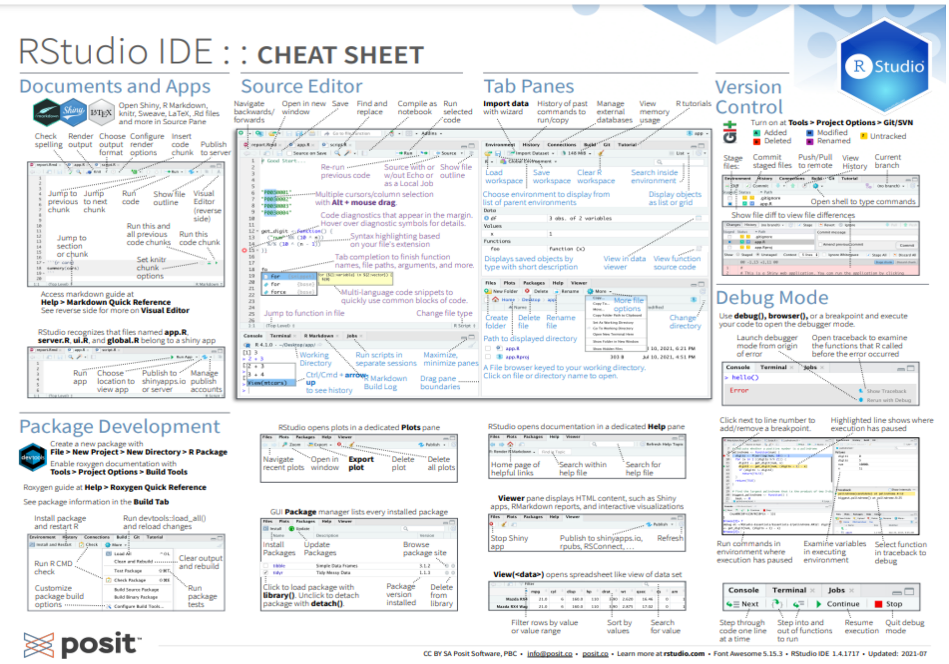
2.3.2.7 Installing R packages

The functionality provided by fresh install of R is only a small fraction of what is possible. In fact, we refer to what you get after your first install as base R. The extra functionality comes from add-ons available from developers. There are currently hundreds of there available from CRAN (Comprehensive R Archive Network) and many others shared via other repositories such as Github. R makes it very easy to install packages from within R. For example, to install the “dslabs” package, which we use to share datasets and code related to this book, you would type:

$$install.packages(\imath dslabs\jmath)$$

$$install.packages(c(\imath tidyverse\jmath, \imath dslabs\jmath))$$

In RStudio, you can navigate to the Tools tab and select install packages. We can then load package into our R sessions using the library function:



RStudio IDE : : CHEAT SHEET

Documents and Apps

Open Shiny, R Markdown, knitr, Sweave, LaTeX, .Rd files and more in Source Pane

Check spelling
Render output
Choose output format
Configure render options
Insert code chunk
Publish to server
Jump to previous chunk
Jump to next chunk
Run code
Show file outline
Visual Editor (reverse side)
Jump to section or chunk
Run this and all previous code chunks
Run this code chunk
Set knitr chunk options

Access markdown guide at **Help > Markdown Quick Reference**
See reverse side for more on **Visual Editor**

RStudio recognizes that files named **app.R**, **server.R**, **ui.R**, and **global.R** belong to a shiny app

Run app
Choose location to view app
Publish to shinyapps.io or server
Manage publish accounts

Source Editor

Navigate backwards/ forwards
Open in new window
Save
Find and replace
Compile as notebook
Run selected code
Re-run previous code
Source with or w/out Echo or as a Local Job
Show file outline
Multiple cursors/column selection with Alt + mouse drag.
Code diagnostics that appear in the margin. Hover over diagnostic symbols for details.
Syntax highlighting based on your file's extension
Tab completion to finish function names, file paths, arguments, and more.
Multi-language code snippets to quickly use common blocks of code.
Jump to function in file
Change file type
Working Directory
Run scripts in separate sessions
Maximize, minimize panes
Ctrl/Cmd + arrow up to see history
R Markdown Build Log
Drag pane boundaries

Tab Panes

Import data with wizard
History of past commands to run/copy
Manage external databases
View memory usage
R tutorials
Load workspace
Save workspace
Clear R workspace
Search inside environment
Choose environment to display from list of parent environments
Display objects as list or grid
Displays saved objects by type with short description
View in data viewer
View function source code
Create folder
Delete file
Rename file
More file options
Change directory
Path to displayed directory

A File browser keyed to your working directory. Click on file or directory name to open.

Version Control

Turn on at **Tools > Project Options > Git/SVN**

Debug Mode

Use **debug()**, **browser()**, or a breakpoint and execute your code to open the debugger mode.

Launch debugger mode from origin of error
Open traceback to examine the functions that R called before the error occurred
Click next to line number to add/remove a breakpoint.
Highlighted line shows where execution has paused
Run commands in environment where execution has paused
Examine variables in executing environment
Select function in traceback to debug
Next
Continue
Stop
Step through code one line at a time
Step into and out of functions to run
Resume execution
Quit debug mode

Package Development

Create a new package with **File > New Project > New Directory > R Package**
Enable roxygen documentation with **Tools > Project Options > Build Tools**

Roxygen guide at **Help > Roxygen Quick Reference**

See package information in the **Build Tab**

Install package and restart R
Run devtools::load_all() and reload changes
Run R CMD check
Clear output and rebuild
Customize package build options
Run package tests

RStudio opens plots in a dedicated **Plots** pane

GUI **Package** manager lists every installed package

Install Packages
Update Packages
Browse package site
Click to load package with library(). Unclick to detach package with detach().
Package version installed
Delete from library

RStudio opens documentation in a dedicated **Help** pane

Viewer pane displays HTML content, such as Shiny apps, RMarkdown reports, and interactive visualizations

View(<data>) opens spreadsheet like view of data set

CC BY SA Posit Software, PBC • info@posit.co • posit.co • Learn more at rstudio.com • Font Awesome 5.15.3 • RStudio IDE 1.4.1717 • Updated: 2021-07

Figure 3 – R-Studio IDE Cheat Sheet for Keyboard Shortcut

```
library(dslabs)
```

Once packages are installed, you can load them into R and you do not need to install them again, unless you install a fresh version of R. Remember packages are installed in R not RStudio. It is helpful to keep a list of all the packages you need for your work in a script because if you need to perform a fresh install of R, you can re-install all your packages by simply running a script.You can see all the packages you have installed using the following function:

```
installed.packages()
```

2.4 BUILDING A R-PACKAGE USING DEVELOPMENT TOOLS

R is a powerful programming language widely used by statisticians and data scientists for its extensive collection of packages. Creating your own R package allows you to encapsulate your code, share it with others, and contribute to the R community. In this project, we provided an in-depth, step-by-step guide on how to build an R package using R Development Tools. By following these instructions, you will gain the necessary skills to create and distribute your own R package effectively.

1. **Package Structure and Documentation:**

2.4.1 Setting Up Your Development Environment:

Before starting package development, ensure you have the necessary tools installed. Begin by downloading and installing R from the official R website (R-PROGRAMMING-LANGUAGE, 2023). Next, install RStudio, a popular Integrated Development Environment (IDE) for R (RSTUDIO, 2023). Once you have R and RStudio installed, open RStudio and proceed to install the "devtools" package by executing the following command in the R console:

```
install.packages(”devtools”)
```

2.4.2 Creating a New Package:

After installing the required tools, create a new R package. In RStudio, click on "File" in the top menu, followed by "New Project." Choose the option "New Directory" and select "R Package." Assign a name to your package and designate a directory to store it. RStudio will generate the necessary files and folders for your package automatically.

2.4.3 Defining the Package Structure:

An R package follows a specific structure to ensure consistency and ease of use. Essential files and folders include "R/" for storing R code, "man/" for documentation, "tests/" for unit tests, and "DESCRIPTION" for package metadata.

Start by adding your R scripts to the "R/" folder. These scripts will contain the functions and code you want to include in your package. Remember to document your functions using Roxygen comments to generate package documentation.

2.4.4 Documenting Your Package:

Documentation is vital for users to understand and utilize your package effectively. Create .Rd files in the "man/" folder to document your package's functions, datasets, and other elements. These files utilize a specific markup language to describe various aspects of your package.

Refer to the official R documentation (CRAN, 2023) for detailed guidance on creating comprehensive documentation.

2. **Building and Checking Your Package:**

2.4.5 Building Your Package

Once you have written the code and documented your package, it's time to build it. Open the R console within RStudio and run the following commands:

```
library(devtools)

build()
```

Executing these commands will generate a package bundle (.tar.gz file) in your project directory.

2.4.6 Checking Your Package:

To ensure your package meets the required standards and guidelines, perform a thorough check using the following command:

check()

The check process identifies potential issues or errors that need to be addressed before sharing or distributing your package.

3. **Adding Functionality to Your Package:**

2.4.7 Adding Functions:

To enhance the functionality of your package, add functions to the "R/" folder. Each function should be defined in a separate .R file and follow the proper naming conventions.

Consider the following tips when creating functions:

- Write clear and concise code.
- Include parameter descriptions and examples in the function documentation.
- Consider adding error handling and robust input validation.

If your package depends on other packages, specify these dependencies in the "DESCRIPTION" file.

2.4.8 Testing Your Package

Testing is a critical aspect of software development, including the creation of R packages. Thorough testing ensures that your package functions as intended, catches potential bugs, and maintains the reliability of your code. In this project, we provided a comprehensive guide on how to test an R package using R Development Tools. By following these steps, you will be able to perform effective testing and ensure the quality of your package.

Thorough testing ensures the reliability and functionality of your package. Create unit tests in the "tests/" folder to validate the behavior of your functions and detect potential errors.
Different types of testing are commonly used for R packages, including:

- **Unit Testing**: Tests individual functions or code units in isolation.
- **Integration Testing**: Verifies the interaction between different components or functions within your package.
- **System Testing**: Ensures that the package works correctly in a complete system or environment.
- **Performance Testing**: Measures the performance and scalability of your package under different conditions.

3 METHODOLOGY

In this chapter, we discussed the database used in the project, the source of our data, study pupolation, variable selection and variable dicitionary, also the methods and techniques used to implement this model, starting with the R-library called "TBDR19Prediction", this library is currently deploy in the CRAN repository and in a github repository (TBDR19PREDICTION.GITHUB, 2023).
The bootstrap and bagging was the technique applied in the model, various samples of the dataset were created, for each sample we implement test and train model.

3.1 DATABASE SOURCE, SELECTION AND DICTIONARY:

In the field of data analysis and prediction, the quality of input data plays a vital role in the accuracy and reliability of the results. The R package "tbdr19prediction" provides a comprehensive set of functions that assist users in data selection and pre-processing, enabling them to enhance the robustness of their predictions. This work explores the various techniques and functionalities offered by the package, highlighting its significance in improving data quality. The database used came from the project "Data Science in Public Health: linkage, data mining and machine learning from a perspective of the Academic Health Department", approved by the Research Ethics Committee of the Instituto de Estudos em Saúde Coletiva da UFRJ, under number CAAE 18964619.0.0000.5286. The present study corresponds to one of the subprojects of the aforementioned project.

3.1.1 Database Source and Study Population:

We used a database gotten from a record linkage between data from the glsSITE-TB (Sistema de Informação sobre Tratamento Especiais da Tuberculose) from 2013 to 2014, and from SINAN (Sistema de Informação de Agravos de Notificação) from 2001 to 2014, and with data from SIM (Sistema de Informação sobre Mortalidade) from 2013 to 2016 (WN, 2021 Oct 9). Tuberculosis notified cases and without duplication from SINAN database were obtained from Rio de Janeiro public hospital and health centers and SIM notified cases were obtained from Rio de Janeiro Health ministry.

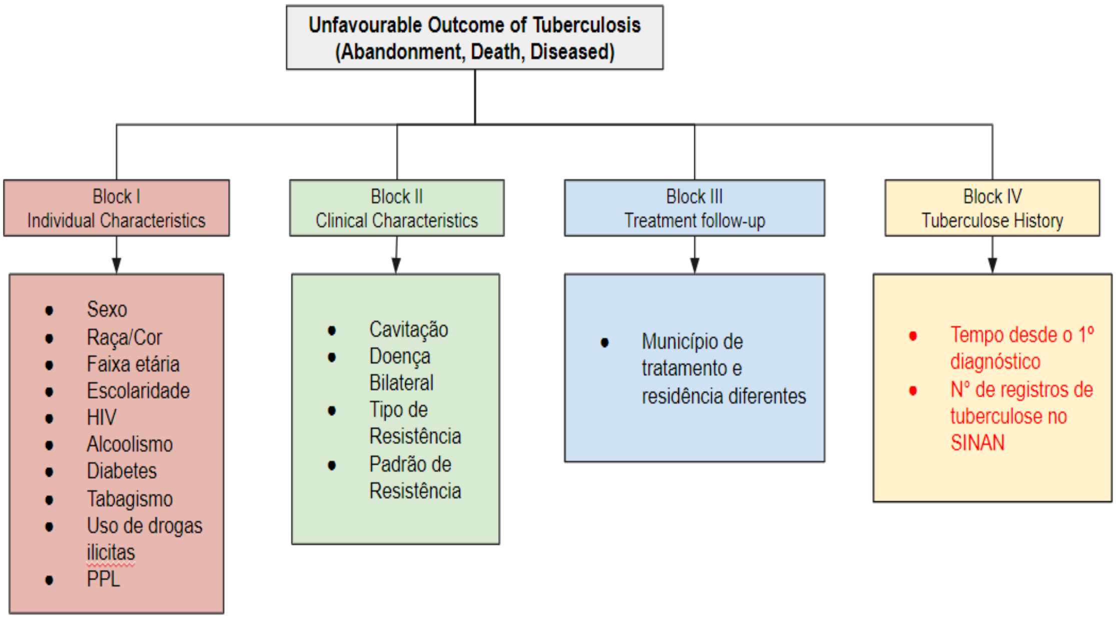
The study population consisted of people diagnosed with notified drug resistant pulmonary TB and residing in the city of Rio de Janeiro between 2013 and 2014. In this study group of individuals, we exclude the individuals notified as terminated in SITE-TB due to a change in diagnosis, also exclude patients under the age of 15 years old.

The next step in any data analysis task is selecting the relevant data for the prediction model. The "tbdr19prediction" package provides several functions to streamline this process. The dataset is received as parameters in the function's definition and may have already been prepared with attributes for which the predictive model could be run. Let's dive into the attributes or variables that were selected for this model and why they were selected.

3.1.2 Dictionary for the Selected Variables

The original database has about 1000 registers and 17 variables selected from Figure 4. This number of variables is too many for the predictive model we are going to be running, and so therefore, for different results and compiled models we will be using a few selected variables in order to get the most optimized results possible. In the table 3, we can observe a pre-selected list of 24 variables and how they are being represented in the dataset.

Figure 4 – Analised Variables: Characterization Model



source: compiled by (BARTHOLOMAY, 2019)

Table 3 – Dictionary for the selected variables

Variables	**Observations**
Sexo (Gender)	sexo (F e M), sexo2 (0=Masculino e 1=Feminino)

Raça/cor(Race/color)	raça (Amarela, Branca, Ignorada, Indígena, Negra, Parda) raca2 (1=Branca, 2=preta e parda, 3=amarela/indígena) raca3 (0=Branca/amarela/indígena/ignorado, 1=preta/parda)
Faixa etária(age group)	fxet3 (1=15-59, 2=60+)
Escolaridade (Education)	escol (12 ou mais, de 1 a 3, de 4 a 7, de 8 a 11, ignorada, nenhum) escol4 (1=0 a 7, 2=8 ou mais)
HIV	hiv_sitetb (Em branco, Em andamento, Não realizado, Negativo, Positivo) hiv3 (0=negativo, 1=positivo, 3=ignorado) comaids (1=positivo)
Alcoolismo (Alcoholic)	alcool (0=não/não sabe, 1=sim) comalcool (1=sim) comalcoo2 (1=sim, 2=não/não sabe)
Diabetes	comdiab (0=não/não sabe, 1=sim) comdiab2 (0=não/não sabe, 1=sim)
Tabagismo (Smoker)	comtabagis (0=não/não sabe, 1=sim) comtabagis2 (0=não/não sabe, 1=sim)
Uso de drogas ilícitas (Use of illicit drugs)	comdrogas (0=não/não sabe, 1=sim) comdrogas2 (0=não/não sabe, 1=sim)
Outras doenças e agravos associados (Other diseases and injuries associated)	Fazem parte dessa variável: silicose, neoplasias, transplantado, usuário de inibidores de TNF alfa e corticoides, convulsão, hepatites virais, insuficiência renal/hemodiálise e transtorno mental
População privada de liberdade (Ex-convicted population)	Esse dado foi extraído da variável local de provável contágio ppl (0=não, 1=sim)
Cavitação (Cavitation)	cavitacao (0=não, 1=sim)
Doença bilateral (bilateral disease)	bilateral (0=não, 1=sim)
Tipo de resistência (resistance type)	Indivíduos com registro do tipo de resistência primária com notificações prévias de resistência (?) no Sinan tiveram o tipo de resistência alterada para adquirida tipores (Adquirida, Nao se aplica, Primaria) tipores3 (0=primaria, 1=adquirida)

Padrão de resistência inicial (Initial resistance pattern)	pdres5 (1=MDR/RR, 2=XDR) pdres (Em branco, Multirresistencia, Resistencia extensiva, Resistente Rifampicina) (deixei como estava na base)
Tipo de esquema inicial (Initial schema type)	esquema (0=individualizado, 1= padronizado)
Tratamento com fluorquinolona (Fluoroquinolone treatment)	
Tratamento com medicamentos injetáveis (Treatment with injectable medications)	
Município de residência e tratamento diferentes (Municipality of residence and different treatment)	mumigual (0=não, 1=sim)
Evolução clínica desfavorável informada (Informed unfavorable clinical evolution)	Pode ter sido informado a qualquer momento do tratamento
Trocou de tipo de esquema (Changed schema type)	É preenchida como sim quando o tratamento atual é diferente do inicial. Não foi possível controlar as alterações entre esquemas individualizados trocesq (0=não, 1=sim)
Teve reação adversa (had an adverse reaction)	reacoes (0=sem registro de reação adversa, 1=somente reações adversas menores, 2=pelo menos uma reação adversa maior)
SINAN Registration	nsinan4 (0=até 3 eventos, 1= 4 eventos ou mais) numregsina (contínua)

Source: Author

3.2 THE MINING TECHNIQUE

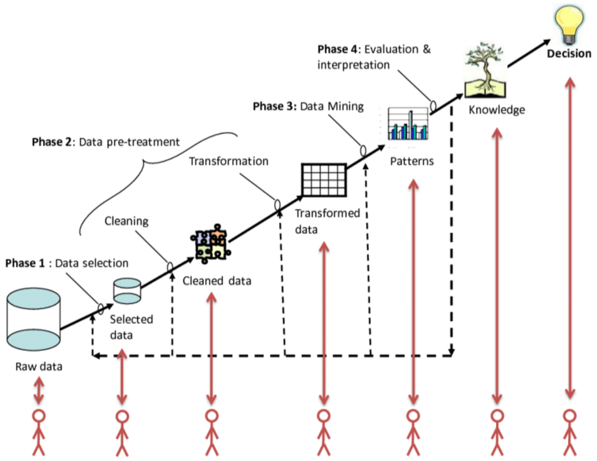
The purpose of this chapter is to provide a comprehensive overview of the mining technique and it's derivative formula to getting the best valid rules that we explore in the next chapter. The following steps are necessary in order to understand an overview of this technique.

- The KDD process for extracting useful Knowledge from volumes of data (FAYYAD et al., 1996).
- The bootstrap and bagging technique using (Test and Train).

3.2.1 Algorithm for Valid Rules using KDD

The model created to obtain the database of valid rules was implemented using the KDD process, which is commonly used to extract valuable insights from large volumes of data. According to (FAYYAD et al., 1996) this process is a systematic and iterative approach that involves several stages in order to transform raw data into useful knowledge.

Figure 5 – The KDD process (Fayyad et al., 1996)



Source: Fayyad et al., 1996

The KDD involves 7 steps, shown in Figure 5, and their sequence is important for obtaining the expected results. KDD as a methodology helps us to accurately extracting information (HAN; KAMBER; PEI, 2012).

1. Understanding the SITE-TB database
2. Selecting important characteristics/variables from the database
3. Cleaning and preparing the data
4. Transformating the data into transactions
5. Apply data mining algorithm
6. Interpretation/Evaluation of the resulting rules
7. Analysing nem knowledge and presentation informations.

3.2.2 The Mining Technique: Bootstrap and Bagging

As in the book (EFRON; TIBSHIRANI, 1994), **Bootstrap Techniques** involves resampling from the available data to create multiple datasets, computing estimates on each resampled dataset, and then analyzing the distribution of these estimates to understand the variability and construct confidence intervals, while **Baggings techniques**, as explained in the article by (BREIMAN, 1996), involves in creating multiple subsets of the original dataset through random sampling with replacement and training separate models on each subset.

Here is a step-by-step overview of the Bootstrap and Bagging method:

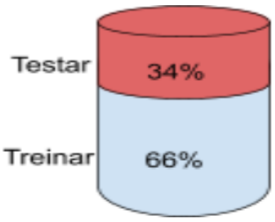
- Data collection
- Random sampling with replacement
- Model training
- Prediction aggregation
- Inference and uncertainty estimation

3.2.3 The Mining Techniques: Test/Train and Apriori

Putting into practical the understanding of bootstrap and bagging techniques, as from (EFRON; TIBSHIRANI, 1994) and (BREIMAN, 1996):

- In the first, we divided our original dataset with 1000 registers and 15 variables into Test and Train model, where 34% was used for Test and 66% for Train, as in Figure 6.

Figure 6 – Divide the DB into 34% Test and 66% Train



Source: Author

- In the second step, after having our database divided into two as shown in figure 7. Utilizing the test database with 34% of our original data, we apply the **Apriori** function with parameters like support and confidence configured into the apriori function as in the Algorithm 2. Running Apriori function into our Test data generates a new dataset of set of rules as the Figure 7a ilustrates.

- The same previous procedure is reapeated to our 64% of Training database, by running the **Apriori** Algorithm 1 to the 64% Train database to generate a new dataset of sets of rules, as ilustrated in Figure 7b.

Figure 7 – Random division into Test/Train

Apriori
Test
34%
Rules
Apriori
Train
66%
Rules

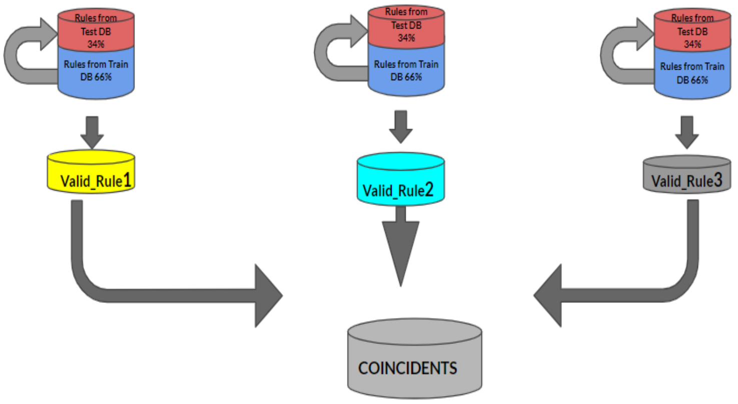
(a) Apriori applied on Test dataset to generate rules.

(b) Apriori applied on Train dataset to generate rules.

source:Author

- In the next step, we created multiple subsets of the original database. For each subset, step 2 and 3 is applied where it was divided into Test/Train and running Apriori Algorithm. As shown in Figure 8, 3 subsets were created randomly as modeled in Bagging Technique (BREIMAN, 1996), For each subset of dataset, the rules generated by 34% of test dataset was used to train the other 66% dataset, the result of the training generated a new dataset called *valid_Rules* becaused it is composed of valid Rules. Following the image example in figure 8, 3 subsets generated 3 datasets of valid rules, from which the *Coincident* dataset was generated by filtering only unique rules from the 3 dataset of valid rules, in other words, the *Coincident* dataset only contain unique rules with no duplication.

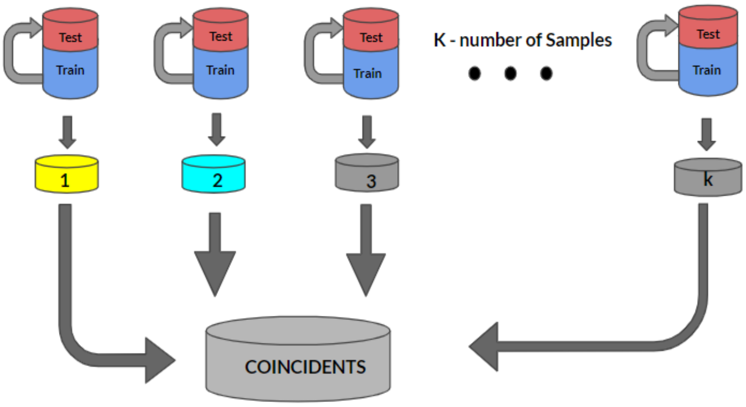
Figure 8 – Mining for Coincidents from 3 sets of Valid Rules.



Source: Author

- As explained in the previous step, In order to obtain the coincident dataset, we need a number of valid rules dataset and since there is no specific number of valid rules dataset needed to obtain coincident dataset, we created a variable K. As shown in Figure 9, K is the number of samples of valid rule datasets need to generate the best Coincident dataset. In the next step we would discuss how to find the ideal value for K.

Figure 9 – Optimizing with K-number of samples.

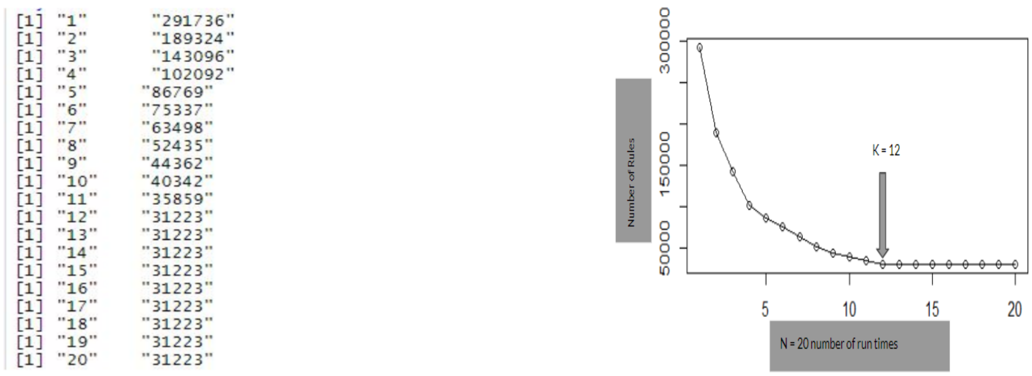


Source: Author

- To find the correct value of K, We run the algorithm several times until it converges to a value, this phenomenal can be explained futher by the Figure 10. The first subset produced 291,736 valid rules, the second subset produced 189,324 valid rules, as shown in figure 10a and figure 10b. The number of valid rules in coincident keeps converging to a value and by the twelveth run we now have a staple set of valid rules for our coincident which is 31,223 valid rules and the value for K is 12, in other words, to generate a refined dataset for coincident we needed to mined 12 sets of valid rules and obtain only unique rules from them.

- The final step in this mining technique, after obtaining the coincident dataset that is made up of all possible valid rules, the coincident dataset is made up of three groups, they are *the Antecident*, *the Consequent* and *the metrics*. The antecident are the rules and every rules contains more that one item, the consequent are the resulting outcomes like Cured, Abandonment, Failed and Death, and the metrics are Support, Confidence, Lift, Count, these metrics are important in the analises of the dataset. As shown in figure 11, the coincident dataset was filtered into 4 different dataset based on their consequent values (Cured, Abandonment, Failed, Death). Various results, knowledge and conclusions were derived from the coincident dataset and would be discussed in the next chapter.

Figure 10 – Random division into Test/Train



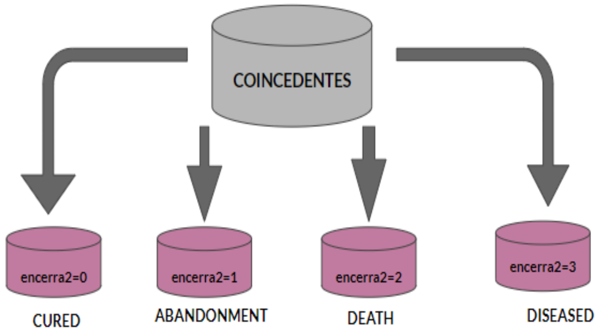
```
[1] "1"      "291736"
[1] "2"      "189324"
[1] "3"      "143096"
[1] "4"      "102092"
[1] "5"     "86769"
[1] "6"     "75337"
[1] "7"     "63498"
[1] "8"     "52435"
[1] "9"     "44362"
[1] "10"    "40342"
[1] "11"    "35859"
[1] "12"    "31223"
[1] "13"    "31223"
[1] "14"    "31223"
[1] "15"    "31223"
[1] "16"    "31223"
[1] "17"    "31223"
[1] "18"    "31223"
[1] "19"    "31223"
[1] "20"    "31223"
```

(a) Discovering the ideal value of K.

(b) Graph plot to find the ideal value of K.

Source: Author

Figure 11 – Extracting Outcomes from Coincidents "Cured - Abandonment - Failed - Death".



Source: Author

3.3 THE "TBDR19PREDICTION" R-LIBRARY

In this project, we created an R-library “tbdr19prediction” for the project Tuberculosis drug-resistant, a predictive model created in 2019, and so comes the name tbdr19prediction. The steps to using this library are detailed in each point like installation, features, data selection and pre-processing.

3.3.1 Installation

The “tbdr19prediction” package provides a framework for working with a predictive model R code. It offers a simple and consistent syntax for running association rule models especially for this project. The models in this library can only be used for predictive models on tuberculosis drug-resistant type of datasets, as the results may vary and be similar in some instances.

The package can be installed from the released version of “tbdr19prediction” from

CRAN (CRAN, 2023) or from the development version from Github (GITHUB, 2023).

To install from CRAN, open your R console and run the following command:

$install.packages(”tbdr19prediction”)$

To install from Github, run the following commands

$install.packages(”devtools”)$

$devtools :: install_github(”chrismomentus/tbdr19prediction”)$

3.3.1.1 Features and functions

The library has a structure for each unit functions which typically follows a three step process:

- A TBDR dataset that the function will be applied to.
- The scope of the function.
- The return dataset type.

A set of functions were written to solve the predictive model algorithm. Let's look into some of the functions and their description to understand their purpose in the project. Commonly used functions includes:

- *‘AprioriTBDR()‘*: This function applies an association rule called apriori algorithm into the defined dataset TBDR. Returns a set of rules in the form of a dataframe.
- *‘prepararRegrasValidas()‘*: Prepare the sets of rules and separate them by metrics and outcomes. This function receives a set of rules in a form of dataset that was generated from function *‘AprioriTBDR()‘* that applied apriori algorithm to our original database. When rules are created they are embedded with a curly bracket, i.e *‘escol4=0,alcool=0,comdiab=0,comdrogas=0,ppl=0‘*. This function eliminates the curly brackets from the rules and also, re-arranges the columns in order of *(‘rules’, ‘count’, ‘lift’, ’confidence’, ‘support’, ‘outcomes’)*. This function returns a well prepared and cleaned dataset of rules, metrics and outcomes.
- *‘criarTabelaAmostragem‘*:This function creates samples of a dataset, and so therefore, it receives a dataset, the desired sample size from (0 to 1) and a value used to set seed for partition reproduction. Returns a sample from the received dataset.

- *‘mutateFilesTBDR‘*: This function is called the coincident mutation because it separates the item of the rules and generates a matrix of data with them. It receives a dataset of a set of coincidences and returns a dataset that was generated in matrix form where each item becomes a column or attributes.

4 RESULTS

The frist result of this work was presented to the (JICTAC, 2020) (Jornada De Iniciação científica Tecnológica, Artística e Cultural), this work was titled "Emerging Knowledge About the Unfavourable Results of Drug-Resistant Tuberculosis Treatment", this presentation received an honourable mention. In the following paragraphs we will be exploring this result.

4.1 CONFIGURATION OF METRICS

In the mining of these results, the following configurations were necessary to obtain the correct Informations. 1000 registers and 13 variables were selected, the variables are Sex, Race/Color, Age, Education, HIV, Use of illicit drugs, PDL, Cavitation, Bilateral Disease, Pattern of Resistance, City of treatment and residency (BARTHOLOMAY, 2019) Of the total records, 64% of the records are related to cure, 16.5% to abandonment, 7.7% to Failed and 12% to death. 66% of the database records were randomly selected for "training" and 34% for "test".

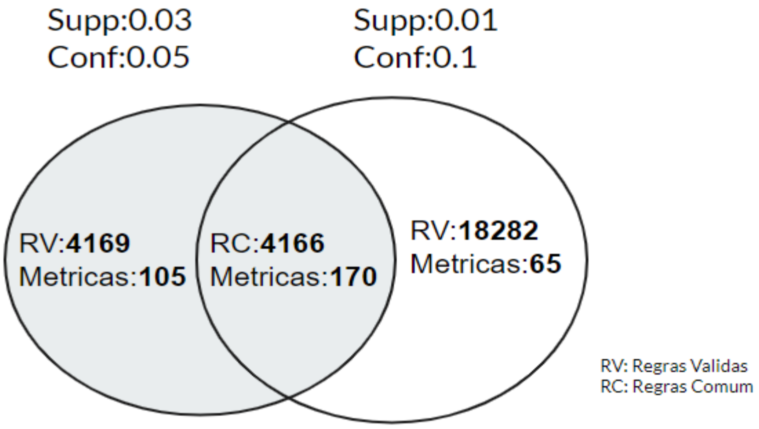
To select the appropriate metrics for our sets of rules, We used 2 methods to help decide these values, it's important to choose a correct value for support and confidence, this will be parameters needed for the apriori function calls as we can see below.

Código 2 – Apriori Function Call with Parameters.

```
rules <- apriori(Dataset, parameter = list(supp = 0.03, conf = 0.05));
```

- The first method for determining the value of support and confidence as parameters in code 2, imagine being in an indecisive situation between various values that you could use, one method would be looking at the intersections of the values, to make happen, compile the apriori function for each of the metrics you wanted to generate your rules. Applying the intersection logic function to these different dataset will generate a new dataset that contains common rules. Lets take an example. the figure 12, we have an intersection between two values of support (0.03 & 0.01) and confidence(0.05 & 0.1), to which apriori was applied to both metrics and a filter of common rules were generated.
- The other method is ploting a truth table of all possible value of support and confidence as parameters for the apriori funtion, for example. In table 4 We have 3 possible value for confidence (0.05, 0.10, 0.15) and 4 possible values for support(0.01,

Figure 12 – Intersection of Common Rules



Source: Author

0.03, 0.05, 0.1). With the truth table we can see important factors like the number of valid rules where created, the metrics and counts of resulting variables like (Cure, Abandon, Failed and Death)

Table 4 – Compilation of Various configurations of Support and Confidence

	Support: 0.01	Support: 0.03	Support: 0.05	Support: 0.1
Conf: 0.05	ValidRule:18377 Cure:13144 Abandon:4460 Failed:116 Death:657	ValidRule:4169 Cure:3608 Abandon:483 Failed:3 Death:75	ValidRule:1103 Cure:1054 Abandon:41 Failed:0 Death:8	ValidRule:69 Cure:69 Abandon:0 Failed:0 Death:0
Conf: 0.10	ValidRule:18282 Cure:13144 Abandon:4460 Failed:23 Death:655	ValidRule:898 Cure:823 Abandon:71 Failed:0 Death:4	ValidRule:1099 Cure:1054 Abandon:41 Failed:0 Death:4	ValidRule:69 Cure:69 Abandon:0 Failed:0 Death:0
Conf: 0.15	ValidRule:17826 Cure:13145 Abandon:4460 Failed:0 Death:222	ValidRule:17826 Cure:13145 Abandon:4460 Failed:0 Death:222	ValidRule:1094 Cure:1054 Abandon:40 Failed:0 Death:0	ValidRule:69 Cure:69 Abandon:0 Failed:0 Death:0

Source: Author

4.1.1 Preliminary Results

The Apriori algorithm was configured with Confidence = 0.05 and Support = 0.05, and executed 50 times on the same training set. It was observed that from the 13th execution the results were maintained. At each execution, the rules were saved and only the matching ones were kept. The most relevant rules were selected for profiles of each unfavorable outcome, based on the highest Lift and confidence, choosing the rules with the fewest items, as they were more comprehensive.

In the 1st run, 300,000 rules were generated and in the 13th, only 29,007 matching rules. Of this total, each outcome obtained Cure (64%) = 16194, Abandonment (16.5%) = 7103, Failed (7.5%) = 3605 and Death (12%) = 2105.

Table 5 – Main Compilation: SUPP[0.05] CONF[0.05]

Metrics	Valid Rules	Count	Lift	Confidence	Support
Cure: 64%	16194	Min:2 Max:69 Méd:9.19	Min:1.10 Max:1.58 Med:1.30	Min:0.72 Max:1 Méd:0.99	Min:0.01 Max:0.23 Méd:0.3
Abandon: 16,5%	7103	Min:2 Max:27 Méd:5.75	Min:1.10 Max:6.49 Méd:2.13	Min:0.17 Max:1 Méd:0.02	Min:0.01 Max:0.10 Méd:0.02
Failed: 7,5%	3605	Min:2 Max:18 Méd:4.77	Min:1.12 Max:5.03 Méd:3.24	Min:0.09 Max:0.33 Méd:0.25	Min:0.01 Max:0.06 Méd:0.02
Death: 12%	2105	Min:2 Max:19 Méd:4.67	Min:1.12 Max:4.49 Méd:1.61	Min:0.12 Max:0.78 Méd:0.17	Min:0.01 Max:0.07 Méd:0.02

Source: Author

4.2 CONFIGURATION: SUPP[0,05], CONF[0,05]

We obtained a new knowlegde when we explored the favourable and un-favourable outcomes of patients, we know cure is a favourable result and while abandonment, Failed and Death are un-favourable. In these results we explored unique profiles of patients and studied their patterns, firstly we would explore favourable profiles from cure and then later un-favourable.

4.2.1 Favourable Results: Cure

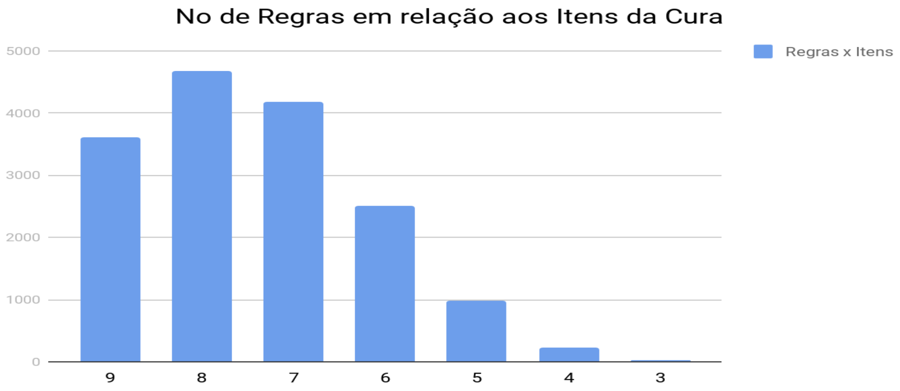
From the main compilation, profiles with Cure as an outcome produced 16194 valid rules, that is 64% of the total number of valid rules, in Table 6 and Figure 13 , there are 3608 rules with 9 items in each rules, 4669 rules with 8 items and the lowest number of items is 3 with 25 valid rules.

Some Items in these rules has unique purpose and makes the difference, In Figure 14, the most outstanding Item was *Escol=12*, this shows that patients with high educational level are liable to Cure of the virus, this variable had a weight of 56.70, the second most outstanding variable is *Drogas-Nao*, this patients that don't use drugs are also liable to

Table 6 – Nº Of Items x Nº of rules

Nº Of Items	**Nº of rules**
9	3608
8	4669
7	4173
6	2517
5	974
4	226
3	25

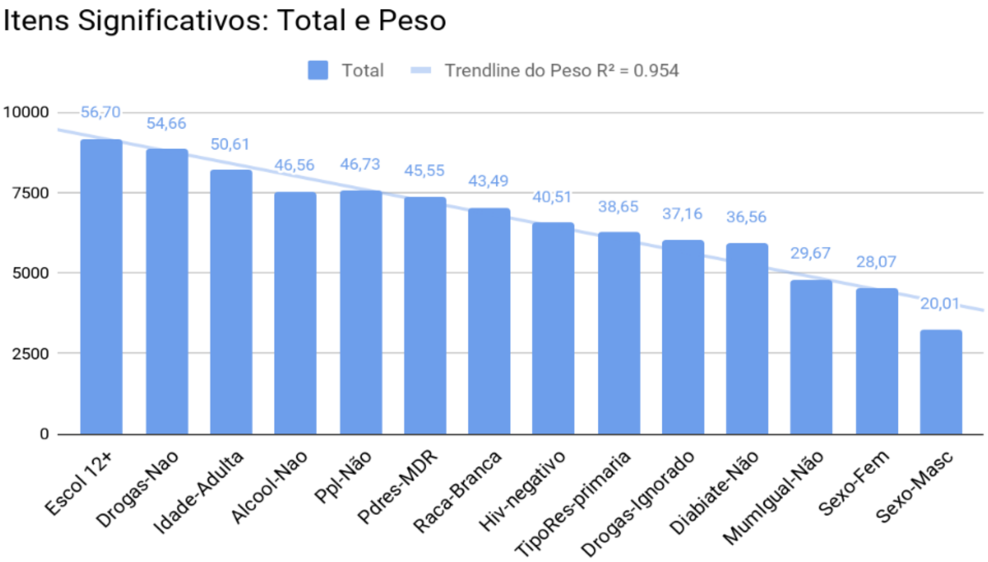
Figure 13 – Plot: Nº of Rules in relations to Items leading to Cure



Source: Author

getting Cure, for time sake, we won't be looking at all these items. the bar chart in Figure 14 list 14 outstanding Items.

Figure 14 – Significant Items: Total x weight (Cure)



Source: Author

4.2.2 Unfavourable Results: Abandonment

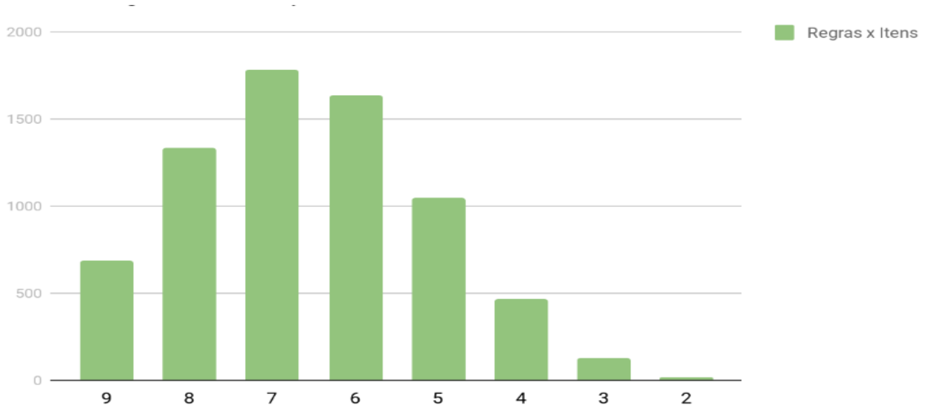
In Table 7, Shows an overview of how Rules and Items are distributed for profiles with Abandonment, With the used configuration, we gather 685 rules with 9 items, 1333 rules with 8 items, 1779 rules with 7 items and the lowest number of items being 2 with 18 valid rules. The bar Figure in 15, show the spread of datas from table 7.

Table 7 – Nº Of Items x Nº of rules (Abandon)

Nº Of Items	Nº of rules
9	685
8	1333
7	1779
6	1637
5	1048
4	470
3	131
2	18

Source: Author

Figure 15 – Plot: Nº of Rules in relations to Items leading to Abandonment

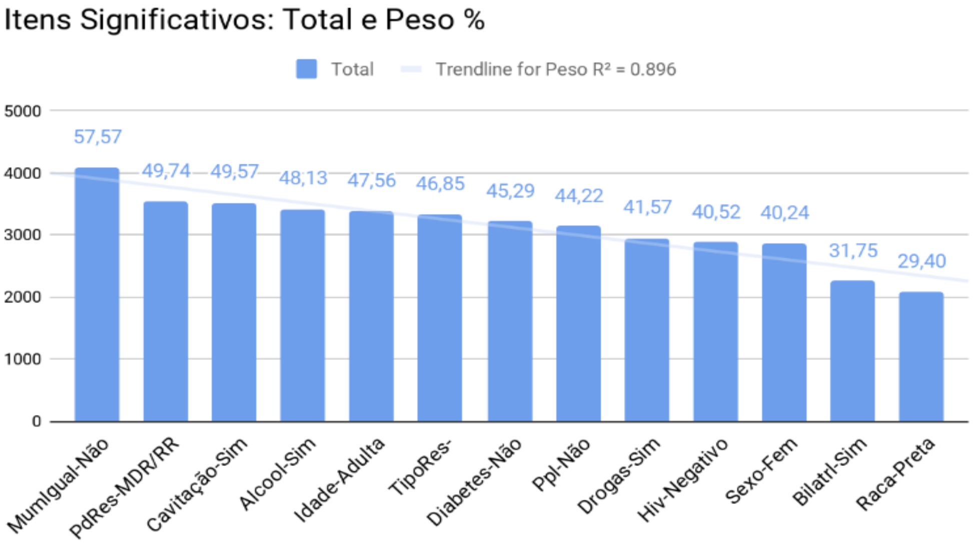


Source: Author

Abandonment is the most occurring Un-favourable Outcome that most patients suffer, Giving up an on going treatment happens alot to most patients than Failing or Death. To understand better what items standout for abandonment, the Figure 16, shows us what unique variables/items stand-out the most for the cause of abandonment, *MumIgual-Não* is a variable for patients that don't live in the town or city where they receive treatment, with the hight weight of 57.57% compared to other items, the second most outstanding item is *PfRes-MDR/RR* (Initial resistance pattern: multidrug-resistance tuberculosis) this variable shows resistance to drugs such as ethambutol, pyrazinamide, or the fluoroquinolones, this variable has a weight of 49.74%. These 13 items were highlighted in the Figure 16 *MumIgual-Não, PdRes-MDR/RR, Cavitação-Sim, Alcool-Sim, Idade-Adulta, TipoRes, Diabetes-Não, Ppl-Não, Drogas-Sim, Hiv-Negativo, Sexo-Fem, Bilatri-Sim, Raça-Preta.*

Exploring the un-favourable outcomes is our main purpose, to accomplish that while we explore abandonment, it would be interesting to compare these varibles of abandonment to other outcomes like Failed and Death, in Table 8, Comparing Abandonment to Failed and Death, we understand *MumIgual-Não* had a weight of 57,57% for Abandon but had 23,70% for (Falencia = Failed) and 16,28% for (Obito = Death), this same comparism is applied to all variable that stood out for Abandonment

Figure 16 – Significant Items: Total x weight (Abandonment)



Source: Author

$LIFT \geq 4.8$

- *raca3=1,mumigual=0,alcool=1,comdrogas=1*
  Race/Color Black or Brown; Receives Treatment from a hospital in a town/city different from his town/city of Residence; Alcoholic and Uses Drugs.

- *sexo2=1,raca3=0,escol4=1,pdres5=1,cavitacao=1,bilateral=0*
  White male with Low Education background (Basic or High School), Multi-resistence ou resistent to rifampicin and gravity of the disease (Cavitation).

- *escol4=0,mumigual=0,hiv3=0,comdrogas=1,bilateral=1*

- *escol4=0,mumigual=0,hiv3=0,comdrogas=1,cavitacao=1*
  These two profiles had confidence of 75% e includes: Patients with medium to high level of education (High School to University), Receives Treatment from a hospital

Table 8 – Comparing Abandonment to Failed & Death

	Failed%	**Death%**
MumIgual-Não	23.70	16.82
PdRes-MDR/RR	48.79	40.52
Cavitação-Sim	41.33	29.83
Alcool-Sim	2.22	0.00
Idade-Adulta	53.12	10.50
TipoRes-Adquirida	45.21	50.83
Diabetes-Não	45.66	14.06
Ppl-Não	50.62	51.12
Drogas-Sim	0.00	0.00
Hiv-Negativo	44.08	24.70
Sexo-Fem	3.74	4.04
Bilatrl-Sim	64.27	47.32
Raca-Preta	23.61	26.70

Source: Author

in a town/city different from his town/city of Residence, HIV Negative Result and bilateral disease or with cavitation.

Table 9 – 4 Unique Profiles with $LIFT \geq 4.8$

Antecedent:ABANDONMENT	**Consequent**	**Count**	**Lift**	**Confidence**	**Support**
raca3=1,mumigual=0, alcool=1,comdrogas=1	encerra2=1	4	6,49	1	0,01
sexo2=1,raca3=0,escol4=1, pdres5=1,cavitacao=1,bilateral=0	encerra2=1	2	6,49	1	0,01
escol4=0,mumigual=0,hiv3=0, comdrogas=1,bilateral=1	encerra2=1	3	4,87	0.75	0,01
escol4=0,mumigual=0,hiv3=0, comdrogas=1,cavitacao=1	encerra2=1	3	4,87	0.75	0,01

Source: Author

4.2.3 Unfavourable Results: Failed

Failed is the second most populated data for un-favourable outcomes with 3,605 valid rules of which 482 rules contains 9 items, 850 rules with 8 items, 969 rules with 7 items and the lowest number of itemset is 2 with 4 valid rules. In the next paragraph we would explore more of these itemset that appeared frequently in common rules as seen in table 10, This information was ploted on a bar chart in figure 17, clearly, the number of rules in relations to itemsets leading to Failed treatment are these unique items.

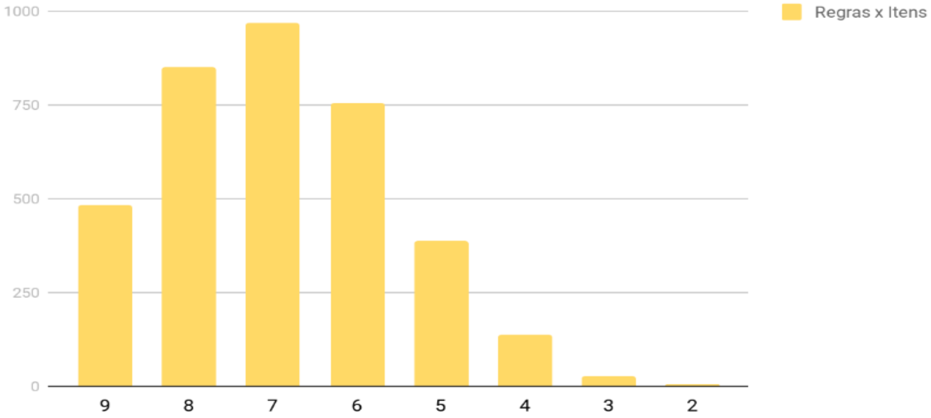
As the second most occurring un-favourable outcomes, Failed imples that the treatment of these profiles of patients failed, these patientes were receiving treatments for

Table 10 – Nº Of Items x Nº of rules

Nº Of Items	**Nº of rules**
9	482
8	850
7	969
6	754
5	389
4	136
3	28
2	4

Source: Author

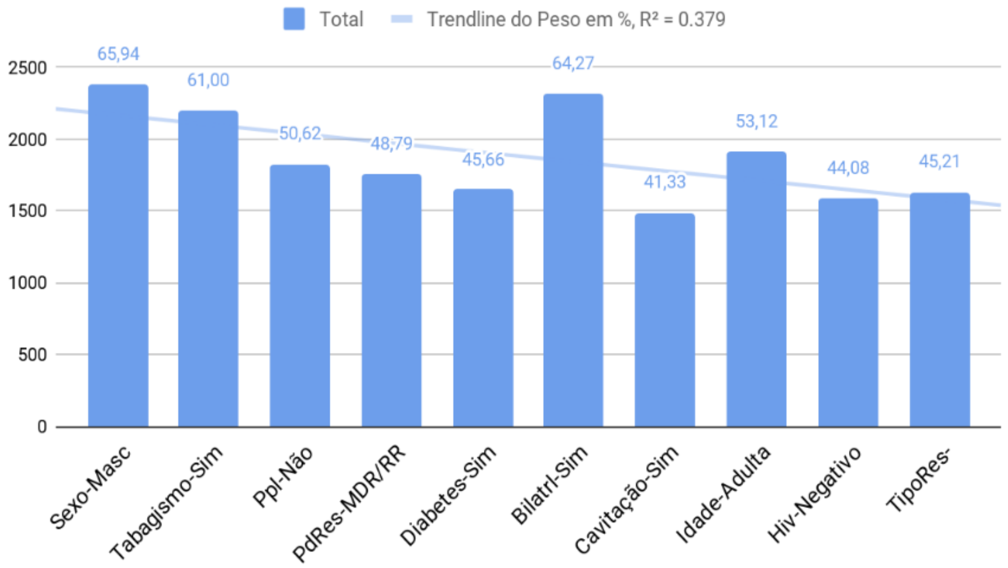
Figure 17 – Plot: Nº of Rules in relations to Items leading to Failed



Source: Author

tuberculosis but the treatment failed, Tuberculosis sometimes might have played a factor but wasn't registered as cured but instead Failed treatment. The figure 18, we will be exploring the unique variables or itemset that results to Failed. Though the bar graph wasn't ploted in incrementing order, but we can still see that *Sexo-Masc* is the first and most re-occurring items from our valid rules with weight of 65,94%, This means male patients suffering from this virus while receiving treatment are at risk of failure, whether or not they were healing and recovering, the second most re-occuring item has a high weight of 64,27% is *Bilatri-Sim* which is a varible for *Doença Bilateral* in otherwords *Bilateral Disease*, with the presence of the variable in a profile it means that the pátient could be suffering from other diseases that affect the lungs, kidney, eyes, ears or joints etc. The third unique item also has a high weight of 61% and that is *Tabagismo-Sim*, This variable when present can change the outcome of a predictive model, this variable indicates profiles smokers, cigarette, tabaco etc. Just to mention a few, we can undertand how this can lead to Failure in treatment. We listed other variables in order of their weights according to figure 18, *Sexo-Masc; Bilatrl-Sim; Tabagismo-Sim; Idade-Adulta; Ppl-Não; PdRes-MDR/RR; Diabetes-Sim; TipoRes-Adquirida; HIV-Negativo; Cavitação-Sim.*

Figure 18 – Significant Items: Total x weight (Failed)



Source: Author

Let's take a paragraph to look at the Table 11, to explore these unique varibles that stoodout for Failed treatment and compare them to Abandonment and Death to understand their weight and impact, *Sexo-Masc* has a very high weight for Failed but only had 0,06% weight for Abandonment, this shows the variable isn't looked or considered by Abandonment, This same variable had a weight of 50,64% for Death, a very high weight and which shows that males are liable to failing more often than females or other genders. *Bilatrl-Sim* had an average high weight for both Abandonment and Death, Like we understood in the previous paragraph, the variable *Bilatrl-Sim* can alter the outcome of a predictive model if present and in most cases it is present. *Sexo-Masc* and *Diabetes-Sim* have 0 weights for Abandonment, which is very interesting and raises the question why ?, doesn't mean they don't appear at all as unique for profiles that abandon treatment ?, the answer to it all is we can't be sure, sometimes its presence in a profile doesn't have any weight to decide outcomes for that profile.

$LIFT \geq 3.9\&Confidenceof33\%$

- *fxet3=1,alcool=1,comtabagis=0,comdrogas=0,cavitacao=1,bilateral=1*
  Teenage/Young adult, Drinks Alcohol but doesn't use illicit drugs or smoker and bilateral disease or with cavitation, this profile is especially for young people.

- *sexo2=0,escol4=1,comdiab=0,comtabagis=0,cavitacao=1,bilateral=1*

- *alcool=1,comtabagis=0,comdrogas=0,cavitacao=1,bilateral=1*
  These two profiles represents Women/Females with low education, nao-diabetic and doesn't smoke.

Table 11 – Comparing Failed to Abandonment & Death

	Abandonment%	**Death%**
Sexo-Masc	0.06	50.64
Tabagismo-Sim	4.45	6.56
Ppl-Não	44.22	51.12
PdRes-MDR/RR	49.74	40.52
Diabetes-Sim	0.00	4.37
Bilatrl-Sim	31.75	47.32
Cavitação-Sim	49.57	29.83
Idade-Adulta	47.56	10.50
Hiv-Negativo	40.52	24.70
TipoRes-Adquirida	46.85	50.83

Source: Author

Table 12 – 3 Unique Profiles with $LIFT \geq 3.9 \& Confidence of 33\%$

Antecedent:Failed	**Consequent**	**Count**	**Lift**	**Confidence**	**Support**
fxet3=1,alcool=1,comtabagis=0, comdrogas=0,cavitacao=1, bilateral=1	encerra2=2	3	4,40	0,37	0,01
sexo2=0,escol4=1,comdiab=0, comtabagis=0,cavitacao=1, bilateral=1	encerra2=2	4	3,91	0,33	0,01
alcool=1,comtabagis=0, comdrogas=0,cavitacao=1, bilateral=1	encerra2=2	3	3,91	0.33	0,01

Source: Author

4.2.4 Unfavourable Results: Death

The outcome *Death* refers to profiles of patientes whose death was caused by tuberculosis, before, during or after treatment. This profile didn't really get cured of the virus and so therefore they died, according to the data mining to obtain rules, there are 2105 valid rules, in Table 13, the most item is 9 for 84 valid rules, a small number of rules compared to how other outcomes portraits but alot can still be derived, this table shows lists of items from 9 to 2 and valid rules from 969 to 4. It is a very interesting table because it allows us to understand why these rules were select by the model as shown in Figure 19.

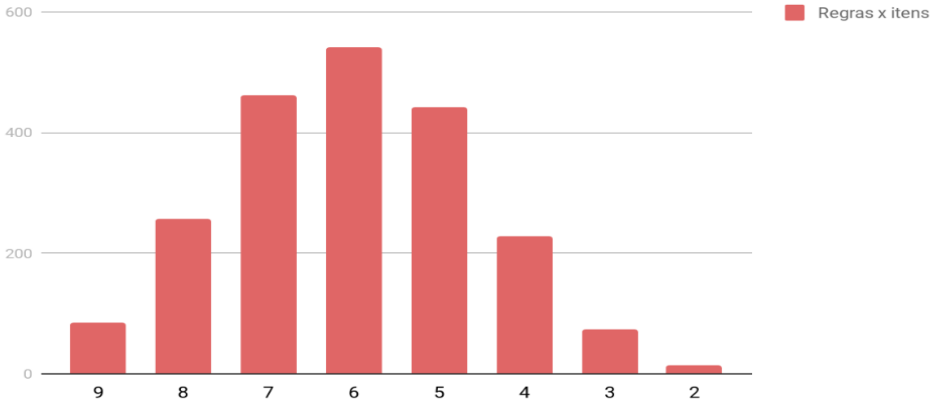
Dying from this virus is tough and painful because it is curable, tuberculosis early stage when diagnosed is very curable, when it becomes drug resistant it also is curable but difficult, from the original database, the profile of patients on the database with Death as their outcome was really low at about 12% of the 1000 registers used. Also, after running our model, a total of 29,007 valid rules were generated of which 2105 of these valid rules are for Death, thats about 7.3% and even lower than the initial value of death from original

Table 13 – Nº Of Items x Nº of rules

Nº Of Items	**Nº of rules**
9	84
8	257
7	462
6	542
5	442
4	229
3	74
2	14

Source: Author

Figure 19 – Plot: Nº of Rules in relations to Items leading to Death

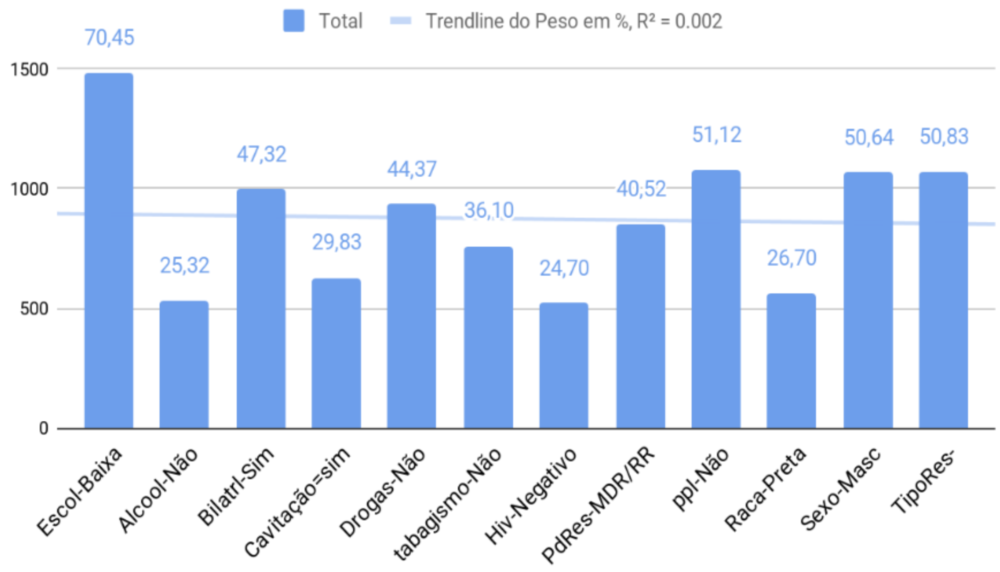


Source: Author

database. In Figure 20, we explore some unique variables or items that leads to patients dying from this disease. the first variable in determining this outcome is *Escol-Baixa*, this indicates very low level of education of patients with a high weight of 70,45%, also means that 70,45% of the valid rules contains this item. The next unique item is *Ppl-Não* with a weight of 51,12%, a little more than half of the valid rules contains this variable, (PPL = No) means indicates that the individual is not an ex-convicted person, in other words have never being to prison. All the unique variables with a very good weight value are *Escol-Baixa; Alcool-Não; Bilatrl-Sim; Cavitação-Sim; Drogas-Não; Tabagismo-Não; HIV-Negativo; Pdres-MDR/RR; Ppp-Não; Raça-Preta; Sexo-Masc; TipoRes-Adquirida.*

The usual comparism was ploted in Table 14, we can explore what these items look like for other outcomes (Abandonment and Failed), the item *Escol-Baixa* has a low weight of 14% when compared to both abandonment and Failed, we can conclude that is not a change altering variable for other outcomes. A few items stoodout from our comparism table such as *Bilatl-Sim* 31,75% abandonment and 64,27% Failed, *Tabagismo-Não* 61.00% Failed, *Ppl-Não* 44,22% Abandonment - 50,62% Failed, *Sexo-Masc* 65,94%.

Figure 20 – Significant Items: Total x weight (Death)



Source: Author

Table 14 – Comparing Death to Abandonment & Failed

	Abandonment%	**Failed%**
Escol-Baixa	14.35	14.62
Alcool-Não	7.10	38.56
Bilatrl-Sim	31.75	64.27
Cavitação=sim	49.57	41.33
Drogas-Não	10.09	41.28
tabagismo-Não	18.92	61.00
Hiv-Negativo	40.52	44.08
PdRes-MDR/RR	49.74	48.79
ppl-Não	44.22	50.62
Raca-Preta	29.40	23.61
Sexo-Masc	0.06	65.94
TipoRes-Adquirida	46.85	45.21

Source: Author

$LIFT \geq 4.0$

- *mumigual=0,hiv3=1,ppl=0,tipores3=1*

  Receives Treatment from a hospital in a town/city different from his town/city of Residence; HIV Negative; PPL Positive (Have being to prison);

- *fxet3=2,escol4=1,mumigual=1,hiv3=0*

  Profiles of patient with Older aged adults, Low level of education, Doesn't Use drugs.

- *escol4=1,hiv3=1,comdrogas=0,ppl=0,tipores3=1*

Old people with Low level of education, HIV Negative, Receives Treatment from a hospital in a town/city different from his town/city of Residence, never being to juventile prison and Aquired Type of Resistant.

Table 15 – 3 Unique Profiles with $LIFT \geq 4.0$

Antecedent:Death	**Consequent**	**Count**	**Lift**	**Confidence**	**Support**
mumigual=0,hiv3=1,ppl=0, tipores3=1	encerra2=3	8	4,48	0,47	0,03
fxet3=2,escol4=1,mumigual=1, hiv3=0	encerra2=3	4	4,23	0,44	0,01
escol4=1,hiv3=1,comdrogas=0, ppl=0,tipores3=1	encerra2=3	6	4,08	0,43	0,02

Source: Author

4.3 CONFIGURATION: SUPP[0,01], CONF[0,1]

A second result was necessary for analises, this time the model was ran using a high value for support and confidence. The effect of this high value when running our model means fewer valid rules in our coincedent dataset but a better and more refine valid rules.

We explored only unfavourable outcomes (Abandon, Failed and Death), the coincedent dataset contained a total of 18,282 valid rules, from which we had 4460 valid rules for Abandonment, 23 valid rules for Failed and 655 for Death. In the following subsections, we would be exploring this unfavourable outcomes for valid rules with different range of lift and count.

4.3.1 Failed

The resulting outcomes for Failed was always low from our original database to datasets derived from the models. In this model there are 23 valid rules from which we explored this results. Filtering this dataset for values of Count between 4 to 8 and values of Lift between (1.16 - 2.3), (1,88 - 1,9), (1,97 - 2,3), these following rules were significant.

- **Lift=(1.16 - 2.3)**:
  raça3=1*: individuals of black or brown race* this item was found in almost all the rules for the exception of one rule. There were 8 rules with the highest values for Count (7 to 8), the Lift of these items ranged from 1.5 to 2.3 and the items are ***bilateral=1****: The presence of a bilateral disease;* ***cavitacao=1****: The presence of Cavitation;* ***alcool=0****:Not alcoholic;* ***comdrogas=0****:Not a drug addict;* ***comtabagis=0****: Not a smoker* this item (comtabagis=0) appeared in 3 important rules, ***fxet3=1****: Individuals between the age of 15 to 59;* ***raça3=1****: individuals of black or brown race;* ***hiv3=0****:HIV Negative;* ***(pdres5=1)****: Pattern of resistance is MDR/RR;* ***tipores3=1****: Type of Resistance is Acquired* This item only appeared in rules with Count equals to 8.

- **Lift=(1,97 - 2,3)**:
  There were 6 rules remaining when we applied these filter for Lift and these rules had items equivalent as to the items above.

- **Lift=(1,88 - 1,9)**:
  sexo2=0*: Male/Men;* ***comtabagis=0****: Not a smoker;* ***raça3=1****: individuals of black or brown race;* ***tipores3=1****: Type of Resistance is Acquired;* In half of these rules, these iteme were very present and significant ***tipores3=1****: Type of Resistance is Acquired;* ***comdrogas=0****:Not a drug addict;* ***ppl=0****: Never Convicted.*

4.3.2 Death

We analized the outcome Death with 655 valid rules from the coincedent dataset, to acquire some new knowledge or information we filtered our valid rules with Lift and Count.

- **Lift=(1,12 - 3,50) and Count=(4 - 24)**
  An important note to take is that, there were 42 valid rules with Lift greater or equal to 3.0, and also, one item was outstanding for rules with count of 11, ***escol4=0***:*Individuals with 0 to 7 years basic education or in other words low education* this particular item appears in all 11 rules. With high value for lift (3.0 - 3.5), ***tipores3.1=1***:*Type of Resistance is Acquired* appeared in half of the rules and ***escol4=0***:*Individuals with 0 to 7 years basic education* this item appeared in 2/3 of the rules.

- **Lift=(3.2 - 3,50)**
  When we applied this filter, we had only 15 rules and from which we had the following outstanding items and they are, ***sexo2.0***: *Male individuals;* ***escol4=0***:*Individuals with basic/low education;* ***bilateral=1***: *The presence of a bilateral disease; and* ***comdrogas=0***:*Not a drug addict.*

  With Lift = 3.5, there were 4 valid rules and in them we found the following outstanding items ***sexo2.0***: *Male individuals;* ***fxet3=1***: *individuals of age 15 to 59 (teanage/adults);* ***escol4=0***:*Individuals with basic/low education;* ***bilateral=1***: *The presence of a bilateral disease;* ***cavitacao=1***: *The presence of Cavitation; and* ***comdrogas=0***:*Not a drug addict.*

  Another information we found is, **mumigual=0** when an individual receives treatment and lives in different town/city, the lift are usually higher than **mumigual=1** when patient receives treatment and lives in the same town/city.

4.3.3 Abandonment

Abandonment has the most registers from our original database, so we were looking forward into seeing more of the outcome in our model. From the coincedent dataset, There were a total of 4460 valid rules for Abandonment. To obtain the best items from the rules we only looked at the value of Count at 4 - 34 and Lift between 1.1 - 3.6.

- **Lift=(1.1 - 3.6)** The following item ***alcool=0***:*Not alcoholic;* appeared in 30 valid rules, but on the other hand ***alcool=1***: *is alcoholic* appear in over half of the valid rules (2,357), they all had lift between 1,5 to 2,5.

- ***bilateral=0***: *No bilateral disease* appears in 72 valid rules with Lift (1,13 - 2,47), but on the other hand, ***bilateral=1***: *The presence of a bilateral disease* appears in about 1/4 of the total valid rules at a number of 1142 rules and has a Lift of (1,13 - 3,57).

- ***cavitacao=1****: The presence of Cavitation* appears in over half of the valid rules at a number of 2232 rules with Lift ranging from (1.13 - 3.63).

- ***comdiab=0****: Negative Diabetes* appears in 2127 rules, that's almost half of the total valid rules with Lift ranging from (1,10 - 3,63), but ***comdiab=1****: Positive Diabetes* doesn't appear in our valid rules.

- ***comdrogas=0****:Not a drug addict* appears in 48 valid rules with Lift (1,13 - 1,88), but on the other ***comdrogas=1****:Is a drug addict* appears in 1527 rules with Lift (1,17 - 3,63).

- 4 rules for ***escol4=0****:Individuals with basic/low education*, in all rules there were ***mumigual=0*** *when an individual receives treatment and lives in different town/city; and* ***fxet3=1****: individuals of age 15 to 59 (teanage/adults)* with Lift (2,47). but also, there were 799 rules with ***escol4=1****:Individuals with high education* with Lift (1,10 - 3,29).

- ***hiv3=0*** *HIV Negative* appeared in 1269 rules with lift (1.13 - 3.59), and there wasn't any rules for HIV positive or Ignored Result.

- ***mumigual=0*** *when an individual receives treatment and lives in different town/city* appears in 1691 rules with Lift (1,10 - 3,63). But ***mumigual=1*** *when patient receives treatment and lives in the same town/city* appears in only 42 rules with Lift (1,19 - 1,74).

- The following items appeared in more or almos half of the total valid rules, ***fxet3=1****: individuals of age 15 to 59 (teanage/adults)* appeared in 2391 rules with Lift (1.10 - 3.63), ***(pdres5=1)****: Pattern of resistance is MDR/RR* appeared in 2281 rules with Lift ranging between (1,10 - 3,63).

- ***raça3=1****: individuals of white race* appeared in very few rules with 27 rules and also, with low Lift of (1,22 - 1,88) but ***raça3=1****: individuals of black or brown race* appeared in 1456 rules with Lift (1,10 - 3,63).

- The following items appeared significantly in many rules, ***ppl=0****: Never Convicted* in 1977 rules with Lift (1,10 - 3,59), ***sexo2=1****: Female/Women* appeared in 1828 rules with Lift (1,10 - 2,82), and ***tipores3=1****: Type of Resistance is Acquired* in 1828 rules with Lift (1,10 - 3,47).

- The item ***comdrogas=1**:Is a drug addict* was a present factor in every rules com Lift $\geq$ 3.0. The Lift increases for rules that possess the following items ***mumigual=0*** *when an individual receives treatment and lives in different town/city* and ***raça3=1**: individuals of black or brown race.* The moment this item ***cavitacao=1**: The presence of Cavitation* is added, Lift increases. In rules that don't posses these two items ***mumigual=0*** *when an individual receives treatment and lives in different town/city* and ***raça3=1**: individuals of black or brown race*, they appear to have ***bilateral=1**: The presence of a bilateral disease* and ***alcool=1**: is alcoholic.*

5 CONCLUSION

In conclusion, dying from tuberculosis is preventable (CARVALHO, 2012), Even so, more than one third of people who died with pulmonary Tuberculosis had this condition as the underlying cause, a relationship that was also observed in a previous study (RS, 2018 Sep 7). The profiles of patients of abandonement stood out having ($lift \geq 4.8$), with 100% confidence: race/color black/brown; treatment in a town different from the town of residence and being a drug and alcohol user; white men with low education (elementary or less); multidrug resistance or rifampicin resistance and disease severity (cavitation). With 75% confidence: having a high school degree or more; treatment in a regions other than the residence; not having HIV infection and bilateral disease or with cavitation.

Death patients had a lift above 3.9 and 33% confidence, the highlight being the severity of the disease (bilaterality with cavitation) associated with: being an alcohol user, but not an illicit drug user or a smoker, especially if you are not elderly; women with low education, non-diabetic and non-smokers.

The profiles of patients that passed away were related to acquired resistance and HIV infection: treatment in a different city/town different from where they lived; patient with low education; non-drug user and not deprived of liberty; elderly with low education; not HIV treated in the municipality of residence.

With this work, it was possible to identify the patterns, with different profiles of the patients that allow the prediction of outcomes/results, allowing future construction of alarming cases according to the characteristics involved in the rules obtained in the training stage.

With this configuration, it was possible to identify the patterns, with different profiles of the patients that allow the prediction of outcomes, allowing future construction of alarms according to the characteristics involved in the rules obtained in the training stage.With the help of Apriori allowed the identification of patient profiles that can compose the future construction of alarms to increase the effectiveness of the treatment.

BIBLIOGRAPHY

A, F.; AJ, H.; CA, P. Risk factors for multidrug resistant tuberculosis in europe: a systematic review. *Thorax*, 2006.

AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: CITESEER. **Proc. 20th int. conf. very large data bases, VLDB**. [S.l.], 1994. v. 1215, p. 487–499.

BARTHOLOMAY, O. P. Vigilância da tuberculose droga resistente no brasil e fatores associados aos desfechos desfavoráveis de tratamento. **Tese Doutorado (Medicina Tropical)- UnB**, 2019.

BARTHOLOMAY, P. et al. Sistema de informação de tratamentos especiais de tuberculose (site-tb): histórico, descrição e perspectivas. **Epidemiologia e Serviços de Saúde**, SciELO Brasil, v. 28, p. e2018158, 2019.

BECKER, R. A. A brief history of s. In: . [S.l.: s.n.], 2004.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, p. 123–140, 1996.

CARVALHO et.al. **Mineração de dados aplicada à fisioterapia. Fisioterapia e Movimento**. [S.l.]: v. 25, n.3, p. 595-605, 2012.

CC, B. et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the xpert mtb/rif test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. **Lancet**, 2011.

CONTROL, C. for D.; (CDC), P. Tuberculosis (tb) - fact sheet. 2023. Disponível em: https://www.cdc.gov/tb/publications/factsheets/general/tb.htm. Acesso em: 24 July. 2023.

CRAN, R.-P. Writing r extensions and deploy to cran. 2023. Disponível em: https://cran.r-project.org/doc/manuals/R-exts.html. Acesso em: 14 July. 2023.

D., H.; S., R.-G.; E., R. Evaluation of the genotype mtbdrplus assay for rifampin and isoniazid susceptibility testing of mycobacterium tuberculosis strains and clinical specimens. **J Clin Microbiol**, 2007.

D, K.; AL et. Drug resistance mechanisms in mycobacterium tuberculosis. **J. Drug Target**, 2019.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. [S.l.]: CRC press, 1994.

FAYYAD, U. et al. The kdd process for extracting useful knowledge from volumes of data. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 39, n. 11, p. 27–34, nov 1996. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/240455.240464.

GITHUB, i. Software development and version control using git. 2023. Disponível em: https://github.com/. Acesso em: 14 July. 2023.

GLYNN, J. et al. Patterns of initial and acquired antituberculosis drug resistance in karonga district, malawi. **The Lancet**, v. 345, n. 8954, p. 907–910, 1995. ISSN 0140-6736. Disponível em: https://www.sciencedirect.com/science/article/pii/S0140673695900160.

HAN, J.; KAMBER, M.; PEI, J. Data mining: Concepts and techniques. In: HAN, J.; KAMBER, M.; PEI, J. (Ed.). **Data Mining (Third Edition)**. Third edition. Boston: Morgan Kaufmann, 2012, (The Morgan Kaufmann Series in Data Management Systems). p. 1–38. ISBN 978-0-12-381479-1. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780123814791000010.

JICTAC. Xlii jornada giulio massarani de iniciação científica, tecnológica, artística e cultural (jictac 2020 – edição especial. **Evento UFRJ**, 2020. Disponível em: https://www.even3.com.br/jictac2020ufrj/.

JONNALAGADA, S.; HARRIES, A.; ZACHARIAH, R. e. a. The timing of death in patients with tuberculosis who die during anti-tuberculosis treatment in andhra pradesh, south india. **BMC Public Health. v. 11**, 2011.

KA, A. et al. Treatment outcomes in patients with multidrug-resistant tuberculosis in north-west ethiopia. **Trop Med Int Health**, 2017.

KJ, S.; S, K.; ML, R. Multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis. **Cold Spring Harb Perspect Med**, 2015.

ORGANIZATION, W. H.; W.H.O. Policy guidance on drug-susceptibility testing (dst) of second-line antituberculosis drugs. 2008.

R-PROGRAMMING-LANGUAGE. The r project for statistical computing. 2023. Disponível em: https://www.r-project.org/. Acesso em: 14 July. 2023.

ROCHA.MS et al. Uso de linkage entre diferentes bases de dados para qualificação de variáveis do sinan-tb e a partir de regras de scripting. **Cad Saúde Pública**, 2019. Disponível em: http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0102-311X2019001404001&lng=en&nrm=iso&tlng=pt.

RODRIGO, T. et al. Factors associated with fatality during the intensive phase of anti-tuberculosis treatment. **PLoS ONE, v. 11, n. 8**, 2016.

RS, R. M. O. G. S. V. A. F. C. C. Effect of inpatient and outpatient care on treatment outcome in tuberculosis: a cohort study. **Rev Panam Salud Publica**, 2018 Sep 7.

RSTUDIO. The r studio ide. 2023. Disponível em: https://en.wikipedia.org/wiki/RStudio. Acesso em: 14 July. 2023.

S., T. et al. Tuberculosis: progress and advances in development of new drugs, treatment regimens, and host-directed therapies. **Lancet Infect Dis**, 2018.

TBDR19PREDICTION.GITHUB. The r-library tbdr19prediction github repository. 2023. Disponível em: https://github.com/chrismomentus/tbdr19prediction. Acesso em: 30 August. 2023.

W.H.O. Global tuberculosis report 2022. 2022. Disponível em: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022. Acesso em: 14 July. 2023.

WITTEN, F. H. **Data Mining, Practical Machine Learning Tools and Techniques**. [S.l.]: ACM New York, NY, USA, 2002. v. 31. 76–77 p.

WN, B. P. P. R. D. F. P. D. de A. Brazilian cohort study of risk factors associated with unsuccessful outcomes of drug resistant tuberculosis. **BMC Infect Dis.**, 2021 Oct 9.

ZHANG, C.; ZHANG, S. **Association rule mining: models and algorithms**. [S.l.]: Springer, 2002.

ZHANG, Y.; WW, Y. Mechanisms of drug resistance in mycobacterium tuberculosis. **Int J Tuberc Lung Dis**, 2009.

ZUMLA, A.; P, C. S. N. Advances in the development of new tuberculosis drugs and treatment regimens. **Nat Rev Drug Discov**, 2013.

GLOSSARY

Apriori An algorithm for frequent item set mining and association rule learning over relational databases.

Baggings To generates different subsets of the training dataset by selecting data points at random and with replacement.

Bootstrap A type of resampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample.

Confidence A percentage value that shows how frequently the rule head occurs among all the groups containing the rule body.

DST (Drug susceptibility testing) DST involves testing the susceptibility of M. tuberculosis isolates to different anti-TB drugs.

LHS The number of transactions in the Left Hand Side of a Rule.

Lift The ratio of the confidence of the rule and the expected confidence of the rule.

MDR-TB (Multidrug-resistant TB) is caused by TB bacteria that are resistant to at least isoniazid and rifampin, the two most potent TB drugs.

R Programming Language A programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing.

RHS The number of transactions in the Right Hand Side of a Rule.

SIM (Sistema de Informações sobre Mortalidade) Established by the ministry of Health Brazil, it is the database of the unification of more than forty models of instruments used over the years to collect data on mortality in the country .

SINAN (Sistema de Informação de Agravos de Notificação com dados sobre tuberculose) aims to collect, transmit and disseminate data generated by the Epidemiological Surveillance System of the three spheres of government, through a computerized network..

SITE-TB (Sistema de Informação de Tratamentos Especiais da Tuberculose) A system for the notification and follow-up of TB cases that are indicated for special treatments, either due to the occurrence of adverse reactions, toxicity or certain comorbidities that make it impossible to use the Basic Scheme, or due to resistance .

Support The frequency of the occurence of an item .

TB A contagious and potentially serious infectious disease caused by the bacterium Mycobacterium tuberculosis.

TBDR A form of tuberculosis caused by bacteria that are resistant to one or more of the primary anti-TB drugs used to treat the disease.

XDR-TB (Extensively drug-resistant TB) is a rare type of multidrug-resistant tuberculosis (MDR TB) that is resistant to isoniazid, rifampin, a fluoroquinolone.