PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Gabriela Maria Rocha Bolaina

ESTUDO SOBRE O COVID 19 NO ESTADO DE MINAS GERAIS

Belo Horizonte
2022

Gabriela Maria Rocha Bolaina

ESTUDO SOBRE O COVID 19 NO ESTADO DE MINAS GERAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2022

SUMÁRIO

1. Introdução

1.1. Contextualização

Os vírus da família *Coronaviridae* causam várias doenças em homens e animais, principalmente no trato respiratório. As partículas virais são esféricas e apresentam projeções em forma de espículas, que geram visualmente uma coroa, vem daí a denominação de coronavírus. Um vírus dessa família atualmente ganhou destaque com a chegada da pandemia do covid-19, o Sars-CoV-2.

O primeiro registro de infecção humana oficial de covid-19 (*coronavirus disease* 2019) foi de um paciente hospitalizado no dia 12 de dezembro de 2019 em Wuhan, China. Mas posteriormente em novos estudos retrospectivos, foi detectado um caso clínico com sintomas da doença em 01/12/19.

No primeiro artigo cientifico publicado por pesquisadores chineses, sugeriu que o novo coronavírus tenha se originado de morcegos (um reservatório já identificado para o Sars-CoV, agente da Sars), visto que materiais genéticos coletados de um infectado continha um vírus cujo genoma mostra relação com o coronavírus causadores da Sars (síndrome respiratória aguda grave) e Mers (síndrome respiratória do Oriente Médio). (GRUBER, Arthur, 2020)

Devido ao surgimento recente da doença, ainda há poucos estudos estatísticos sobre variáveis de influência no contagio e óbito por covid-19, visto isso, o objetivo deste trabalho é encontrar relações entre o contagio e óbito por covid-19 em pessoas residentes em MG e as variáveis de estudo: sexo, faixa etária, comorbidades, etnia, necessidade de internação e UTI.

Para atingir esse objetivo, usaremos ferramentas poderosas da estatística: os testes de hipóteses. Os testes de hipótese fazem parte da estatística inferencial, a qual tem seu uso para formar conclusões e fazer inferências sobre as populações baseada em dados coletados por amostras.

Um exemplo de teste de hipótese é o teste de Qui-Quadrado para independência, que utiliza tabelas de contingência como base. Um dos principais objetivos de se construir tabelas de contingência é analisar a distribuição conjunta de duas variáveis qualitativas e descrever a associação entre elas. Caso haja suspeita de uma certa dependência entre as variáveis, como

por exemplo ser fumante e surgimento de câncer de pulmão, construindo tal tabela, podemos buscar evidencia estatística de que essas duas variáveis possuem certo grau de associação.

Ao fazer esse tipo de análise, buscando evidencia estatística, estamos realizando um teste de hipóteses. O teste que melhor se adequa a esse tipo de pesquisa é o Qui-Quadrado de independência, nele testam-se as seguintes hipóteses:

$$\begin{cases} H_0: As\ variáveis\ são\ independentes \\ H_1: As\ variáveis\ não\ são\ independentes \end{cases}$$

Um p-valor próximo de zero, indica que a hipótese de independência (não associação) é verdadeira. Entretanto, para a validade do teste, é necessário que algumas pressuposições sejam respeitadas:

- Os dados são selecionados aleatoriamente
- Todas as frequências esperadas são maiores ou iguais a 1
- Não mais de 20% das frequências esperadas são inferiores a 5%
- Utilizado preferencialmente para amostras grandes, com tamanho maior ou igual a 30.

Satisfazendo-se esses critérios, o teste é eficaz na análise de associação entre as variáveis. Em caso de quebra do segundo ou terceiro requisitos, há a possibilidade de agrupa-los em uma única categoria.

Quando o teste de Qui-Quadrado de independência não tem os requisitos de tamanho de amostra ou frequência esperada, mesmo agrupando classes quando possível, atingidos podemos recorrer ao teste exato de Fisher. Ele também é um teste de significância estatística utilizado para a análise de tabelas de contingência.

Embora na pratica o teste exato de Fisher seja utilizado para amostras pequenas, ele é valido para todos os tamanhos de amostras. Esse é um teste da categoria de testes exatos, chamados assim por conta da significância do desvio de uma hipótese nula que pode ser calculada exatamente, ou seja, ele não depende de uma aproximação que se torna exata no limite conforme o tamanho da amostra cresce para o infinito, como a maioria dos testes estatísticos.

O teste exato de Fisher pode ser utilizado em dados categóricos para examinar a associação entre dois tipos de classificação. Aqui está uma limitação do teste, que diferentemente do teste de Qui-Quadrado que não estipula um limite de classificações, o teste exato de Fisher pode ser apenas utilizado em tabelas de contingência de tamanho 2x2 (duas

linhas e duas colunas). No cálculo do p-valor do teste, assim como no Qui-Quadrado, as hipóteses nulas e alternativa levadas em consideração são, respectivamente, de independência e dependência. Ou seja:

$$\begin{cases} H_0: As\ variáveis\ são\ independentes \\ H_1: As\ variáveis\ não\ são\ independentes \end{cases}$$

Uma outra maneira de avaliar estatisticamente grupos de risco para uma determinada doença é utilizando o teste de comparação de proporções. Os testes para proporções são utilizados quando temos duas variáveis, nominais ou ordinais, e buscam responder se a proporção esperada é estatisticamente igual à observada, ou seja:

$$\begin{cases} H_0: A\ proporção\ na\ amostra\ é\ igual\ à\ esperada \\ H_1: A\ proporção\ na\ amostra\ é\ diferente\ da\ esperada \end{cases}$$

Após analises estatísticas, que ajudam no entendimento e comportamentos das bases analisadas, será também aplicado modelo de machine learning para fazer a classificação dos dados. Para aplicar esse modelo, as bases de óbitos e casos confirmados serão unidas e teremos duas possíveis evoluções: óbito ou recuperação. Visto isso, o modelo visará classificar os pacientes em recuperados ou que vieram a óbito. O modelo utilizado neste trabalho foi a árvore de decisão.

1.2. O problema proposto

Por não apresentar vacinas suficientes para todos nem medicamentos disponíveis para o tratamento da Covid-19, é importante entender quais características apresentam as pessoas mais suscetíveis ao óbito e contagio.

Para entender esse comportamento, foram usadas as bases de dados do Governo do Estado de Minas Gerais sobre residentes que foram contagiados e quais foram a óbito em decorrência da covid-19, a qual nos ajudará a encontrar um grupo de risco para a doença. Para isso, iremos analisar dados da triagem de pacientes, no período de janeiro à dezembro.

Para um melhor entendimento do comportamento da doença, será criado um modelo de machine learning de classificação, que buscará classificar pacientes em possíveis recuperações ou óbitos.

Vale ressaltar que embora o sistema de saúde atendesse residentes de outros estados, os dados levados em consideração foram apenas dados de residentes de MG.

1.3. Objetivos

O objetivo deste trabalho é encontrar possíveis relações do COVID-19 com as variáveis de estudo: sexo, faixa etária, comorbidades, etnia, necessidade de internação e UTI. Dessa maneira, espera-se encontrar um grupo de maior risco para a doença, tanto para contagio quanto para óbitos. Além de usar modelo de machine learning para avaliar a possibilidade da evolução do paciente, entender se ele tem mais chances de se recuperar ou vir a óbito.

2. Processamento/Tratamento de Dados

Os dados utilizados neste trabalho foram obtidos no Portal de Dados Abertos - Governo de Minas Gerais (https://dados.mg.gov.br/) e da Rede SUAS (http://blog.mds.gov.br/redesuas/) (fonte: IBGE).

A primeira fonte de dados corresponde aos óbitos por COVID-19.

Tabela 1: Dicionário de dados óbitos por COVID-19.

Nome da coluna	Descrição	Tipo
Paciente	Cada paciente é identificado com um único número para preservação de sua identidade	Numérico
Sexo	Sexo do paciente	Binário: M ou F (Respectivamente Masculino ou Feminino).
Idade	Idade do paciente	Numérico
Municipio_Residencia	Município em que o paciente reside	Texto
Data_Obito	Data do óbito do paciente	Data no formato YYYY-MM-DD
Comorbidade	Informa se o paciente possui ou não comorbidade	Binário: Sim ou Não.

A segunda fonte de dados corresponde aos casos confirmados por COVID-19.

Tabela 2: Dicionário de dados casos confirmados por COVID-19.

Nome da coluna	Descrição	Tipo
URS	Unidade Regional de Saúde do município de residência do paciente	Texto
Micro	Microrregião de saúde do município de residência do paciente	Texto
Macro	Macrorregião de Saúde do município de residência do paciente	Texto
ID	Cada paciente é identificado com um número para preservação de sua identidade. O ID não representa contagem numérica dos casos confirmados.	Numérico
Data_Notificacao	Data de notificação do caso	Data
Classificacao_caso	Classificação do caso	Texto
Sexo	Sexo do paciente	Binário: Masculino ou Feminino

Idade	Idade do paciente	Numérico
Faixa_etaria	Faixa Etária do paciente	Texto
Municipio_Residencia	Município em que o paciente reside	Texto
Codigo	Código IBGE do município de residência do paciente	Numérico
Comorbidade	Informa se o paciente possui ou não comorbidade	Binário: Sim ou Não.
Evolucao	Evolução do paciente	Texto
Internacao	Informa se o paciente com caso confirmado de COVID 19 foi internado, com possibilidade de preenchimento SIM ou NÃO.	Binário: Sim ou Não.
UTI	Informa se o paciente com caso confirmado de COVID 19 internado, precisou de UTI (Unidade de Terapia Intensiva) com possibilidade de preenchimento SIM ou NÃO.	Binário: Sim ou Não.
Etnia	Etnia do paciente	Texto
Data_Atualizacao	Data de upload do arquivo	Data
Origem_da_Informacao	Fonte do dado correspondente a cada linha do arquivo	Texto

Com intenção de enriquecimento de dados, foi acrescida a terceira fonte de dados corresponde a dados demográficos dos municípios.

Tabela 3: Dicionário de dados demográficos dos municípios.

Nome da coluna	Descrição	Tipo
IBGE	Código IBGE do município de residência do paciente	Numérico
UF	Estado	Texto
Município	Município	Texto
Região	Região que o município se encontra	Texto
População	População em 2010	Numérico
Porte	Porte do município	Texto
Capital	Informa se o município é ou não a capital do estado	Texto

Com intenção de classificação de dados, foi acrescida a quarta fonte de dados corresponde a idade e faixa etária correspondente. (Fonte: Elaborado pela autora).

Tabela 4: Dicionário de dados idade e faixa etária.

Nome da coluna	Descrição	Tipo
Idade	Idade	Numérico
Faixa_Etaria	Faixa etária correspondente a idade	Texto

Com intenção de estudo ao longo do tempo, foi acrescida a quinta fonte de dados corresponde as datas. (Fonte: Elaborado pela autora).

Tabela 5: Dicionário de dados idade e faixa etária.

Nome da coluna	Descrição	Tipo
Data	Data	Data no formato DD-MM-YYYY

Temos os seguintes relacionamentos entre as fontes de dados:

Tabela 6: Relações nas bases de dados

Tabela 1	Tabela 2	Campos	Relação
Municípios	Casos confirmados	IBGE e CODIGO	1:*
Municípios	Óbitos	IBGE e IBGE	1:*
Casos confirmados	Datas	Data_Notificacao e Data	*:1
Casos confirmados	Faixa etária	Idade e Idade	*:1
Óbitos	Faixa etária	Idade e Idade	*:1
Óbitos	Datas	Data_Obito e Data	*:1

A coluna Campos indica respectivamente o campo correspondente na tabela 1 e tabela 2, assim como a coluna Relação. A coluna relação indica a cardinalidade dos relacionamentos, sendo todos 1 para muitos (1:* ou *:1). Podemos ver os relacionamentos na imagem a seguir:

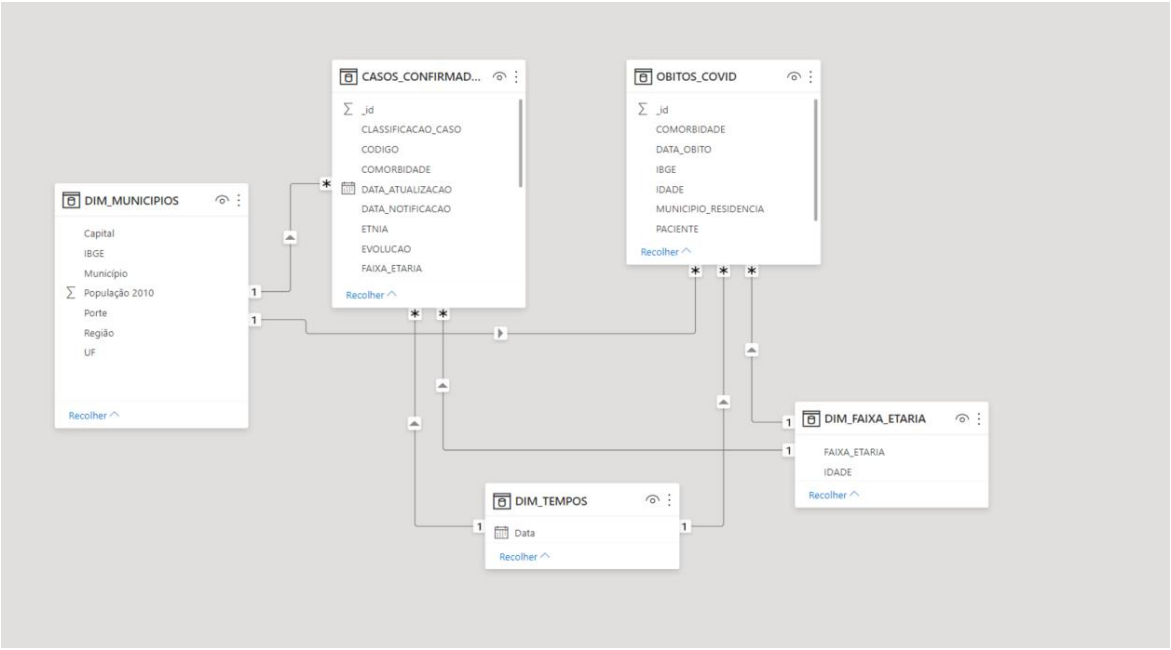


Imagem 1: Relacionamentos na base de dados.

As bases obtidas estavam muito organizadas e não necessitou de muito tratamento, as únicas bases que precisaram ser tratadas foram as de óbitos e casos confirmados.

Para ambas as bases, os registros foram filtrados e residentes de outros estados foram excluídos da base de análise, visto que o local de estudo é o estado de Minas Gerais, as bases resultantes apresentaram, desse modo, 2825 e 45729 registros validos para óbitos e casos confirmados, respectivamente.

Alguns registros da base de confirmados apresentaram códigos do IBGE em branco, os quais foram *inputados* manualmente. Como a base de óbitos não apresentava essa coluna, a mesma foi criada e os valores inseridos com base em ligações com a dimensão de cidades.

Para fins de análise estatística, valores ausentes (NA ou Não Informado), foram filtrados da análise em especifico, não apresentando exclusão total, foram excluídos apenas quando precisava-se da informação em especifico. Optou-se pelo não preenchimento pois tratam-se de dados pessoais e a base de dados é grande.

3. Análise e Exploração dos Dados

A etapa inicial de qualquer análise de dados é a análise descritiva, essa é o estudo dos dados coletados. Ela é usada para organizar, resumir e descrever as características dos conjuntos de dados.

São várias as ferramentas da analise descritiva, tais como vários estilos de gráficos e tabelas e medidas de síntese como porcentagens, índices e médias. Vale ressaltar que a condensação de dados resulta na perda de informação, entretanto é mínima comparada ao ganho de clareza na interpretação.

Outro objetivo da descrição é identificar anomalias e até mesmo o registro incorreto de valores, assim como dados discrepantes (Que se afastam muito da tendencia geral do conjunto).

Inicialmente analisamos a base de dados sobre os óbitos.

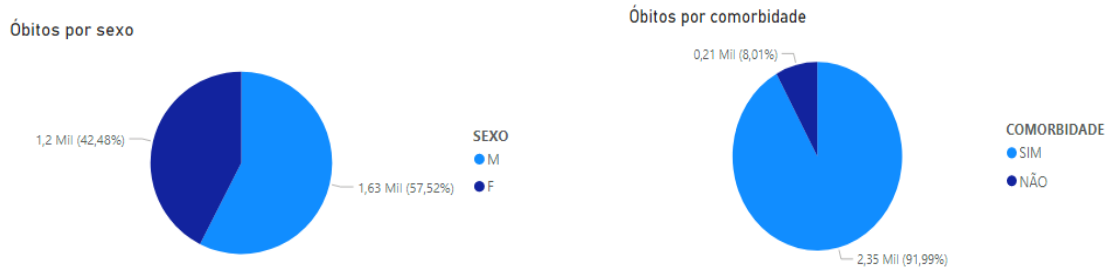


Imagem 2 – Óbitos por sexo e Óbitos por comorbidade

Visualmente, podemos notar que enquanto a variável Sexo apresenta estar distribuída como o esperado, o mesmo não ocorre com a variável Comorbidade, a qual parece apresentar forte relação entre óbito e presença de comorbidade.

Agora analisa-se a quantidade de óbitos frente a faixa etária:

Visualmente, o aumento da idade pode estar relacionado com a quantidade de óbitos, sendo os pacientes de 60 a 89 anos os mais suscetíveis a doença.

Tabela 7: Óbitos por faixa etária

Faixa Etária	Qtd de óbitos	Freq. relativa de óbitos
<1ANO	2	0,07%
1 A 9 ANOS	3	0,11%
10 A 19 ANOS	3	0,11%

20 A 29 ANOS	27	0,96%
30 A 39 ANOS	70	2,48%
40 A 49 ANOS	178	6,30%
50 A 59 ANOS	355	12,57%
60 A 69 ANOS	619	21,91%
70 A 79 ANOS	728	25,77%
80 A 89 ANOS	657	23,26%
90 OU MAIS	183	6,48%

Visualmente, para o óbito o grupo de risco aparenta ser: pessoas com comorbidades e com idades entre 60 e 89 anos, o sexo parece não interferir no curso da doença.

Após analisarmos os óbitos, iremos analisar os pacientes infectados pela COVID-19.

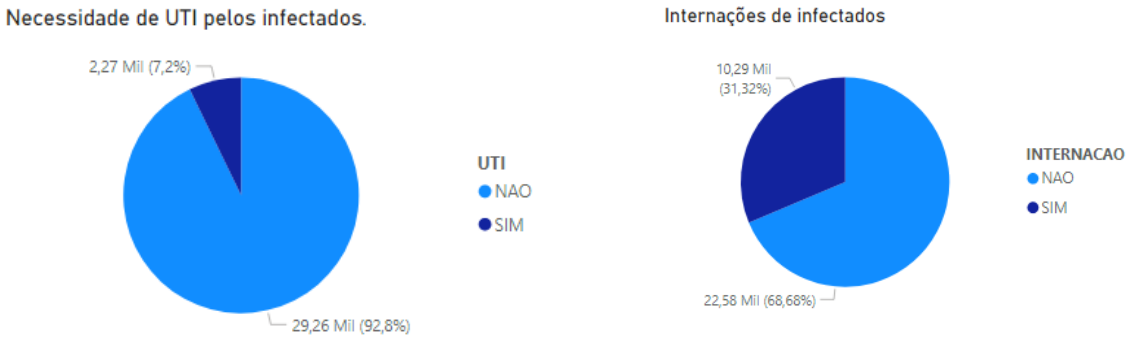


Imagem 3: Necessidade de internação ou UTI pelos infectados

Imagem 4: Sexo dos contagiados

Imagem 5: Comorbidade dos contagiados

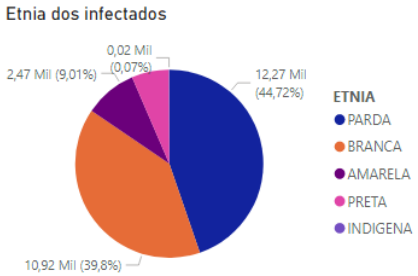


Imagem 6: Etnia dos contagiados

Visualmente, assim como nos óbitos, a variavel sexo parece não ter influência sobre o paciente ser infectado pelo vírus, mas ao contrario do que ocorreu em óbitos, comorbidade também parece não ter influência sobre o contágio. Nota-se que as etnias que apresentam maior incidencia da doença é a branca e a parda. Vê-se também que a minoria dos infectados precisa de UTI ou internação.

Tabela 8: Casos confirmados por faixa etária

Faixa Etária	Qtd de infectados	Freq. relativa de infectados
<1ANO	148	0,33%
1 A 9 ANOS	978	2,18%
10 A 19 ANOS	1688	3,76%
20 A 29 ANOS	6845	15,25%
30 A 39 ANOS	10879	24,24%
40 A 49 ANOS	9100	20,28%
50 A 59 ANOS	6745	15,03%
60 A 69 ANOS	4298	9,58%
70 A 79 ANOS	2453	5,47%
80 A 89 ANOS	1412	3,15%
90 OU MAIS	330	0,74%

Como vê-se na tabela acima, diferente do que ocorre com óbitos, a faixa de idades mais suscetíveis ao contagio da doença é entre 30 a 49 anos.

Com intuito de encontrar relação entre as variáveis de estudo entre os infectados, levantou-se a hipótese: Há alguma relação entre as variáveis de estudo entre os infectados? A combinação apresentou a seguinte visualização:

Tabela 9 – UTI x Sexo

UTI/Sexo	Feminino	Masculino
Não	13762	15492
Sim	971	1298

Tabela 10 – Internação x Sexo

Internação/Sexo	Feminino	Masculino
Não	10607	11962
Sim	4770	5519

Aparentemente, não existe relação entre o sexo do paciente e a necessidade de internação ou UTI.

Tabela 11 – Faixa etária x Internação

Faixa etária/Internação	Não	Sim
<1ANO	80	14
1 A 9 ANOS	567	115
10 A 19 ANOS	1242	62
20 A 29 ANOS	4279	389
30 A 39 ANOS	6193	1117
40 A 49 ANOS	4738	1690
50 A 59 ANOS	3025	2129
60 A 69 ANOS	1400	2124
70 A 79 ANOS	595	1507
80 A 89 ANOS	287	935
90 OU MAIS	52	207

Tabela 12 – Faixa etária x UTI

Faixa etária/UTI	Não	Sim
<1ANO	86	6
1 A 9 ANOS	649	12
10 A 19 ANOS	1275	12
20 A 29 ANOS	4518	65
30 A 39 ANOS	6916	203
40 A 49 ANOS	5838	370
50 A 59 ANOS	4455	461
60 A 69 ANOS	2790	485
70 A 79 ANOS	1574	365

80 A 89 ANOS	861	239
90 OU MAIS	181	50

A tabela parece mostrar que existe relação entre a necessidade de internação ou UTI e a faixa etária do paciente.

Tabela 13 – Etnia x Internação

Etnia/Internação	Não	Sim
AMARELA	1649	211
BRANCA	5711	3081
INDIGENA	11	3
PARDA	5770	4197
PRETA	785	661

Tabela 14 – Etnia x UTI

Etnia/Internação	Não	Sim
AMARELA	1800	42
BRANCA	7779	764
INDIGENA	14	0
PARDA	8695	832
PRETA	1229	129

Visualmente parece haver relação entre a necessidade de internação ou UTI e a etnia do paciente.

Tabela 15 – UTI x Comorbidade

UTI/Comorbidade	Não	Sim
Não	8576	4905
Sim	610	1234

Tabela 16 – Internação x Comorbidade

UTI/Comorbidade	Não	Sim
Não	6115	1673
Sim	3504	4984

As tabelas parecem mostrar evidencias de que há relação entre a necessidade de internação ou UTI e a presença de comorbidade no paciente.

Pela a analise descritivas dos dados, os dados aparentam apresentar relações significantes. Como na base de óbitos em que a presença de comorbidade e faixa etária de 60 a 89 aparentam ser o grupo de risco.

Já a base de infectados mostrou que comorbidade, diferentemente do caso anterior, não apresenta ralação entre o contagio ou não da doença. Mas já as etnias branca e parda e a faixa etária de 30 a 49 anos apresentam mais incidência da doença, o que indicada um possível grupo mais suscetível a contrair a covid-19.

A base para aplicar modelos de machine learning foi composta com dados da união das bases de casos confirmados e de pacientes que vieram a óbito, para entender as diferenças entre eles, fez-se a seguinte analise:

Tabela 17 – Evolução x Comorbidade

Evolução/Comorbidade	Sim	Não	Total
Óbito	17,57%	1,58%	19,15%
Recuperado	34,97%	45,87%	80,85%
Total	52,55%	47,45%	100,00%

A tabela mostra que pacientes com comorbidade apresentam maior chances de virem a óbito e pacientes sem comorbidades maiores chances de recuperação.

Tabela 18 – Evolução x Média de idade

Evolução	Média de idade
Óbito	70,09
Recuperado	47,77

A tabela mostra que pacientes com uma idade mais elevada apresentam maior chances de virem a óbito.

Tabela 19 – Evolução x Sexo

Evolução/Sexo	Masculino	Feminino	Total
Óbito	10,94%	8,21%	19,15%
Recuperado	44,79%	36,06%	80,85%
Total	55,74%	44,26%	100,00%

A tabela mostra que pacientes do sexo masculino apresentam maiores chances de virem a óbito e que pacientes do sexo feminino.

4. Criação de Modelos de Machine Learning

Inicialmente, iremos fazer inferência sobre a base de óbitos. Para estudar as variaveis sexo e comorbidade, foi utilizado o teste de comparação de proporções. Para ambas as variaveis, quando o teste foi aplicado, houve como retorno um p-valor muito próximo de 0. Disso, podemos concluir que há diferença estatisticamente significativa na proporção de óbitos entre homens e mulheres, e em pessoas que apresentam comorbidades e as que não apresentam.

Algoritmo utilizado:

```
## Sexo
prop.test(1200,2825,0.5,correct=F)
prop.test(1625,2825,0.5,correct=F, alternative = c("greater"))

## Comorbidade
prop.test(2354,2559,0.5,correct=F)
prop.test(2354,2559,0.5,correct=F, alternative = c("greater"))
```

Para testar a hipotese do grupo de risco ser entre 60 e 89 anos agrupou-se segundo o esquema abaixo:

Tabela 17 - Óbitos por faixa etária agrupados

Faixa Etária	Qtd de óbitos	Freq. relativa de óbitos
60 a 89 anos	2004	70,94%
Menos de 60 ou mais de 80 anos	821	29,06%

Dessa maneira, pode-se testar se este é o grupo com mais propensão de vir a obito. Para fazer este teste, considerou-se que a população com a faixa de idade entre 60 e 89 anos no estado de Minas Gerais é de 11%, segundo o senso de 2010 do IBGE. O teste de comparação de proporção apresentou o p-valor muito do proximo de 0, sendo assim, podemos concluir que há diferença estatisticamente significativa na proporção esperada de óbitos para a doença e a proporção de habitantes nesta faixa de idade, podendo considerar então como um fator de risco a idade entre 60 e 89 anos.

Algoritmo utilizado:

```
## Faixa etaria
```

```
prop.test(2004, 2825, 0.11, correct=F)
prop.test(2004, 2825, 0.11, correct=F, alternative = c("greater"))
```

Analisando a base de infectados, a qual contem mais variavies, podemos fazer mais inferencia sobre os dados. Inicialmente, foi utilizado o teste de comparação de proporções. Para as variaveis: necessidade de UTI, Necessidade de internação, Sexo e comorbidades, quando o teste foi aplicado, houve como retorno um p-valor muito próximo de 0. Disso, podemos concluir que há diferença estatisticamente significativa na proporção de contaminados que precisam de UTI ou internação, entre homens e mulheres, e em pessoas que apresentam comorbidades e as que não apresentam.

Algoritmo utilizado:

```
### Necessidade de UTI
prop.test(2269, 31531, 0.5, correct=F)
prop.test(2269, 31531, 0.5, correct=F, alternative = c("less"))
### Necessidade de Internação
prop.test(22575, 32868, 0.5, correct=F)
prop.test(22575, 32868, 0.5, correct=F, alternative = c("less"))
prop.test(22575, 32868, 0.5, correct=F, alternative = c("greater"))
### Sexo
prop.test(24051, 45719, 0.5, correct=F)
prop.test(24051, 45719, 0.5, correct=F, alternative = c("greater"))
### Comorbidade
prop.test(7263, 17094, 0.5, correct=F)
prop.test(7263, 17094, 0.5, correct=F, alternative = c("greater"))
```

Para testar a hipotese do grupo de maior risco ser entre 30 e 49 anos agrupou-se segundo o esquema abaixo:

Tabela 18 – Casos confirmados por faixa etária agrupados

Faixa Etária	Qtd de óbitos	Freq. relativa de óbitos
30 a 49 anos	2004	44,55%
Menos de 30 ou mais de 49 anos	821	55,45%

Dessa maneira, pode-se testar se este é o grupo com mais propensão de contrair a Covid-19. Para fazer este teste, considerou-se que a população com a faixa de idade entre 30 e 49 anos no estado de Minas Gerais é de 29%, segundo o senso de 2010 do IBGE. O teste de comparação de proporção apresentou o p-valor muito do proximo de 0, sendo assim, podemos concluir que há diferença estatisticamente significativa na proporção esperada de

contaminados para a doença e a proporção de habitantes nesta faixa de idade, podendo considerar então como um fator de risco de contagio a idade entre 30 e 49 anos.

Algoritmo utilizado:

```
### Faixa etaria
prop.test(19979, 44847, 0.29, correct=F)
prop.test(19979, 44847, 0.29, correct=F, alternative = c("greater"))
```

Além de testar cada variavel separadamente, foi estudada a interação entre elas por meio de teste de independencia.

A primeira interação estudada foi entre necessidade de internação em UTI e o sexo do paciente. A tabela de contigencia entre tais variaveis não atingiu o presuposto da frequencia minima esperada, logo, foi necessario utilizar o Teste Exato de Fisher, o qual retornou um p-valor muito proximo de 0. Disso, concluimos que necessidade de internação em UTI e o sexo do paciente, são variaveis dependentes.

Também foi analisada a interação entre necessidade de internação e sexo, como os presupostos para o teste de Qui-Quadrado foram atingidos, não foi necessario aplicar o Teste Exato de Fisher. O p-valor retornado foi de 0,2282, disso conclui-se que internação e sexo do paciente não são dependentes.

Com intuito de entender a relação entre necessidade de internação em UTI e a presença ou não de comorbidade, foi aplicado o Teste Exato de Fisher nessas variaveis, já que os presupostos do teste de Qui-Quadrado não foram atendidos. O teste retorno o p-valor de 0,2828, o que indica que as variaveis são independentes entre si.

A última interação analisada foi entre necessidade de internação e presença ou não de comorbidade, pelo teste de Qui-Quadrado, o qual retornou um p-valor muito proximo de zero e teve os presupostos atingidos, podemos concluir que as variaveis são dependentes entre si.

Após os padrões de comportamento serem entendidos na analise estatistica, fez-se o treinamento do modelo, o qual foi utilizado a árvore de decisão. Para isso, a base teve que ser codificada da seguinte forma:

Tabela 19 – Codificação dos dados para os modelos de machine learning

Variável	Codificação
Sexo	Feminino – 0 Masculino - 1
Comorbidade	Não – 0 Sim - 1
Faixa etária	Risco – 1

Sem risco – 0

Lembrando que pela analise estatística, a faixa etaria de risco para o grupo de óbitos é de 60 a 89 anos e para recuperados de 30 a 49 anos. A base foi dividida em duas partes: treinamento e validação com 80 e 20% dos dados, respectivamente. Nesse ponto, deve-se resaltar tbm que a base de casos confirmados foi filtrada para exibir apenas pacientes que se recuperaram.

Árvores de decisão são métodos de aprendizado de máquinas supervisionado não-paramétricos, muito utilizados em tarefas de classificação e regressão. Em geral no mundo da programação, árvores são estruturas de dados formadas por um conjunto de elementos que armazenam informações, os nós.

Além do mais, toda árvore possui um nó chamado raiz, que possui o maior nível hierárquico (o ponto de partida) e ligações para outros elementos, denominados filhos. Esses filhos podem ou não possuir seus próprios filhos que também podem possuir os seus. O nó que não possui filho é conhecido como nó folha ou terminal.

Vale ressaltar que uma árvore de decisão é que uma árvore que armazena regras em seus nós, e os nós folhas representam a decisão a ser tomada, no caso desse trabalho, se o paciente em questão irá se recuperar ou vir a falecer.

O objetivo de uma árvore de decisão é tomar uma decisão através do caminho a partir do nó raiz até o nó folha.

Uma parte importante da modelagem de machine learning é a validação do modelo, o método utilizado nesse trabalho foi a da validação cruzada. Tal método consistem em uma técnica para avaliar a capacidade de generalização de um modelo, ela é muito empregada em problemas onde o objetivo da modelagem é a predição. Em suma busca-se estimar o quão preciso é este modelo na prática, ou seja, o seu desempenho para classificar novos conjunto de dados.

Para empregar-se este tipo de validação, é realizado o particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, os quais são divididos em dados de treinamento e dados de validação ou teste, que são respectivamente utilizados para estimação dos parâmetros do modelo e para validação do modelo.

Os modelos de machine learning foram todos feitos no Knime, conforme imagem abaixo:

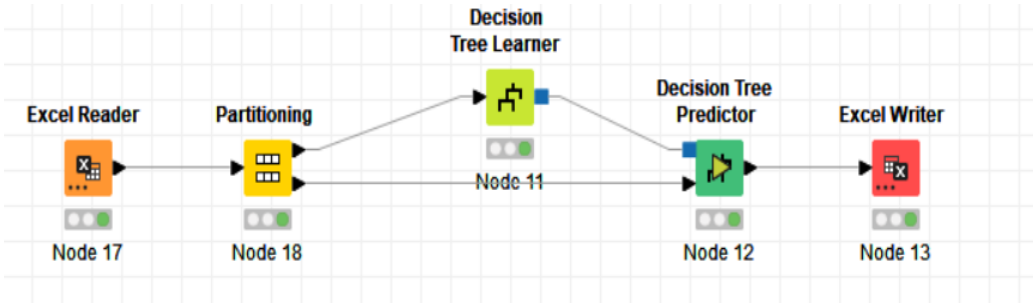


Imagem 7: Modelagem no Knime

Foram feitos 3 modelos de árvore de decisão, os quais seguiram o modelo da imagem a cima. Primeiramente foi feito o input de dados (Excel Reader), depois o particionamento dos dados para fins de validação (Partitioning), após isso usou-se uma parte para treinar o modelo (Decision Tree Learner) e a outra para validação (Decision Tree Predictor) e finalmente os dados foram extraídos (Excel Writer).

O Knime oferece diversos modos de configuração para execução do algoritmo da árvore de decisão, com base nessas configurações foram escolhidos os parâmetros dos três modelos ajustados neste trabalho.

Primeiramente, devemos escolher o método de decisão que procura executar a melhor divisão do recurso de entrada em cada iteração, para isso temos duas opções o Gini index e Gain Ratio. O Gini index se baseia na ideia de minimizar os erros de classificação e o Gain ratio que se baseia na ideia de escolher o recurso mais informativo.

Outro parâmetro que devemos configurar é o “Average split point”, o splint (ou divisão), um atributo numérico que é sempre uma divisão binária que divide o conjunto de dados em dois subconjuntos. Ele serve para decidir onde fazer a divisão. Marcando como sim ela é definida como a média do valor mais alto da partição inferior e no valor mais alto da partição inferior como o ponto de divisão, se não selecionada a divisão será no valor mais alto da partição inferior.

Para ajustes do modelo também é configurado o “Pruning method” que é uma técnica de poda da árvore, no Knime encontra-se o MDL (Comprimento mínimo de descrição), pode-se decidir aplica-lo ou não. Outra opção também eficaz é aplicar o “Reduce error pruning” que se inicia nas folhas da árvore e substitui cada nó por sua classe mais popular, mas apenas se a precisão geral da previsão não diminuir.

Outra configuração foi para limitar o tamanho da árvore durante o treinamento, definimos um número maior para o número mínimo de registros por nó (Min number records per node), este é um critério de parada para o algoritmo. Assim que o número de registros em um nó for menor que esse número predefinido, o algoritmo evita mais divisões.

Temos duas configurações restantes: o número de registros a serem armazenados para visualização (The number of records to store for view) e o número de segmentos (The number of threads). Porém essas configurações estão relacionadas à velocidade de execução do nó e não afetam o modelo. Eles, respectivamente, definem quantos registros são armazenados para criar a visualização e para realçá-los, o número padrão (10) tenta limitar os recursos usados pelo algoritmo e o número de segmentos.

Abaixo vemos o primeiro modelo, o qual atingiu uma precisão de previsão de 97,37%(Validação cruzada).

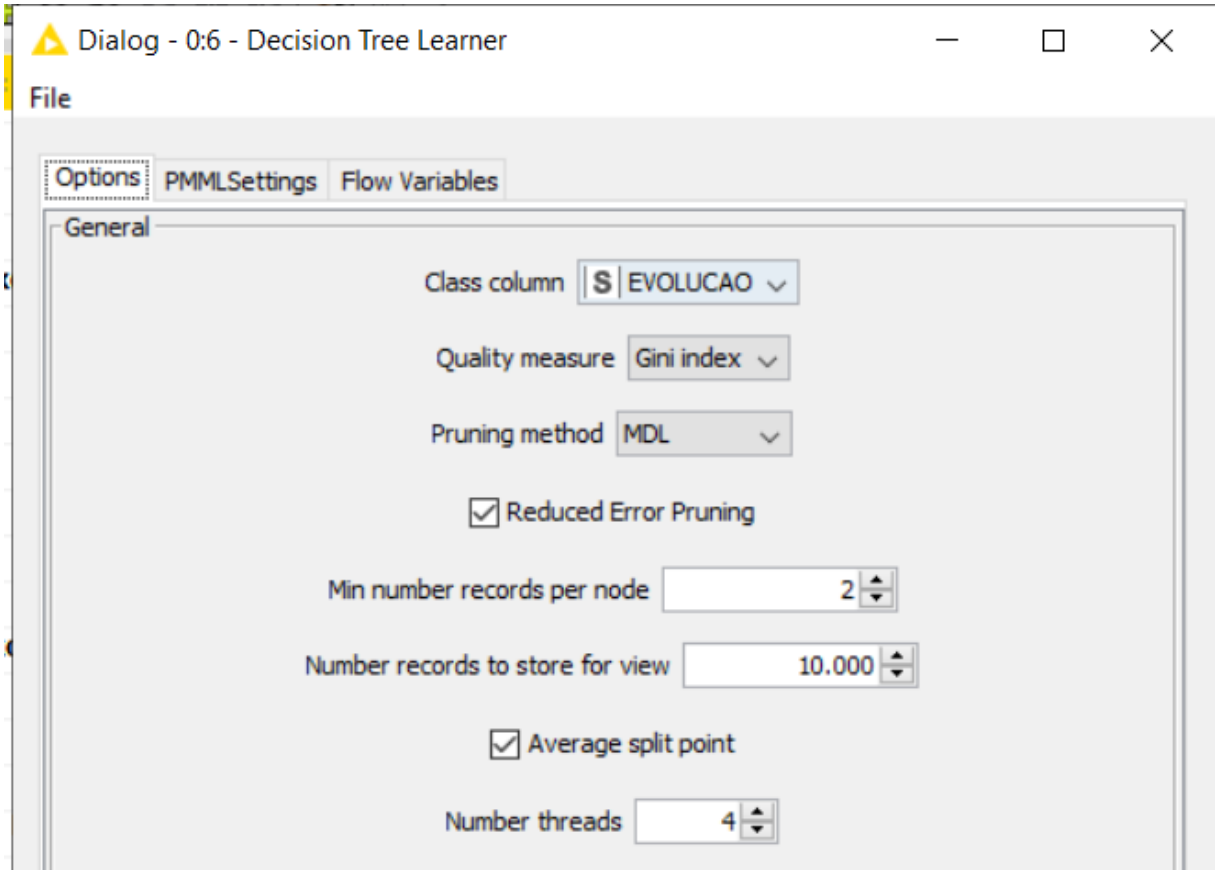


Imagem 8: Primeira modelagem da árvore de decisão no Knime

Para o segundo modelo, retira-se o “Pruning method”, ou a técnica de poda da árvore. Vemos que a precisão da previsão cai para 96,32% (Validação cruzada).

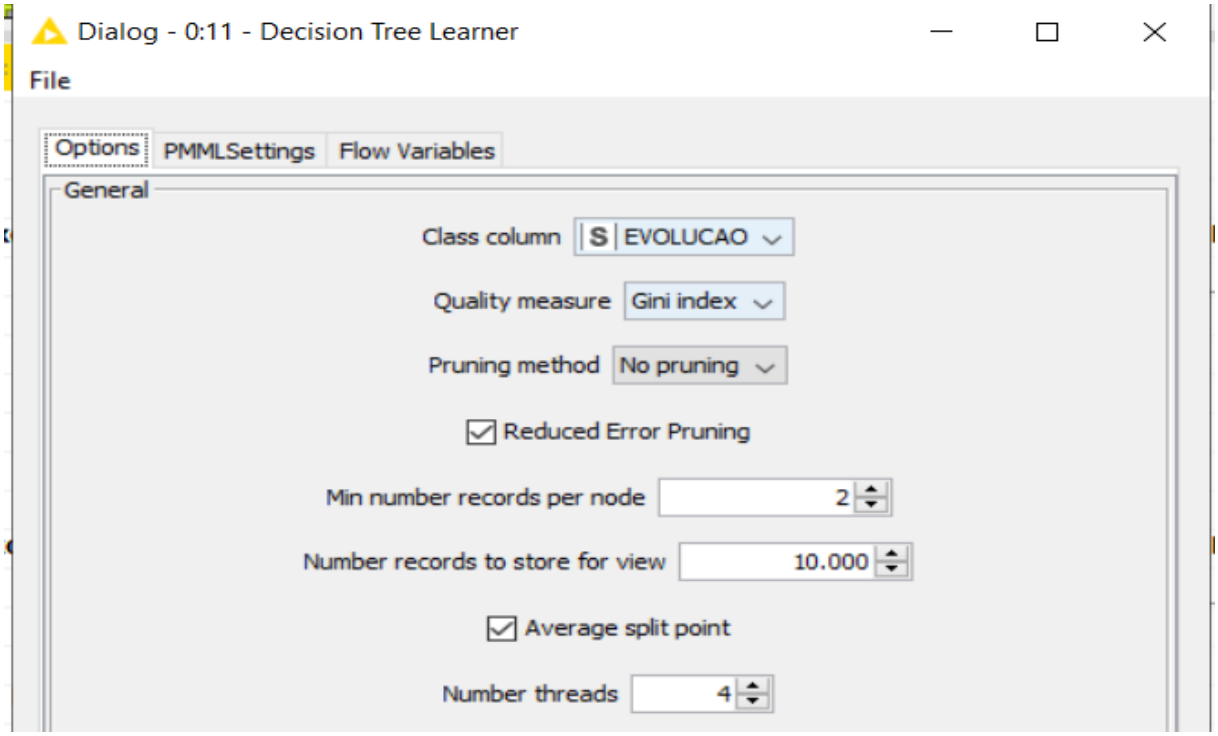


Imagem 9: Segunda modelagem da árvore de decisão no Knime

Já na terceira modelagem além de eliminar a técnica de poda, também se elimina o “Reduce Error Pruning” que substitui cada nó por sua classe mais popular, a precisão dele sobe, comparado ao segundo modelo, para 96,93% (Validação cruzada).

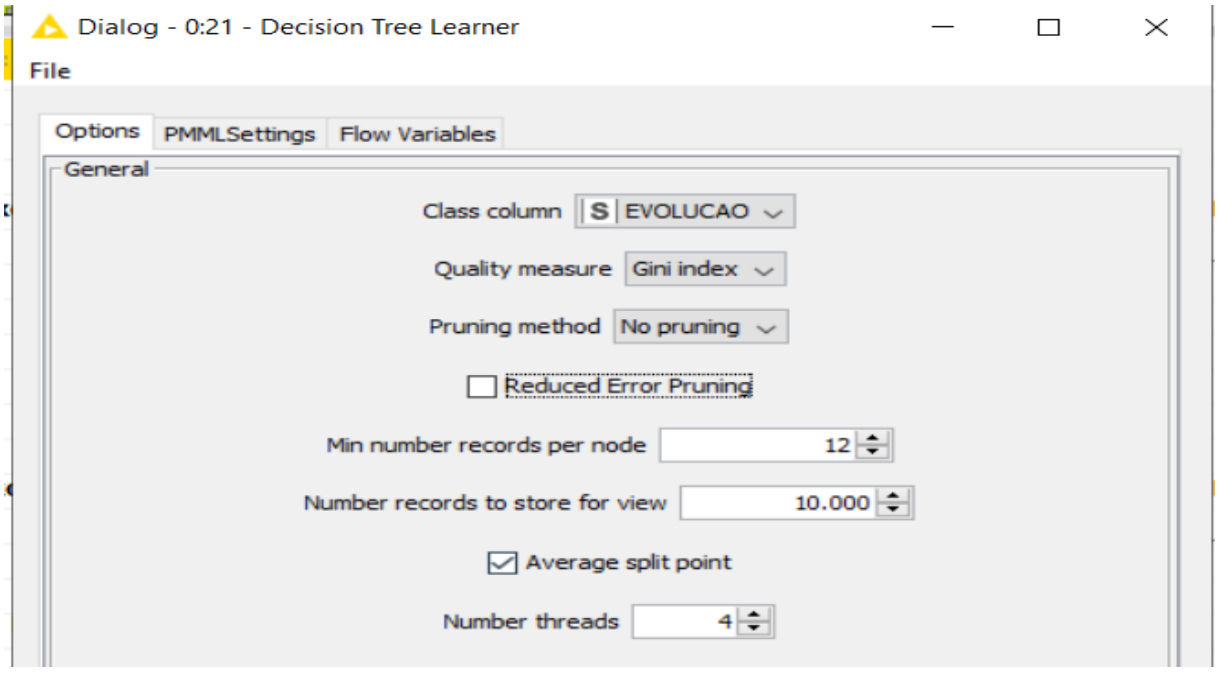


Imagem 10: Terceira modelagem da árvore de decisão no Knime

5. Interpretação dos Resultados

A partir dos dados analisados, podemos concluir que o grupo de risco para óbitos em decorrência da covid-19: são homens, pessoas com comorbidades e a faixa etária entre 60 e 89 anos.

Assim como na base de óbitos, podemos concluir que há maior risco de contagio em homens, pessoas sem comorbidades e entre 30 e 49 anos. Além disso, menos da metade de contaminados não precisa de UTI, mas mais da metade precisa de internação.

Foi também estudado a interação das variáveis, o que mostrou relação significativamente estatística entre sexo e necessidade de internação e entre internação e ausência ou presença de comorbidade do paciente.

Além do mais, mostrou-se que a modelagem utilizada nos modelos de machine learning foram muito precisas em classificar pacientes que vieram a óbito e que se recuperaram com capacidade máxima atingida superior a 97% de classificações corretas.

6. Apresentação dos Resultados

A parte de visualização foi inteiramente feita no Power BI, como mostram as imagens abaixo.

Ao acessar o recurso visual, o usuário é levado a tela de seleção do que deseja se informar:

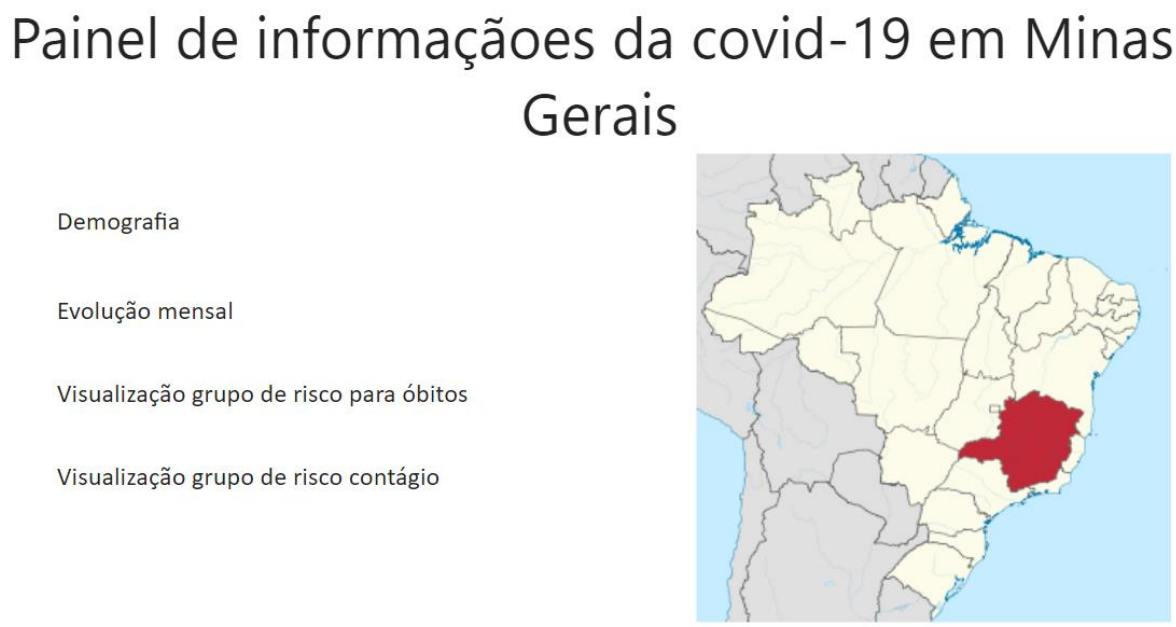


Imagem 11: Tela de seleção do Power BI

Clicando em “Demografia”, pode-se observar no mapa onde há a maior incidência de contaminados e casos de óbitos:

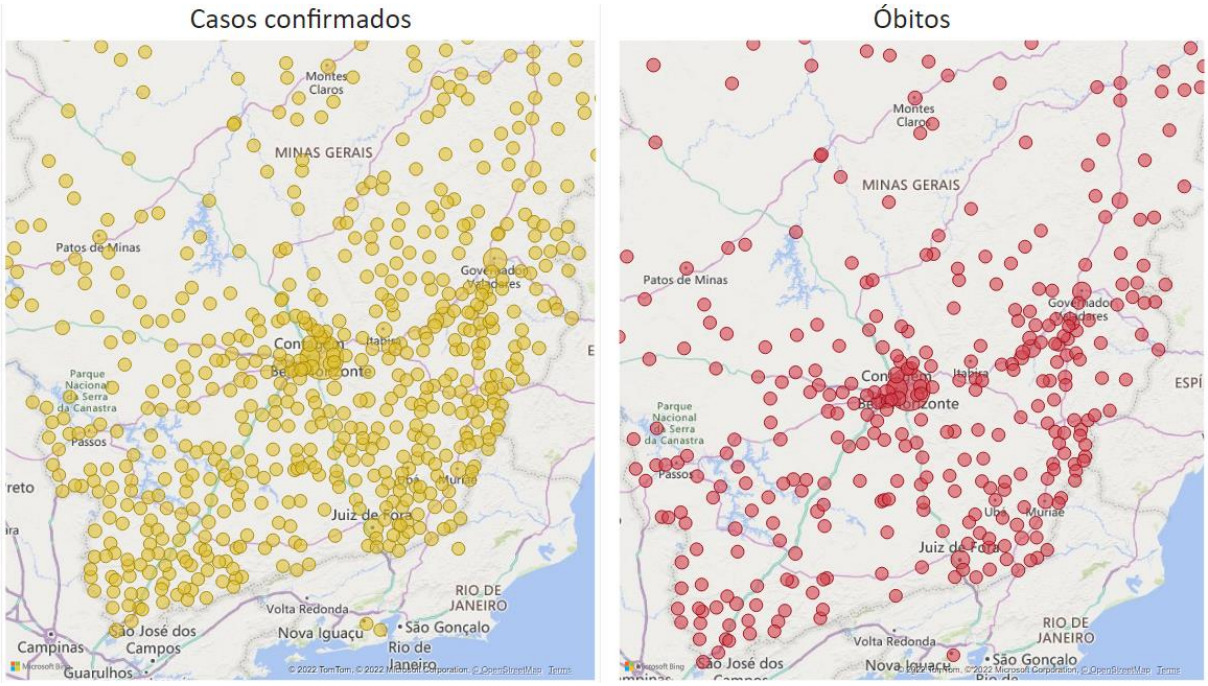


Imagem 12: Demografia da Covid-19

Já em “Evolução mensal”, podemos ver a quantidade histórica do número de contaminados e de óbitos mensalmente.

Imagem 13: Evolução mensal

Agora em “visualização grupo de risco de óbitos”, podemos ver em destaque as características de risco comprovadas estatisticamente.

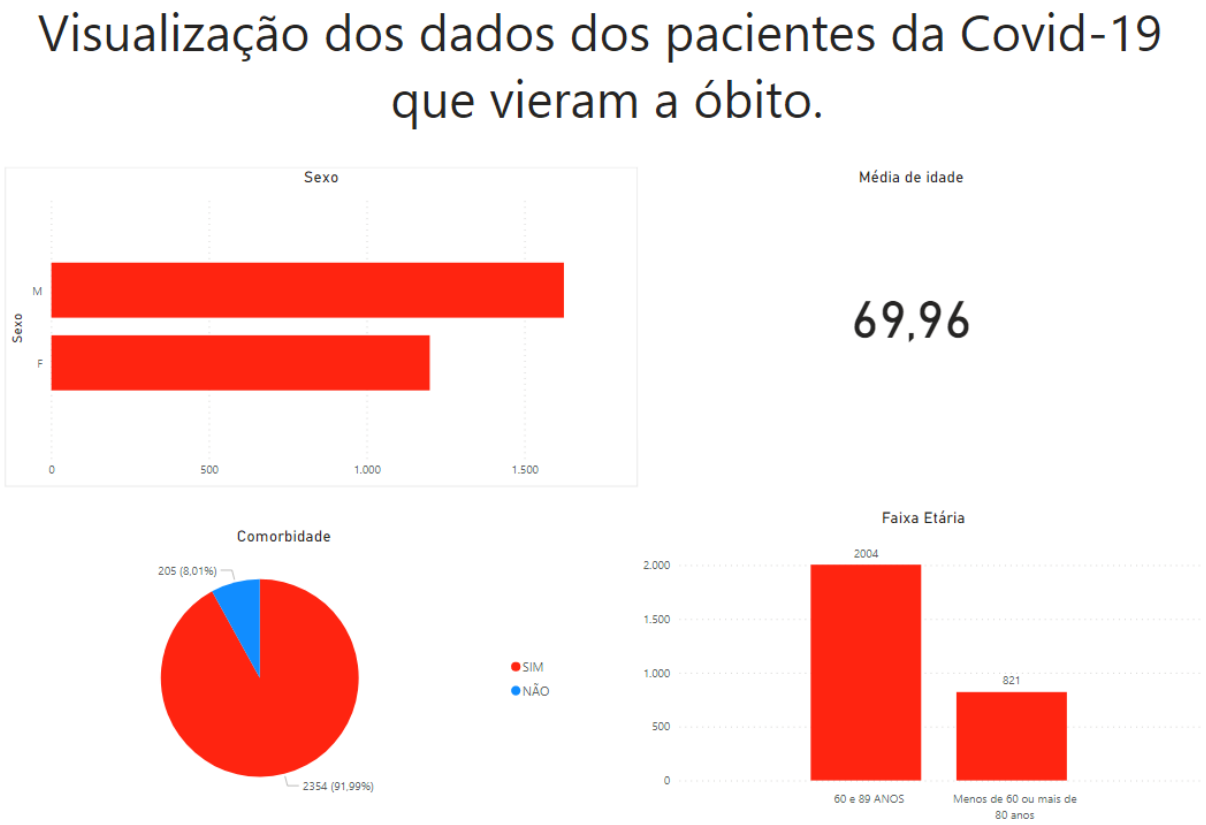


Imagem 14: Descritivo óbitos

Selecionando “visualização grupo de risco contágio”, podemos ver em destaque as características de risco comprovadas estatisticamente.

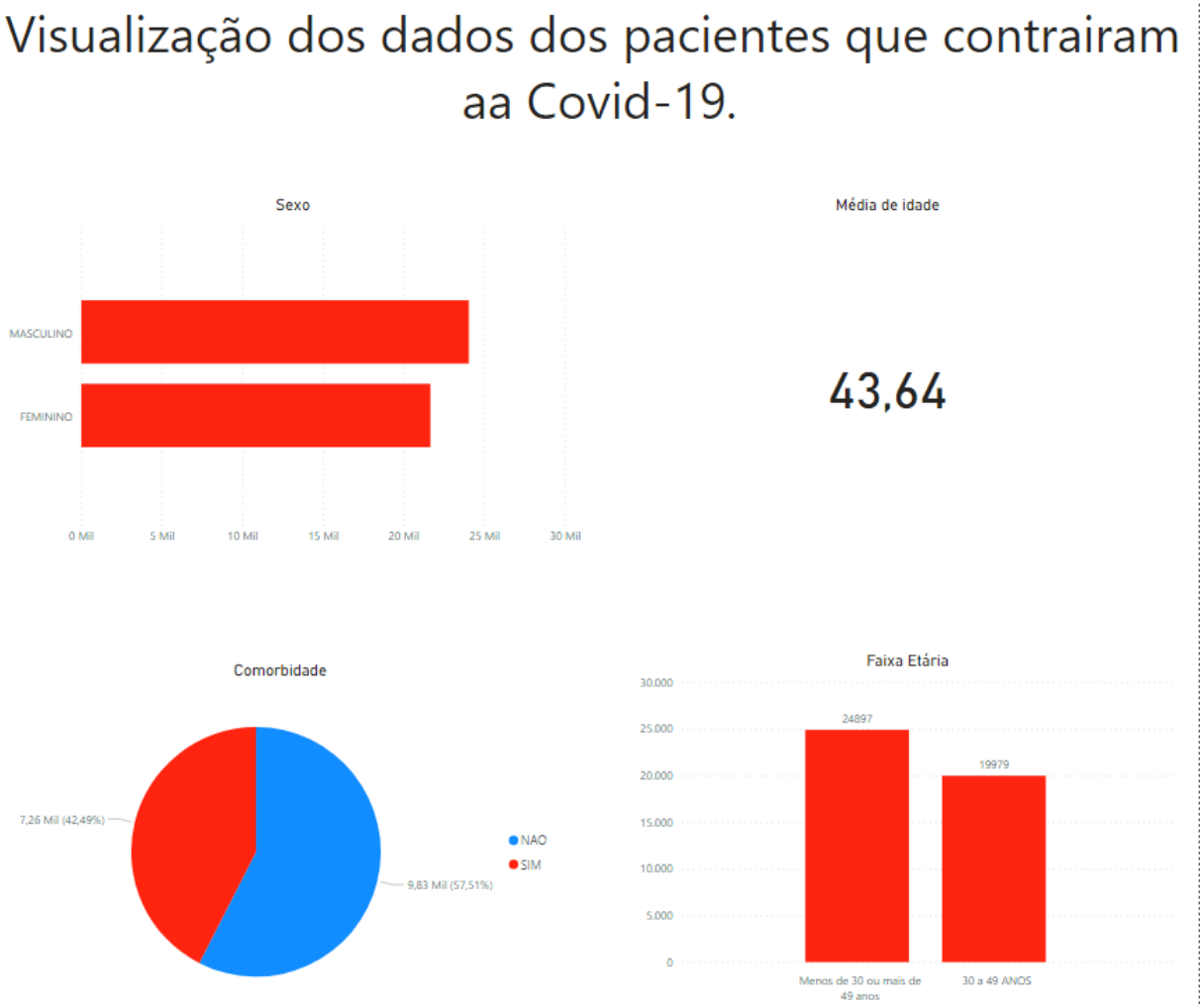


Imagem 15: Descritivo contágios

7. Links

Para acessar ao vídeo explicativo, por favor acesse este link do youtube: https://youtu.be/iOyOk7B4jGE

O repositório com os dados, scripts e PBI se encontra nesse link: https://drive.google.com/drive/folders/1o4n_LHp2-0LtZaLqW_1zw3Ow5CvbRBXl?usp=sharing

REFERÊNCIAS

GRUBER, Arthur. **Covid-19: o que se sabe sobre a origem da doença.** Jornal da USP. https://jornal.usp.br/artigos/covid2-o-que-se-sabe-sobre-a-origem-da-doenca/. Acesso em 12/07/2021.

IBGE – Instituto Brasileiro de Geografia e Estatística. **População 2010.**

Portal de Dados Abertos - Governo de Minas Gerais. https://dados.mg.gov.br/. Acesso em 02/07/2021.

Portal de Dados Abertos - Governo de Minas Gerais. https://dados.mg.gov.br/. Acesso em 02/07/2021.

Portal de Dados Abertos - Governo de Minas Gerais. **Casos confirmados por COVID-19.** https://dados.mg.gov.br/dataset/casos-confirmados-covid-19. Acesso em 02/07/2021.

Portal de Dados Abertos - Governo de Minas Gerais. **Óbitos confirmados por COVID-19.** https://dados.mg.gov.br/dataset/obitos-confirmados-covid-19. Acesso em 02/07/2021.

Rede Suas. **Lista de municípios Brasileiros e Informações adicionais.** http://blog.mds.gov.br/redesuas/lista-de-municipios-brasileiros/. Acesso em 06/07/2021.

APÊNDICE

Programação/Scripts

Script R utilizado para analises estatísticas.

```
############## Base de óbitos ########################
## Sexo
prop.test(1200,2825,0.5,correct=F)
prop.test(1625,2825,0.5,correct=F, alternative = c("greater"))
## Comorbidade
prop.test(2354,2559,0.5,correct=F)
prop.test(2354,2559,0.5,correct=F, alternative = c("greater"))
##### Faixa etaria
# var = read.csv("C:/Users/Gabi/Google Drive/TCC PUC 2021/datasetqui.csv")
prop.test(2004, 2825, 0.11, correct=F)
prop.test(2004, 2825, 0.11, correct=F, alternative = c("greater"))
############## Base de contaminados ####################
### Necessidade de UTI
prop.test(2269, 31531, 0.5, correct=F)
prop.test(2269, 31531, 0.5, correct=F, alternative = c("less"))
### Necessidade de Internação
prop.test(22575, 32868, 0.5, correct=F)
prop.test(22575, 32868, 0.5, correct=F, alternative = c("less"))
prop.test(22575, 32868, 0.5, correct=F, alternative = c("greater"))
### Sexo
prop.test(24051, 45719, 0.5, correct=F)
prop.test(24051, 45719, 0.5, correct=F, alternative = c("greater"))
### Comorbidade
prop.test(7263, 17094, 0.5, correct=F)
prop.test(7263, 17094, 0.5, correct=F, alternative = c("greater"))
### Faixa etaria
prop.test(19979, 44847, 0.29, correct=F)
prop.test(19979, 44847, 0.29, correct=F, alternative = c("greater"))
```

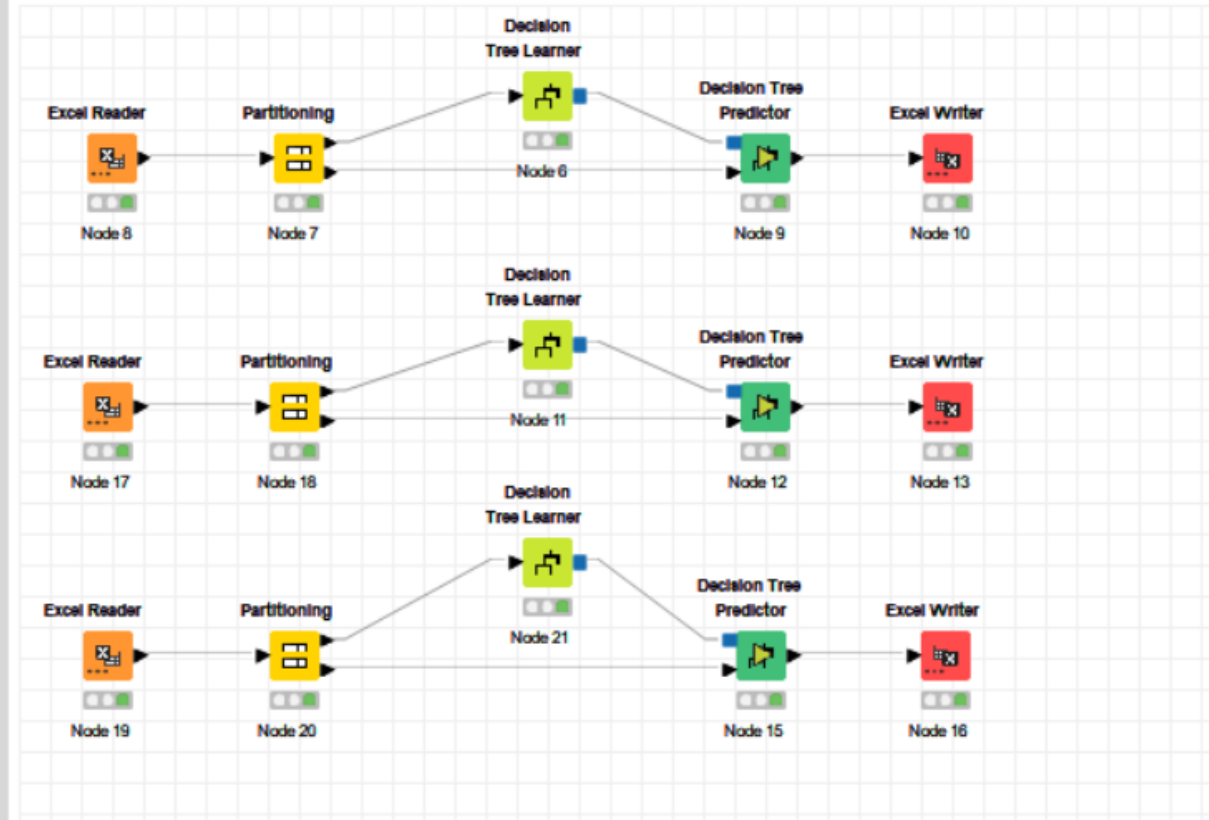
```
#################### Qui-quadrado de independência ##########################
# Passo 1: Carregar os pacotes que serão usados
if(!require(dplyr)) install.packages("dplyr")
library(dplyr)
if(!require(rstatix)) install.packages("rstatix")
library(rstatix)
if(!require(psych)) install.packages("psych")
library(psych)
if(!require(corrplot)) install.packages("corrplot")
library(corrplot)
############################ UTI x Sexo ############################
UTI_Sexo = read.csv("C:/Users/Gabi/Google Drive/TCC PUC
 2021/UTI_SEXO.csv", stringsAsFactors = T)
#View(UTI_Sexo)
names(UTI_Sexo)
summary(UTI_Sexo$SEXO)
summary(UTI_Sexo$ï..UTI)
UTI_Sexo$SEXO <- factor(UTI_Sexo$SEXO, levels = c("FEMININO",  "MASCULINO"))
tabela <- table(UTI_Sexo$ï..UTI, UTI_Sexo$SEXO)
tabela
## Realização do modelo
quiqua2 <- chisq.test(tabela)
quiqua2
# Análise das frequências esperadas
# Pressuposto: frequências esperadas > 5
quiqua2$expected
N <- count(UTI_Sexo[1])
100*(360.7656/N)
#Não atingiu, usar fisher
fisher.test(tabela)
########################## Internação x Sexo ##########################
```

```
INTERNACAO_SEXO=read.csv("C:/Users/Gabi/Google Drive/TCC PUC
2021/INTERNACAO_SEXO.csv", stringsAsFactors = T)
#View(INTERNACAO_SEXO)
names(INTERNACAO_SEXO)
summary(INTERNACAO_SEXO$SEXO)
summary(INTERNACAO_SEXO$ï..INTERNACAO)
INTERNACAO_SEXO$SEXO <- factor(INTERNACAO_SEXO$SEXO, levels = c("FEMININO",
"MASCULINO"))
tabela <- table(INTERNACAO_SEXO$ï..INTERNACAO, INTERNACAO_SEXO$SEXO)
tabela
## Realização do modelo
quiqua2 <- chisq.test(tabela)
quiqua2
# Análise das frequências esperadas
# Pressuposto: frequências esperadas > 5
quiqua2$expected
N <- count(INTERNACAO_SEXO[1])
100*(3953.546/N)
############################## UTI x Comorbidade ##############################
UTI_COMORBIDADE = read.csv("C:/Users/Gabi/Google Drive/TCC PUC
2021/UTI_COMORBIDADE.csv", stringsAsFactors = T)
#View(UTI_COMORBIDADE)
names(UTI_COMORBIDADE)
summary(UTI_COMORBIDADE$COMORBIDADE)
summary(UTI_COMORBIDADE$ï..UTI)
UTI_COMORBIDADE$COMORBIDADE <- factor(UTI_COMORBIDADE$COMORBIDADE,
levels = c("SIM","NAO"))
tabela <- table(UTI_COMORBIDADE$COMORBIDADE, UTI_COMORBIDADE$ï..UTI)
tabela
## Realização do modelo
quiqua2 <- chisq.test(tabela)
quiqua2
```

```
# Análise das frequências esperadas
# Pressuposto: frequências esperadas > 5
quiqua2$expected
N <- count(UTI_COMORBIDADE[1])
100*(738.6829/N)
#Não atingiu, usar fisher
fisher.test(tabela)

##################### Internação x Comorbidade ############################
INTERNACAO_COMORBIDADE = read.csv("C:/Users/Gabi/Google Drive/TCC PUC
2021/INTERNACAO_COMORBIDADE.csv", stringsAsFactors = T)
#View(INTERNACAO_COMORBIDADE)
names(INTERNACAO_COMORBIDADE)
summary(INTERNACAO_COMORBIDADE$COMORBIDADE)
summary(INTERNACAO_COMORBIDADE$ï..INTERNACAO)



INTERNACAO_COMORBIDADE$COMORBIDADE <-
factor(INTERNACAO_COMORBIDADE$COMORBIDADE, levels = c("SIM","NAO"))
tabela <- table(INTERNACAO_COMORBIDADE$COMORBIDADE,
INTERNACAO_COMORBIDADE$ï..INTERNACAO)
tabela
## Realização do modelo
quiqua2 <- chisq.test(tabela)
quiqua2
# Análise das frequências esperadas
# Pressuposto: frequências esperadas > 5
quiqua2$expected
N <- count(UTI_COMORBIDADE[1])
100*(3185.348/N)
```

Gráficos

Workflow do Knime no qual foram feitos os modelos de machine learning.



Tabelas

Para fins de cálculo de porcentagens de faixa etária, foi gerada a tabela abaixo.

Faixa etária	Homens	Mulheres	Total
Mais de 100 anos	739	1904	2.643
95 a 99 anos	3332	7576	10.908
90 a 94 anos	12469	24269	36.738
85 a 89 anos	34862	56569	91.431
80 a 84 anos	76292	112030	188.322
75 a 79 anos	129276	168843	298.119
70 a 74 anos	191852	233376	425.228
65 a 69 anos	251626	290172	541.798
60 a 64 anos	339165	376213	715.378
55 a 59 anos	441415	479713	921.128
50 a 54 anos	548830	584829	1.133.659
45 a 49 anos	628195	666388	1.294.583
40 a 44 anos	671738	702039	1.373.777
35 a 39 anos	694342	722116	1.416.458
30 a 34 anos	790229	805450	1.595.679
25 a 29 anos	851586	853105	1.704.691

20 a 24 anos	874104	859390	1.733.494
15 a 19 anos	868022	851253	1.719.275
10 a 14 anos	858109	830051	1.688.160
5 a 9 anos	726034	702961	1.428.995
0 a 4 anos	649660	627206	1.276.866