Agrupamentos de pacientes pertencentes ao espectro autista utilizando-se técnicas de “Clustering”

Aluno: Israel José Monteiro Carvalho, Bacharel em Administração de Empresas, Especialista e Marketing. Rua: Alberto Wiest, 259 – Costa e Silva; 89220-570 Joinville, Santa Catarina, Brasil. Email: israeljmcarvalho@gmail.com; Orientador: Anna Carolina Martins, Professora Orientadora e Mestre em Economia Aplicada. Rua Alexandre Herculano, 120 - T6 - Vila Monteiro; 13418-445 Piracicaba, São Paulo

Resumo

Atualmente, pacientes diagnosticados dentro do espectro autista são acompanhados primariamente por médicos especialistas em neurologia ou psiquiatria. Além deste acompanhamento, em grande parte dos casos, é sugerido que haja acompanhamento terapêutico em áreas como psicoterapia, fonoaudiologia, terapia ocupacional, terapias de educação física, pedagogia dentre outras. As instruções que estes profissionais precisam para elaborarem seus planos de terapia geralmente são fornecidas pelos médicos. Todas as informações que os médicos têm sobre seus respectivos pacientes são oriundas de relatos dos pais (ou responsáveis) ou de suas próprias conclusões ao analisarem pessoalmente os pacientes durante as consultas. Percebe-se que há uma grande dificuldade dos médicos analisarem seus pacientes sob a ótica de outros pacientes, ou grupo de pacientes com características afins. Parte desta dificuldade dá-se por ainda não ser uma prática o registro das diversas coletadas variáveis dos pacientes em banco de dados, mas sim em prontuários isolados. Desta forma não há a possibilidade de analisar e interpretar informações de um paciente sob a ótica de outro, ou mesmo analisar o comportamento das variáveis de um grupo de pacientes com caraterísticas semelhante. Com base neste contexto, esta obra acadêmica foi fundamentada tendo como intenção a apresentação de como a técnica de “Clustering”, pertencente ao grupo “Unsupervised Machine Learning”, pode ser útil na classificação e agrupamentos de pacientes com base em suas similaridades em grupos distintos, o que expandiria exponencialmente o poder analítico dos profissionais da área médica e consequentemente aos demais profissionais envolvidos no tratamento de pacientes dentro do espectro autista.

Palavras-chave: autisto; tea; análise de comportamento de pacientes.

Agrupamentos de pacientes pertencentes ao espectro autista utilizando-se técnicas de “Clustering”

Palavras-chave: (visualização; agrupamentos; compreensão; *insights;* similaridades).

Introdução

Atualmente, quando se fala sobre pacientes pertencentes ao espectro do autismo, as especialidades médicas primárias que acompanham estes pacientes são Psiquiatria e Neurologia. Indiferente à especialidade, é consenso a falta de informações de qualidade disponíveis aos médicos para tomarem decisões a respeito da melhor escolha do tratamento (principalmente medicamentoso) de seus respectivos pacientes. Grande parte desta dificuldade dá-se pelo fato de que os progenitores (ou responsáveis pelos pacientes) são as principais fontes de informação. Isto é, muitas vezes por possuírem íntima ligação emocional com os pacientes, estão naturalmente sujeitos a extrema variabilidade do comportamento dos pacientes e tornam se impossibilitados de avaliarem o paciente de forma plenamente imparcial. Nesse sentido demonstram-se incapazes de elaborar um relato sem viés (hora oriundo de fases difíceis que os pacientes estejam passando, hora gerando excentricidade desproporcional com possíveis evoluções do mesmo paciente).

Exemplificando o exposto anteriormente, se o médico perguntar ao responsável do paciente como está a rotina de sono do paciente e se coincidentemente nas últimas 5 noites o responsável não conseguiu dormir, pois o paciente passou por noites com muita agitação psicomotora, muito provavelmente o responsável reportará ao médico que a rotina de sono do paciente está péssima. Isso acontece mesmo que nos outros 85 dias anteriores aos 5 e posteriores a consulta médica anterior, o paciente tenha dormido perfeitamente.

Segundo Belfiore e Fávero (2017), a utilização de algoritimos de “Clustering” (também chamados de Análise de Conglomerados) são muito úteis na intenção de verificar se há existência de comportamentos semelhantes entre observações em relação a determinadas variáveis. De posse deste conceito propõe-se neste trabalho a realização de simulações de algoritimos de “Clustering”, “Unsupervised Machine Learning”, para avaliar as semelhanças entre os pacientes com o mesmo distúrbio afim de obter-se agrupamentos de pacientes com características semelhantes, proporcionando para os profissionais da área médica novas formas de analisar seus pacientes.

Diferentemente das técnicas supervisionadas de “Machine Learning”, como por exemplo Regressões Linear, Logística Binária, Logística Multinomial, Poisson dentre outras, a clusterização não apresenta caráter preditivo para observações não presentes na amostra, desta forma esta técnica é considerada uma técnica exploratória (ou de interdependência). Caso surjam novas observações, ou variáveis, ou mesmo caso altere-se o valor das variáveis existentes, faz-se necessário aplicar novamente a técnica pois possivelmente o cálculo da distância entre as observações será alterado e tal alteração poderá resultar em uma

distribuição diferente das observações quando comparada a distribuição inicialmente sugerida pelo algoritmo.

O agrupamento de observações pode ser uma excelente estratégia quando deseja-se classificar as observações de uma população em grupos homogêneos onde serão contidas observações semelhantes e onde deseja-se que cada grupo seja o mais distinto possível um do outro. Desta forma é proporcionado ao processo de tomada de decisão ricos mecanismos onde são explicitados não apenas o comportamento, mas também a relação de interdependência entre as observações do “dataset” analisado.

Este tipo de técnica não supervisionada pode realizar de forma eficaz, e principalmente sem ponderação arbitrária, o agrupamento de observações levando em consideração variáveis métrica ou até mesmo binárias.

A principal forma de agrupamento é o cálculo da distância (também chamadas de dissimilaridade) ou de semelhança (também chamadas de similaridade) entre os valores e, portanto, não se deve utilizar variáveis qualitativas na aplicação de tal técnica.
Medidas de distância são geralmente utilizadas quando as variáveis do “dataset” forem essencialmente métricas enquanto as medidas de semelhança são geralmente utilizadas quando as variáveis forem binárias.

O cálculo das distâncias pode ser realizado através da utilização de diversas medidas distintas, sendo a mais comum a Distância Euclidiana, que por sua vez é uma generalização do Teorema de Pitágoras. Outras medidas de dissimilaridade que podem ser utilizadas são: Distância Euclidiana Quadrática, Distância de Minkowski, Distância de Manhattan (também conhecida por distância absoluta ou bloco), Distância de Chebychev (também conhecida por distância infinita ou máxima) e Distância de Canberra. Cada método pode gerar valores diferentes de distância cabendo ao pesquisador a escolha de qual método está mais alinhado com os objetivos estipulados na pesquisa. De acordo com Bussab et ai. (1990), a utilização de diferentes critérios de medidas de distância e de esquemas de aglomeração podem resultar em diferentes formações de grupos, cabendo ao pesquisador a escolha do melhor mix de ferramentas. Salienta-se que as escolhas do pesquisador interferem diretamente na homogeneidade das observações dentro de cada cluster.

Recomenda-se que o pesquisador documente com clareza e transparência as premissas escolhidas na elaboração de seu estudo (esquemas de aglomeração não hierárquicos ou hierárquicos e métodos de encadeamento) de forma a conseguir associar o resultado obtido com a respectivas premissas utilizadas possibilitando desta forma que simulações possam ser feitas e facilmente compreendidos seus resultados

A técnica de clusterização foi desenvolvida para receber “input” de dados de variáveis quantitativos e tem como “output” variáveis qualitativas (os grupos efetivamente criados). O

"output" obtido após execução desta técnica pode servir de "input" em outras técnicas multivariadas, tanto exploratórias, quanto confirmatórias, onde sejam requeridas variáveis qualitativas como "input".

Aconselha-se que a utilização das possíveis variáveis qualitativas constantes no "dataset" seja empregada no processo de interpretação e explicação dos grupos formados (output do processo de clusterização). Obviamente tais variáveis podem ser objeto de outras técnicas de "Machine Learning" que possuam "inputs" de dados oriundos de variáveis categóricas como por exemplo Análise de Correspondência.

Evidencia-se também a redução da dimensionalidade da base de dados e a possibilidade de validar constructos previamente estabelecidos como sendo dois outros importantes objetivos buscados no emprego desta técnica.

Quando for observado que as variáveis selecionadas para a execução do algoritmo de clusterização apresentarem diferentes métricas ou ordens de grandeza, é fortemente recomendado que seja realizada a padronização dos valores antes de calcular as distâncias para evitar possíveis viés nos resultados. Um dos processos mais utilizados na realização de tal processo é o "Z-Scores" pois a forma da distribuição da variável original não é alterada. Este método resultará em novas variáveis padronizadas que possuirão média igual a 0 e variância igual a 1.

A aplicação de tal método será feito a partir da utilização de um "Dataset*"* com observações e variáveis quantitativas que não correspondem a dados de pacientes reais, mas que são em seu caráter totalmente aderentes a potenciais casos médicos, refletindo com muita propriedade a realidade encontrada em muitos pacientes. Este "Dataset*"* será elaborado em conjunto com um profissional médico psiquiatra atuante na cidade de Joinville em conjunto com profissionais multidisciplinares que realizam terapias neste tipo de pacientes, sendo estes profissionais: Terapeutas Ocupacionais, Psicoterapeutas, Pedagogos, Fisioterapeutas, Educadores Físico.

Através da execução destes algoritmos pleiteia-se a possibilidade de identificar "clusters" de pacientes com a menor variabilidade possível entre as observações e a maior variabilidade possível entre os "clusters". A avaliação dessa variabilidade auxiliaria os profissionais da área médica na tentativa de entendimento do comportamento de seus pacientes não dependendo mais apenas do reporte dos responsáveis dos assistidos.

Obviamente o resultado encontrado, devido ao fato de ser utilizado "Dataset" que não representa casos reais, não refletirá "clusters" de pacientes reais, mas como sabido e objetivado neste estudo, pleiteia-se mostrar o grande potencial desta ferramenta aos profissionais da área médica que tratam pacientes dentro deste espectro.

Recomenda-se que o pesquisador documente com clareza e transparência as premissas escolhidas na elaboração de seu estudo (esquemas de aglomeração não hierárquicos ou hierárquicos e métodos de encadeamento) de forma a conseguir associar o resultado obtido com a respectivas premissas utilizadas possibilitando desta forma que simulações possam ser feitas e facilmente compreendidos seus resultados

Com o aumento do poder computacional e com a evolução dos softwares nos últimos anos, surgiram novas técnicas para analisar e realizar os agrupamentos (clusterização) para as mais variadas áreas do conhecimento, no entanto o exposto neste trabalho acadêmico, até a presente data, são considerados os métodos mais empregados na busca da resolução dos desafios de agrupamentos de observações em grupos homogêneos quanto olhado para seu interior, e o mais heterogêneo possível quando comparador um grupo com o outro.

Material e Métodos

Afim de obter-se a distribuição das duzentas observações contidas na amostra em grupos heterogêneos entre si foi utilizada a técnica “Clustering”, pertencente ao grupo “Usupervised Machine Learning “ utilizando o software R para execução dos algoritmos. O conceito de “Clustering”, que é agrupar observações em grupos onde é almejada a menor variabilidade possível entre as observações dentro do mesmo grupo e a maior variabilidade possível entre os grupos, reflete exatamente o principal objetivo deste trabalho.

O método de cálculo selecionado para calcular as distâncias entre observações foi a Distância Euclidiana. Através desta metodologia de cálculo foi construído para todas as simulações a matriz de distâncias utilizando a função “dist()” nativamente disponibilizada pelo software R. Desta forma, foram executados vários protótipos utilizando-se tanto o método hierárquico quanto o não hierárquico. O método hierárquico de acordo com Belfiore e Fávero (2017) agrupa a cada iteração do algoritmo observações a outras observações ou grupos previamente formados em função do método de encadeamento escolhido bem como também no método de cálculo de distância escolhido. Neste esquema de aglomeração é possível observar a formação dos clusters passo a passo até a formação de um único cluster contendo todas as observações. Já o método não hierárquico, o número de grupos é definido preliminarmente pelo pesquisador. A grande vantagem da utilização do método hierárquico de aglomeração é a possibilidade de visualizar os agrupamentos sendo realizados, podendo o pesquisador utilizar-se de ferramentas gráficas, com por exemplo o gráfico de dendograma, tomar analisar a formação dos grupos. Este método não performa tão bem com grandes quantidades de observações tornando-se difícil a análise gráfica. Outro ponto relevante que deve ser levando em consideração é que o método hierárquico é computacionalmente mais demandante que o não hierárquico, onde já é definido inicialmente o número de clusters. Quanto as vantagens de utilização do método não hierárquico, podemos citar que ele é computacionalmente menos demandante que o método hierárquico, ele performa melhor com grande quantidade de observações quando comparado com o método de aglomeração hierárquico, e como principal vantagem, a possibilidade de previamente definir o número de grupos a serem criados com base na realidade de negócios na qual o pesquisador fundamentou sua pesquisa. Em muitos casos, o número de clusters precisa ser adequado a realidade de negócios mesmo que porventura haja a possibilidade de chegar-se a menor variabilidade intra-grupo em números maiores de clusters, a realidade de negócios onde o pesquisador impões como restrição o número de clusters, seja por motivos financeiros, ou operacionais, nestes casos a utilização do método de aglomeração não hierárquico apresenta-

se mais apropriado. Esta é justamente a realidade desta pesquisa, onde o número de cluster não deveria, segundo orientação dos médicos envolvidos neste trabalho, ultrapassar a quantidade de 3 cluster.

Mesmo com a premissa de números de cluster sendo definida pelos clientes desta pesquisa, impondo restrições aos resultados apresentados pelo método hierárquico, a título comparativo, foram realização diversos protótipos utilizando-se o método hierárquico. Foram empregados nestes protótipos alguns métodos para escolha do número ideal de “Clusters”, sendo eles o Coeficiente R^2, o método “*Elbow*” e o método "*Silhouette*". O coeficiente $R^{2,}$ de acordo com Belfiore e Fávero (2017) é uma medida de ajuste de um modelo estatístico linear generalizado aos valores observados de uma variável aleatória podendo variar entre 0 e 1 sendo em muitas ocasiões expresso em termos percentuais. Desta forma, buscou-se obter o maior valor possível deste coeficiente levando em consideração a viabilidade do número de clusters apresentado. O método R^2 visa demonstrar a variabilidade total das observações tanto da amostra antes de ser dividida em grupos como da soma das variabilidades das observações dentro dos grupos e dos grupos entre si. O principal objetivo é identificar o quanto a variabilidade total diminui à medida que são formados os grupos. Outra aplicação muito interessante para este método é calcular o coeficiente R^2 para cada quantidade de grupos pleiteada tornando possível identificar uma boa sugestão de quantidade de grupos ao verificar-se que a variabilidade total passa a reduzir pouco comparado como aumento do número de grupos. Por óbvio, quanto maior for a quantidade de grupos, menor será a variabilidade total, bem como quanto maior a quantidade de grupos, maior será o R^2. A grande vantagem da utilização deste coeficiente é identificar o momento em que o acréscimo na quantidade de grupos não reflete de forma tão intensa na redução da variabilidade dentro dos grupos.

O Método “Elbow”, de acordo com Thorndike (1953) e Ketchen e Shook (1996) , visa determinar o número de clusters em um conjunto de dados. Tal método consiste em traçar a variação explicada em função do número de clusters e escolher o cotovelo da curva do gráfico apresentado como o número de clusters a utilizar. Isto é, ao observar o gráfico das duas grandezas citadas, sugere-se o número de clusters a serem adotados. Este método contempla a elaboração de um gráfico bidimensional a quantidade de grupos em um eixo e a soma dos quadrados total dentro do cluster no outro eixo. O gráfico gerado por este método assemelha-se com um cotovelo humano, por isto o nome do método. Pleiteia-se ao analisar este gráfico, visualizar um número sugestivo de clusters em que a variabilidade ainda represente uma significativa redução quando comparado ao número de clusters imediatamente anterior. De forma análoga ao método R^2, o gráfico do método “Elbow” visa auxiliar na escolha do número ideal de cluster ao observar a relação entre variabilidade total e número de clusters.

Já o Método "Silhouette", segundo Rousseeuw (1987), refere-se a um método de interpretação e validação de consistência dentro de agrupamentos de dados. A técnica fornece uma representação gráfica sucinta de quão bem cada objeto foi classificado. O coeficiente de silhueta é a medida da relação entre um ponto e os membros de seu próprio grupo. O coeficiente de silhueta de todo o set é definido pela média dos coeficientes calculados para cada ponto.

Como explicitado anteriormente, o número de clusters a ser formado, para ser viável seu estudo, sofreria a restrição de não ultrapassar 3 grupos. Esta restrição foi imposta pela área médica envolvida na pesquisa. De posse desta restrição, evidenciou-se que a utilização do método não hierárquico seria mais apropriada visto que neste, é pré definida a quantidade de grupos a serem formadas, sendo desta forma escolhido este método para apresentação dos resultados aos clientes.

Além dos métodos acima citado, também foram experimentados números aleatórios (k) de “clusters” sempre ponderando o resultado obtido sob a ótica da área de negócio (no caso em questão, os médicos Neurologistas e Psiquiatras que tratam de pacientes dentro do espectro autista). Com muito critério, foram arbitradas aproximadamente quinhentas observações, nove variáveis quantitativas sendo elas (idade em anos, peso em quilos, índice de Glicemia em mg/dl, índice de Prolactina em mg/ml, índice de Colesterol HDL em mg/dl, índice de Colesterol LDL em mg/dl, índice de Triglicérides em mg/dl, dosagem diária de utilização da medicação Risperidona em mg, dosagem diária de utilização da medicação Valproato Sódico em mg, e duas variáveis qualitativas sendo elas: sexo biológico e classificação do diagnóstico médico para o paciente, sendo estas classificações: Grau Leve, Grau Moderado e Grau Severo. Estas variáveis qualitativas, por motivos óbvios, não participaram na realização dos cálculos, mas contribuíram na elaboração de parecer explicativo de cada “Cluster” criado após cada simulação.

Como a grande maioria das variáveis quantitativas escolhidas para este estudo apresentaram escalas muito diferentes, foi utilizado a metodologia de padronização, utilizando a função “scale()” do software R, dos valores antes da elaboração da matriz de distâncias afim de neutralizar possíveis distorções nos resultados. Todas as variáveis utilizadas, tanto qualitativas como quantitativas fazem total sentido no acompanhamento de praticamente todo o paciente dentro do espectro autista e foram cuidadosamente escolhidas para integrarem o “Dataset” do estudo.

Também foram realizados cálculos de análise de “Cluster” não hierárquico, “k-means”, onde foram selecionados os centróides iniciais, lidos as observações e atualizado os centróides até ocorrer a convergência na intenção de ao final de todas as rodadas, cada observação estar associada ao centróide mais próximo da mesma. Na utilização deste método

foram realizadas diversas iterações na intenção de encontrar a melhor distribuição das observações com seus respectivos centróides.

Como o número de observações utilizados foi consideravelmente grande, este método, “k-means” apresentou-se mais intuitivo quando comparado a análise do dendograma gerado no método hierárquico facilitando a visualização dos grupos e consequentemente compreensão da heterogeneidade entre eles. Preservamos a premissa de reprodução dos estudos realizados ao incluirmos no campo “Apêndices” tanto o “Dataset” como os códigos da linguagem R utilizados para a realização dos cálculos.

Após a definição do número de clusters e a execução do algorítimo que distribui as duzentas observações da amostra nos grupos, iniciou-se a análise descritiva na intenção de efetuar sumarização dos resultados bem como facilitar leitura e análise do resultado encontrado. Segundo Belfiore e Fávero (2017), análise descritiva é uma técnica que visa sintetizar e descrever as principais características observadas numa amostra. Esta técnica pode utilizar-se de tabelas, gráficos e medidas resumo (apenas quando analisadas variáveis quantitativas) na intenção de proporcionar ao pesquisar melhor entendimento dos dados. Através da aplicação desta técnica é possível obter estatísticas sobre a amostra e de posse dessas estatísticas torna-se possível a elaboração hipóteses que auxiliará a compreensão do comportamento dos dados.

As medidas-resumo podem ser divididas em medidas de posição (também chamada de localização), medidas de dispersão (ou variabilidade) e medidas de forma. Com relação as medidas de posição, podemos citar a média, mediana, moda, quartis, decis e percentis. Estas medidas ajudam a visualizar como os valores estão dispostos (localização dos dados em relação ao eixo dos valores assumidos pela variável ou característica em estudo. As medidas de posição são subdivididas em medidas de tendência central e medidas separatrizes. A principais medidas de tendência central são a média, mediana e moda, já as principais medidas separatrizes são os quartis, os decis e os percentis. Se ordenarmos os dados de forma crescente, o primeiro quartil descreverá os primeiros 25% das observações, o segundo quartil descreverá a primeira metade das observações (mesmo conceito da mediana), o terceiro quartil corresponderá a 75% das observações e por fim o quarto quartil representará a totalidade das observações. O gráfico “Boxplot” é uma boa ferramenta para análise visual de como está distribuído os valores da amostra. Nele são apresentadas 5 medidas de posição, sendo elas, o valor mínimo, o primeiro quartil, o segundo quartil (mediana), o terceiro quartil e valor máximo.

Com relação as medidas de dispersão (ou variabilidade) podemos destacar a amplitude, o desvio médio, a variância, o desvio padrão, o erro padrão e coeficiente de variação e através destas medidas é possível analisar a dispersão ou variabilidade dos dados.

A variância é uma medida de dispersão ou variabilidade que demonstra o quanto os dados estão dispersos em relação à média, portanto quanto maior for a variância, maior será a dispersão dos dados. Com relação as medidas de forma, podemos citar a assimetria e curtose como sendo as principais medidas.

Com base nas observações da amostra, foi selecionada a Média como medida de tendência central na intenção de compreender o comportamento conjunto das variáveis e as medidas Variância e Desvio Padrão para avaliarem a o quanto os valores estavam distantes de média, ou seja, para mensurar a dispersão das observações dentro de seus respectivos grupos.

É valido ressaltar que os resultados obtidos através da técnica de clusterização, tanto através do método hierárquico como do não hierárquico podem ser comparadas a fim de proporcionar ao pesquisador analisar qual opção é a mais aderente ao seu problema de negócios. A mesma afirmação é verdadeira quanto a escolha das medidas de cálculo de distâncias ou mesmo aos métodos de escolha do número ideal de cluster. A técnica de agrupamentos não tem por objetivo tácito definir um número ideal de cluster nem tão menos suas respectivas composições, mas sim auxiliar o pesquisador na escolha das melhores ferramentas disponíveis pela referida técnica de forma a solucionar da melhor forma possível o desafio fruto do objeto da atual pesquisa.

Resultados Preliminares

Foram realizados diversos experimentos utilizando tanto métodos de aglomeração hierárquico como não hierárquico bem como testados os quartos métodos de encadeamento citado na sessão anterior a esta.

Conforme a Figura 1, pode-se observar um gráfico de dendograma onde foi utilizado o método de encadeamento “Complete Linkage”. Compõem este dendograma todas as duzentas observações analisadas nesta pesquisa.

Figura 1

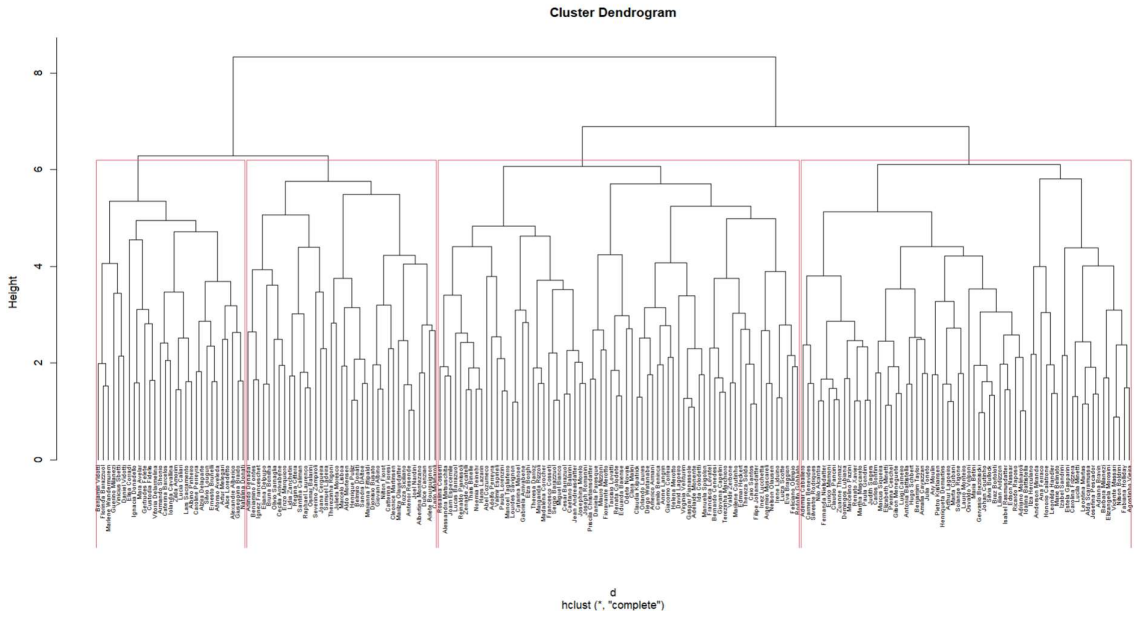

Figura 1
Gráfico Dendograma Cluster Hierárquico “Complete Linkage”
FONTE: ISRAEL JOSÉ MONTEIRO CARVALHO (2022)

Utilizando o método de encadeamento “ward.D” foi gerado outro dendograma afim de avaliar o resultado frente ao método de encadeamento anteriormente utilizado. Neste dendograma também foi utilizado todas as observações da amostra.

O dendograma do segundo método utilizado pode ser observado na Figura 2.

Figura 2

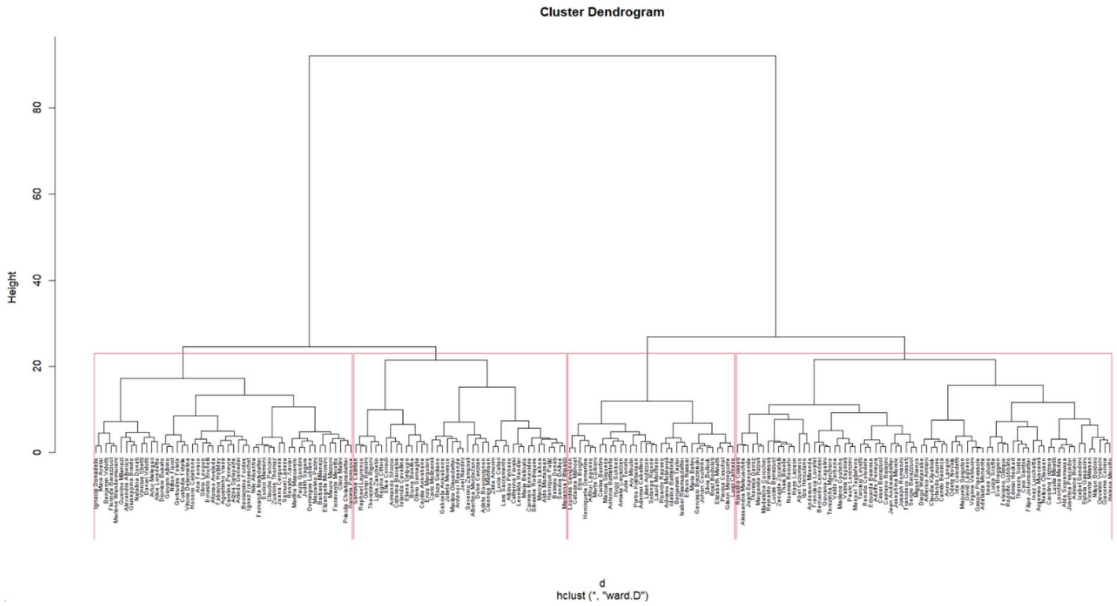

Figura 2

Gráfico Dendograma Cluster Hierárquico “ward.D”
FONTE: ISRAEL JOSÉ MONTEIRO CARVALHO (2022)

Como pode-se concluir, a interpretação de ambos os gráficos se tornou difícil devido ao grande número de observações pertencentes a amostra, o que reafirma uma das limitações da utilização do método hierárquico, que é a difícil interpretação dos resultados gerados quando a amostra possui grande quantidade de observações. Visto que analisar a formação dos clusters a medida que elas vão acontecendo é a grande vantagem da utilização do método hierárquico, somado com a restrição de número de grupos a serem criados imposta pelo cliente (médicos), optou-se por aprofundar as análises nas opções ofertadas pelos métodos não hierárquico de aglomeração.

Mesmo o número de clusters inicialmente sendo restringido pelo cliente, a fim se obter mais detalhamento sobre a qualidade dos grupos, principalmente quanto a variabilidade associada a escolha do número de clusters, foi calculado o coeficiente R^2 bem como criado um gráfico de Elbow, apresentado na Figura 3. Ao analisa-lo tanto o coeficiente R^2 quanto o gráfico Elbow, percebeu-se que a variabilidade total é drasticamente reduzida quando a amostra é separada em 2 grupos, mas ao incrementar um ou mais grupos, a variabilidade não expressa redução tão significativa.

Figura 3

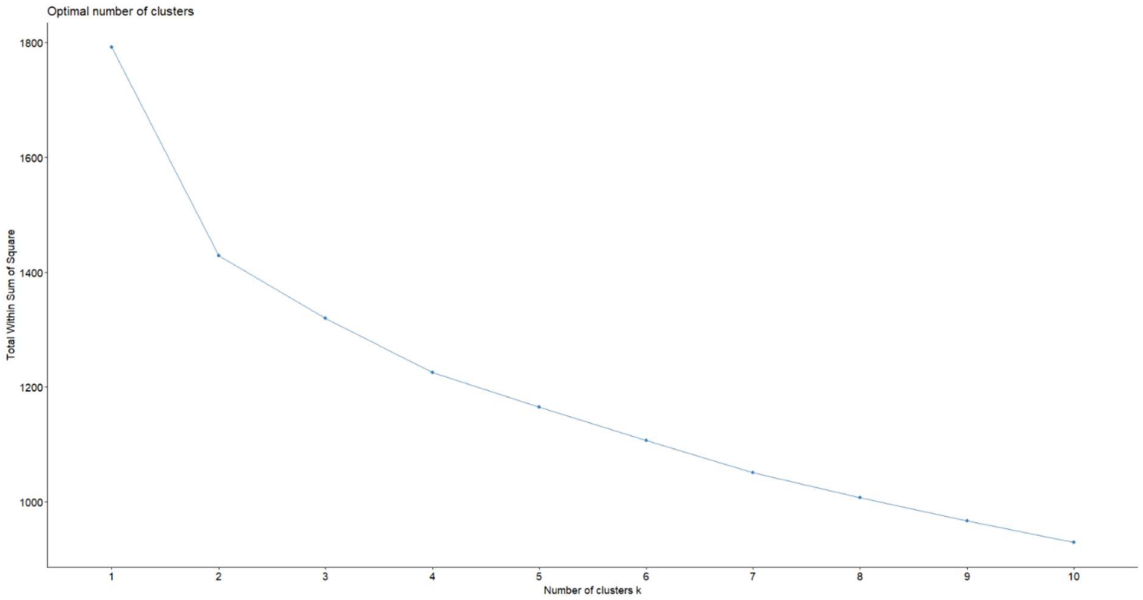

Figura 3

Gráfico Elbow
FONTE: ISRAEL JOSÉ MONTEIRO CARVALHO (2022)

A mesma conclusão é obtida ao analisar os dendogramas, principalmente o que utilizou-se do método de encadeamento “ward.D”. Neste sentido, quando comparada a restrição do número de grupos estabelecida previamente com a as sugestões de número de grupos através dos gráficos de dendograma, Elbow e do cálculo do R^2, percebeu-se que mesmo trabalhando com número pequeno de grupos, a variabilidade seria significativamente reduzida quando comparada com a variabilidade total da amostra.

A título de curiosidade, observou-se que o resultado apresentado pelo método hierárquico utilizando a função hclust() sobre um “dataframe” com suas variáveis padronizadas através da função scale(), sendo utilizado o método “Euclidian” para calcular as distâncias e o método “Complete Linkage” como método de encadeamento, é o mesmo resultado apresentado pelo método não hierárquico “k-Means” utilizando-se a função kmeans(). Nos experimentos desta pesquisa, em ambos os métodos todos as observações foram alocadas aos mesmos grupos, e referente as estatísticas descritivas, as médias de cada variável dentro de cada respectivo grupo foi exatamente igual para os dois métodos, hierárquico (respeitando as premissas acima descriminadas) e não hierárquico. Ao avaliar a dispersão dos dados através do Desvio Padrão, as diferenças entre os métodos surgiram apenas após a nona casa decimal.

Diante de todo o exposto anteriormente, optou-se para a apresentação da análise de agrupamentos, o método não hierárquico “k-Means”, onde foi utilizado como parâmetro principal de input desta técnica o número máximo de clusters informado pelo cliente, quarto clusters.

Após executado o algoritmo de clusterização, foi gerado estatísticas descritivas para elaboração de hipóteses. Foram utilizados a média e o desvio padrão para interpretar os resultados de cada variável frente a seus respectivos grupos.

A amostra foi dividida em 4 grupos sendo que no primeiro foram alocadas 70 observações, no segundo 64, no terceiro 37 e no quarto 29, totalizando as 200 observações da amostra.

Devido à grande heterogeneidade das observações da amostra, observou-se através de análise do desvio padrão a existência de uma considerada dispersão dos dados para o grupo 1 e variável “Idade Anos” o desvio padrão de 2,5 anos sendo que a média de idade dos pacientes deste grupo é 4,3 anos. Já ao observar o grupo 4, a dispersão é muito menos quando observada a mesma variável, média de idade igual 12,4 anos e o desvio padrão igual 1,8 ano. Testes realizados com número de grupos maiores apresentaram variabilidade significativamente menor do que com apenas 4 grupos, mas foi mantido a quantidade de grupos inicialmente definida pelos médicos.

Ao observar os valores fruto das estatísticas descritivas, diversas hipóteses sobre o comportamento dos dados foram elaboradas na tentativa de criar um racional onde fosse facilitada a tomara de decisão. Por óbvio, como citado anteriormente, os dados constantes nesta amostra foram arbitrados única e exclusivamente para convidar os clientes a pensarem sobre uma forma diferente de análise dos pacientes, ou seja, analisa-los não apenas sob uma ótica individual, mas sim também sob a luz dos comportamentos de outros pacientes com caraterística semelhantes.

A Tabela 1 apresenta as médias de todas as variáveis para seus respectivos grupos. Foram destacados com as cores vermelhas e verdes valores médios extremos “ruins” e “bons” para cada variável que representa o resultado de um exame laboratorial na intenção de proporcionar ao leitor a identificação visual do quanto cada variável está alinhada com os valores desejados. Obviamente a caraterística de cada variável foi respeitada na hora da coloração de seus valores. As variáveis em que são desejados valores baixos são: Glicemia, Prolactina, LDLr e Triglicérides. Apenas a variável HDLb deseja-se que possua valores elevados. É válido ressaltar que estes não são desejos reais absolutos adotados pela medicina. Na prática existem valores de referência que podem mudar conforme faixa etária, sexo biológico, ou outros critérios. Foram adotadas as premissas acima apenas para a construção de possíveis hipóteses. Apesar de não refletirem exatamente a forma que um

profissional médico analisa o resultado de um exame laboratorial, a construção demonstrada possibilita a discussão com relativa fundamentação.

Tabela 1 – Média

Grupo	n	Idade Anos	Peso kg	Glicemia mgdl	Prolactina ngmL	HDLb mgdl	LDLr mgdl	Triglicérides mgdl	Risperidona DoseDiária (mg)	Valproato DoseDiária (mg)
1	70	4,3	20,2	110	40	53	111	202	13	850
2	64	6,8	30,8	116	31	76	101	202	16	1355
3	37	12,3	54,2	100	41	64	134	182	20	2418
4	29	12,4	63,5	125	32	62	124	241	13	3224

Fonte: Resultados originais da pesquisa

Ao analisar os dados apresentados na Tabela 1, a primeira situação que foi chamou a atenção foi o peso médio dos pacientes do grupo 3 (54,2kg) ser tão diferente do peso médio dos pacientes do grupo 5 (63,5 kg). São 9,3kg de diferença média sendo que a idade média entre os pacientes destes dois grupos é praticamente a mesma, 12,3 anos e 12,4 anos. Percebeu-se que aparentemente o peso ideal para uma criança de 12 anos aproxima-se mais da média apresentada no grupo 3 do que no grupo 4, o que sugere um certo grau de obesidade nos pacientes do grupo 4. Observou-se também que embora os pacientes do grupo 4 apresentem obesidade, o valor médio de LDLr, também chamado de colesterol ruim, apresentou valor inferior quando comparado com os valores apresentados pelos pacientes do grupo 3, o que por si só não faz sentido se não for considerada outras variáveis que possivelmente podem influenciar nos níveis de LDLr. A mesma observação serve ao ser analisado os níveis de HDLb, também conhecido como colesterol bom. Busca-se sempre valores elevados para este indicador e na tabela resumo estudada os pacientes tanto com obesidade quanto os não obesos apresentaram valores muito próximos, 62 e 64 mgdl.

Percebeu-se também que embora as médias de idades entre os pacientes do grupo 3 e 4 são praticamente iguais, o consumo diário das medicações Risperidona e Valproato é muito diferente, possivelmente sendo mais um fator que fez com que o algoritmo tenha separado estes pacientes com a mesma média de idade em 2 grupos. Ainda analisando os pacientes do grupo 4, quando comparados com os 70 pacientes do grupo 1, embora ambos os grupos apresentem mesma média de consumo da medicação Risperidona, 13mg/dia, a diferença na dosagem diária de Valproato é extrema, o que talvez possa fomentar a hipótese de que o consumo de Valproato em dosagens mais altas pode afetar na intensidade da fome dos pacientes, visto que os pacientes do grupo 1 não apresentam sinal de obesidade mesmo tomando a mesma dosagem média de Risperidona por dia que os pacientes do grupo 4 (pacientes obesos). É sabido que a resposta dos pacientes as medicações possivelmente seja

fruto não apenas da escolha das drogas isoladamente, mas sim também das associações entre elas visto que muitos pacientes do espectro autista não tomam apenas uma medicação. Sob esta ótica, avaliando a dosagem conjunta de Risperidona e Valproato percebe-se que o grupo 2, onde constam 64 pacientes, apresentou um consumo médio intermediário de Risperidona e Valproato, e também apresentou os melhores níveis de Prolactina (menores níveis registrados), HDLb (maiores valores médios entre todos os grupos) e LDLr (menores níveis médios de colesterol). É sabido também que possivelmente essas duas medicações causem influência sobre alguns dos índices laboratoriais aqui estudados, portanto, tentar identificar um grupo de pacientes onde estes índices estão bons pode ser uma boa estratégia para a compressão dos perfis dos pacientes.

Referente as dosagens conjuntas das duas medicações, percebeu-se que com o aumento do peso médio dos grupos, aumenta-se as dosagens de ambas, Risperidona e Valproato, com uma única exceção que é o consumo de Risperidona pelos 29 pacientes do grupo 4, onde é encontrada a maior média de peso (63,5kg) mas a dosagem média de Risperidona apresentou o menor valor (13mg/dia), consumo este igual ao consumo dos 70 pacientes do grupo 1 onde a média de peso é a menor de todos os grupos (20,2kg).

Referente a Glicemia, percebeu-se que o grupo 3 apresentou o menor índice médio registrado (100mgdl). Percebeu-se também que foi neste grupo o registro da maior dosagem média de Risperidona (20mg/dia) e uma dosagem média relativamente alta de Valproato (2418mg/dia). Analogamente ao excelente índice glicêmico observado nestes 37 pacientes, foi registrado também melhor índice de Triglicérides (182mgdl). Em dicotomia a estes dois índices, os índices de Prolactina e LDLr apresentaram as piores médias entre grupos, respectivamente 41mgdl e 134mgdl.

Ao analisar o HDLr, percebeu-se que em três dos quatro grupo estudados, o grupo 1, onde a média de idade é a menor de todos os grupos (4,3 anos) bem como a média de peso é a menor de todas (20,2kg), apresentou média (53mgdl) substancialmente pior que os outros grupos. Chama a atenção que este é o único índice laboratorial que apresentou valores altamente indesejados, todos os outros apresentaram valores relativamente bons quando comparados aos outros grupos.

Em complemento a medida de tendência central utilizada nestas análises, é muito importante demonstrar a dispersão dos dados dentro do grupo. Como citado anteriormente, foi utilizado o Desvio Padrão para avaliar tal comportamento. A Tabela 2 demonstra o Desvio Padrão para todas as variáveis e grupos utilizados nesta pesquisa.

Tabela 2 – Desvio Padrão

Grupo	n	Idade Anos	Peso kg	Glicemia mgdl	Prolactina ngmL	HDLb mgdl	LDLr mgdl	Triglicérides mgdl	Risperidona DoseDiária (mg)	Valproato DoseDiária (mg)
1	70	2,5	13,0	32	14	15	31	41	9	634
2	64	3,0	13,7	28	14	13	30	45	7	677
3	37	2,1	10,0	31	14	17	48	45	7	673
4	29	1,8	11,8	28	15	14	40	29	9	824

Fonte: Resultados originais da pesquisa

Com relação a dispersão dos dados, a variável ValproatoDoseDiária (mg) no grupo 1 apresentou o maior desvio padrão quando comparado com a média (média 850mg/dia e desvio padrão 634mg/dia). Ao analisar individualmente as variáveis deste grupo, percebe-se que o consumo deste medicamente é muito heterogêneo entre os pacientes, sendo percebido 5 pacientes com 0mg de consumo por dia, o que obviamente compromete a média de consumo deste medicamento aumentando a dispersão dos dados. Em contrapartida ao desvio padrão identificado na variável e grupo acima, percebeu-se o inverso na variável Triglicérides dentro do grupo 4, onde a média é 241mgdl e o desvio padrão 29gmdl. A compreensão deste fator é extremamente relevante pois permitirá aos pesquisadores considerar a média deste índice como efetivamente representativa para os indivíduos deste grupo. Ainda com relação ao pequeno desvio padrão desta variável dentro do grupo em questão, é de grande importância o conhecimento destas medidas visto que a média de Triglicérides do grupo 4 (241mgdl) é a maior dentro todos os outros grupos, e com certeza os médicos terem a possibilidade de confiar que o valor apresentado pela média é apropriado para se analisar todos os 29 pacientes do grupo é um ponto muito relevantes a ser considerado.

Possivelmente caso fosse trabalhado com um número maior de clusters esta dispersão seria minimizada ao melhorar a distribuição das observações sobre este número maior de clusters, mas ao atendermos a premissa de número máximo de cluster igual a 4 grupos, fatalmente algumas observações seriam agrupadas com valores de variáveis bem distintas umas das outras.

Estas são algumas hipóteses levantadas por uma pessoa que não tem quaisquer conhecimentos científicos sobre medicina. É válido ressaltar que se um profissional da área médica, que por natureza possui conhecimentos sobres os exames e medicamentos contidos nesta pesquisa, analisar os resultados demonstrados nas medidas resumo aqui apresentadas, provavelmente seriam elaboradas hipóteses muito mais ricas e explicativas ou mesmo descartadas hipóteses que não conforme sua teoria subjacente não condizerem com a realidade.

Considerações Finais

Conforme apresentado nesta obra, cremos que a utilização da técnica de clusterização pode ser de extrema utilidade na análise de comportamento dos pacientes não apenas dentro do espectro do autismo bem como para pacientes pertencentes a quaisquer outras síndromes ou portadores de qualquer patologia. A possibilidade de analisar o comportamento das variáveis de um paciente frente ao comportamento das variáveis de diversos outros pacientes com características semelhantes pode ser algo realmente transformador no modo como são estudados os casos médicos. Cita-se que a utilização de técnicas de agrupamento são apenas uma fração das inúmeras possibilidades de análises que seriam viabilizadas através de outras técnicas de “Machine Learning”, tanto supervisionadas com não supervisionadas. A expectativa destes resultados é que profissionais da área médica (psiquiatras e neurologistas) que são plenamente aptos para a leitura, interpretação e possíveis descobertas de padrões, desses resultados, percebam a grandeza exponencial que é avaliar diversas variáveis, muitas vezes inimaginadas, correlacionadas e agrupadas entre pacientes com mesmas características. Ter a possibilidade de agrupar pacientes com características afim, com a mínima variabilidade possível entre cada um deles, gerando clusters com a maior variabilidade possível entre si é uma abordagem extremamente promissora na forma de análise deste tipo de paciente.

Referências

BOLFARINE, H.; BUSSAB, W. O. **Elementos de amostragem**. São Paulo: Edgard Blücher, 2005.

FÁVERO, L. P.; BELFIORE, P **Manual de Análise de Dados**: Estatística e Modelagem Multivariadas com Excel®, SPSS® e Stata. Rio de Janeiro: Elsevier, 2017

KETCHEN, David J.; SHOOK, Christopher L. (1996). **The application of cluster analysis in Strategic Management Research: An analysis and critique**. Strategic Management Journal. 17 (6): 441–458. doi:10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819> 3.0.CO;2-G

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

MOORE, D. S.; McCABE, G. P.; DUCKWORTH, W. M.; SCLOVE, S. L. **Estatística empresarial: como usar dados para tomar decisões**. Rio de Janeiro: LTC Editora, 2006.

NAVIDI, W. **Probabilidade e estatística para ciências exatas**. Porto Alegre: Bookman, 2012

NEUFELD, J. L. **Estatística aplicada à administração usando Excel**. São Paulo: Prentice Hall, 2003.

OLIVEIRA, F. E. M. **Estatística e probabilidade**. 2. ed. São Paulo: Atlas, 2009.

ROUSSEEUW, P J **Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis**. Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

THORNDIKE, R. L (December 1953). **Who Belongs in the Family?**. Psychometrika. 18 (4): 267–276. doi:10.1007/BF02289263

Apêndices

Tabela 1. “Dataset” utilizado na realização dos algoritimos de “Clusters”

(continua)

Paciente	Sexo	Classificação	Idade Anos	Peso kg	Glicemia mgdl	Prolactina ngmL	HDLb mgdl	LDLr mgdl	Triglicérides mgdl	Risperidona DoseDiária (mg)	Valproato DoseDiária (mg)
Priscila Christiansdatter	F	Leve	6	31,3	69	28	76	94	202	3	1159
Izabel Sandrini	F	Severo	8	44,3	79	12	88	166	175	24	1853
Lucas Barrazuol	M	Leve	8	44,8	102	36	36	118	160	0	2197
Ary Moulin	M	Leve	5	24,2	129	44	93	141	253	21	1167
Virginia Venturim	F	Severo	2	7,7	67	48	51	120	208	28	370
Josephina Morelo	F	Leve	6	27	150	50	40	128	133	19	1046
Tarcisio Lovatti	M	Moderado	5	30,5	93	41	37	90	255	21	1408
Maria Betini	F	Moderado	8	33,8	151	31	69	71	153	20	0
Irene Liduino	F	Leve	3	12,3	72	20	72	70	157	17	435
Carlos Bonfim	M	Moderado	7	28,3	120	30	78	120	201	17	1741
Therezinha Rigoni	F	Leve	15	52,8	105	60	39	173	136	16	2376
Joseph Romanini	M	Moderado	5	30,5	140	58	35	144	178	25	1580
Zenaide Zocatelli	F	Leve	5	27,3	82	34	43	94	205	5	1719
Caio Santos	M	Leve	2	9	106	21	49	67	238	11	457
Estela Gasparini	F	Moderado	5	17,7	70	12	91	144	151	29	554
Cecilia Arrivabene	F	Moderado	11	53	114	16	41	191	261	16	2067
Paula Gondim	F	Moderado	8	31,2	116	52	58	113	264	13	1340
Feliciano Olimpio	M	Leve	3	13,7	67	43	74	53	133	8	520
Leonilda Handler	F	Moderado	12	40	81	17	62	62	148	15	1878
Pietro Kristiansen	M	Leve	4	24,5	140	59	84	153	201	24	870
Claudio Pancini	M	Severo	10	50,7	130	40	75	118	256	19	2248
Odete Novais	F	Leve	4	18	67	59	46	135	264	11	1008
Antenor Resende	M	Severo	10	42,3	65	54	75	146	148	5	2412
Jean Andreasdatter	M	Severo	5	23,8	149	58	60	113	174	29	827
Marilza Olausdatter	F	Moderado	13	51	109	41	93	117	130	4	1786
Benjamin Vidotti	M	Severo	14	52,1	140	58	66	165	268	9	3385
Beniamino Fontes	M	Leve	14	65,7	63	18	58	135	228	17	3162
Enzo Marquioro	M	Moderado	15	68,2	132	19	54	173	247	24	2606
Roza Siciliano	F	Leve	10	43,2	68	48	79	152	132	16	1730
Mauro Mizzoni	M	Leve	11	37	71	40	94	106	264	3	1813
Joana Cergueira	F	Moderado	13	55,9	98	50	64	185	253	26	2145
Guerino Milanezi	M	Leve	12	70,2	141	11	46	181	255	6	4212
Elisa Corradi	F	Moderado	15	85,8	80	39	81	155	241	16	5577
Osvaldo Volponi	M	Leve	3	11,3	83	23	94	52	204	29	354
Alexandre Alberico	M	Moderado	13	70,8	121	16	64	66	268	0	3892
Ilza Herculano	F	Leve	7	36,7	150	14	81	91	177	0	0
Francesco Cosseti	M	Severo	8	41	156	50	40	155	247	14	1806
Arlindo Denadai	M	Moderado	10	44,7	62	17	37	85	205	21	2111
Solange Rigone	F	Leve	1	7,1	127	47	78	87	153	26	397
Custodia Fidelis	F	Moderado	14	73,2	74	43	42	118	273	16	3292
Judith Sagers	F	Leve	9	30,2	115	37	62	99	264	9	1572
Filipe Johansdatter	M	Moderado	2	7	96	14	47	74	226	3	455
Paulo Lorenzini	M	Moderado	5	17,6	148	52	62	98	237	13	845
Arthur Lagoeiro	M	Leve	2	8,5	115	43	94	102	154	6	512
Anderson Miranda	M	Leve	15	51,5	138	11	76	113	166	0	0
Cesar Barrazuol	M	Moderado	5	27,1	98	58	40	102	162	26	1428
Thereza Solda	F	Severo	1	5,7	147	25	39	56	232	6	251
Giovana Capelini	F	Leve	2	7,7	142	52	35	151	222	10	393
Henrique Puljiz	M	Moderado	12	46,7	125	44	57	86	157	21	2506
Belmiro Donati	M	Leve	10	53,5	143	36	53	83	160	27	2583
Julieta Moresco	F	Moderado	12	56,2	63	56	55	104	150	14	3594
Olivio Sarnaglia	M	Moderado	15	53,1	101	12	39	187	183	28	2264
Josefina Pescador	F	Leve	7	39,4	90	26	52	105	229	15	1617
Valdir Zerbone	M	Leve	2	5	121	36	43	88	271	16	217
Honorio Calatrone	M	Moderado	11	46	83	20	69	98	189	5	1610
Diego Watanabe	M	Severo	1	6,7	105	35	45	106	181	28	348
Valmir Espineli	M	Severo	6	34,2	139	45	68	68	237	3	1266
Albertina Montezano	F	Moderado	13	54,1	136	53	88	137	132	20	3062
Margarida Rizzoli	F	Leve	3	11,1	140	25	54	141	172	24	464
Adriano Molinaroli	M	Moderado	8	35,9	144	37	79	88	217	29	1309
Ademar Caballero	M	Leve	4	23,3	146	58	90	73	159	19	1186
Lilian Ardizzon	F	Moderado	5	25,7	95	35	84	86	151	22	1188
Lars Sacramento	M	Moderado	13	72	157	30	79	77	189	28	2944
Marcelo Lisboa	M	Leve	13	41,8	151	46	43	77	145	3	2046

Dulce Canzian	F	Severo	14	73,6	128	59	79	113	171	26	2629
Camilo Beninca	M	Moderado	3	10,5	155	44	79	81	219	28	447
Benjamim Taylor	M	Leve	8	43,3	139	16	87	112	132	13	1516
Caetano Balarini	M	Moderado	5	17,3	113	57	37	79	181	20	814
Domingos Lubiana	M	Leve	6	28,7	109	50	69	84	249	23	1253
Ernesto Brunelli	M	Leve	12	60,5	85	26	84	111	235	15	2480
Argemiro Muscareli	M	Moderado	2	8,1	63	24	64	136	261	8	300
Claudia Kilpatrick	F	Severo	4	17,8	143	53	43	123	249	29	619
Sergio Barazzuol	M	Leve	7	36	157	36	37	115	183	10	1586
Giacomo Correia	M	Moderado	2	7,9	121	23	70	127	240	29	309
Florinda Barazzuol	F	Moderado	13	59,2	114	41	71	196	234	6	2781
Idalina Battistella	F	Moderado	8	30,7	156	26	78	84	201	27	1587
Carolina Frizzera	F	Leve	4	14,9	70	22	42	128	255	27	813
Ignez Franceschet	F	Severo	14	49,6	87	29	49	156	219	15	2530
Marta Curbani	F	Moderado	4	15,1	115	52	82	119	160	8	726
Fabiana Holliday	F	Severo	8	32,6	92	20	70	98	220	24	1927
Noemia Reinehr	F	Leve	7	24,5	98	51	39	129	181	5	1053
Arlete Bourguignon	F	Severo	13	82,9	99	51	94	92	186	24	3921
Armando Shonio	M	Leve	13	70,2	146	12	67	170	192	22	3963
Americo Armani	M	Moderado	1	4,8	140	31	36	118	226	23	229
Lourdes Savignon	F	Severo	2	7	84	54	76	144	168	10	406
Vicente Massari	M	Severo	5	26	94	25	77	65	277	16	1402
Isabel Rasmusdatter	F	Moderado	8	51	108	25	84	79	192	24	1669
Gilson Meneguete	M	Severo	3	12,5	130	16	63	101	234	8	811
Caroline Perreyra	F	Leve	11	35	118	27	55	135	226	11	1820
Artur Melegari	M	Leve	12	68	138	40	60	136	195	3	2858
Giovani Albino	M	Moderado	14	52,1	66	42	65	74	231	12	3177
Adolfo Perreyra	M	Moderado	7	27,9	122	55	48	54	159	12	1143
Ricardo Raposo	M	Leve	6	39,6	157	27	85	54	151	26	1981
Thais Binelle	F	Leve	6	21,5	79	31	53	131	217	7	923
Gervasio Breziniscki	M	Leve	4	13,7	125	34	73	81	154	15	643
Adriana Bravin	F	Leve	5	23,8	89	17	50	114	252	13	1022
Alessandra Matsuschita	F	Moderado	8	31,9	134	18	65	151	162	12	1947
Alzira Delaparte	F	Moderado	11	48,8	118	29	70	118	220	20	2806
Giuseppina Braido	F	Moderado	13	49,9	124	13	52	112	261	4	2893
Nilo Alcantra	M	Moderado	10	50,6	125	48	91	119	268	14	2277
Vinicius Sbetti	M	Leve	8	44,1	157	30	75	156	232	4	2028
Renata Colodete	F	Severo	8	30,6	87	40	47	101	186	20	1704
Amalia Cavazzan	F	Severo	5	19,4	157	38	93	115	186	11	1026
Aldo Mortensen	M	Leve	13	59,2	148	50	51	56	130	15	2721
Barbara Milanezi	F	Leve	9	46,3	64	31	83	145	268	18	2404
Gaspar Pianassole	M	Moderado	1	4,6	63	45	42	130	201	19	189
Gabriela Arrivabene	F	Moderado	6	35,1	95	54	85	148	130	20	2149
Henriqueta Demartins	F	Moderado	6	23,8	124	43	92	107	141	10	1355
Santina Caliman	F	Moderado	7	50,8	69	43	51	207	213	29	2519
Marcus Siqueira	M	Moderado	3	15,4	92	26	69	130	192	13	983
Alice Loredetto	F	Moderado	12	56,2	156	44	67	69	259	0	2920
Severino Zampiroli	M	Leve	15	69,7	114	46	87	155	166	25	0
Nilton Fiorot	M	Moderado	13	62,6	63	45	76	104	264	16	3069
Gertrudes Felete	F	Leve	14	79,4	90	51	48	90	279	6	4127
Eliana Delpupo	F	Leve	12	54,7	61	19	74	198	162	25	2832
Carlo Molinaroli	M	Moderado	13	62,7	65	49	90	156	132	28	3271
Bernardo Cenedesi	M	Severo	4	15,1	130	42	52	137	186	9	711
Lucila Minetti	F	Moderado	3	11,2	61	22	42	119	265	17	572
Orlando Laures	M	Moderado	2	8,5	141	51	46	88	226	28	371
Nelson Olausen	M	Moderado	3	12,7	62	33	87	137	240	6	572
Anne Langrin	F	Leve	2	7,1	158	41	52	68	246	20	262
Lidia Zanchetin	F	Moderado	14	52,8	98	52	57	192	144	28	2156
Fioravante Merotto	M	Leve	7	35,4	82	55	90	93	253	0	0
Natalina Donati	F	Severo	14	59,5	144	20	44	94	279	12	2321
Antonia Battistella	F	Moderado	5	17,6	155	30	68	134	142	11	898
Raphael Laurenco	M	Leve	8	44,3	107	47	40	195	189	24	1813
Helena Merotto	F	Moderado	2	6,9	94	39	78	105	267	18	329
Daniel Vidotti	M	Moderado	11	60	157	17	93	138	279	7	2580
Ida Colodetti	F	Moderado	1	5,8	93	47	52	145	277	24	217
Marilene Coutinho	F	Moderado	5	20,7	124	44	38	84	250	10	869
Rodolfo Gobbato	M	Moderado	4	23	61	42	45	63	151	5	1198
Caterina Barcelos	F	Leve	15	81,3	127	29	64	111	208	27	4719
Ignacia Donadello	F	Moderado	14	74,4	154	51	45	122	262	11	4166
Madalena Corocher	F	Moderado	5	25,2	157	22	59	144	135	24	1373
Marcelino Pazini	M	Leve	9	33,8	77	53	70	94	211	19	1660
Eduarda Baiense	F	Leve	6	22,9	111	41	48	76	199	17	1322
Bruna Bonilha	F	Leve	14	53,8	76	13	70	192	211	19	3101
Reinaldo Pianissoli	M	Moderado	8	47,1	117	32	45	113	195	10	0
Catarina Manoel	F	Leve	2	9,1	84	59	58	140	147	9	555

Fernanda Nielsdatter	F	Severo	10	40,7	134	58	91	126	273	21	2188
Johan Coutinho	M	Moderado	4	20,5	141	27	75	93	133	15	821
Silvia Bullock	F	Moderado	4	20,9	148	20	85	75	154	15	960
Adelaide Monente	F	Moderado	1	6,2	66	38	37	129	198	27	264
Eurico Mamoni	M	Leve	11	48,8	138	56	78	113	222	19	1978
Elizabeth Minetti	F	Moderado	4	20,2	118	12	80	61	205	13	1290
Elizangela Morosini	F	Moderado	6	30,2	65	38	68	139	273	12	1810
Regina Cittera	F	Leve	12	72	132	40	57	194	147	30	2255
Inez Lucchetta	F	Leve	2	6,4	87	17	63	156	172	0	0
Pierina Ceschel	F	Leve	4	23,4	120	19	66	52	183	12	983
Luiza Scotte	M	Moderado	3	13	63	12	66	75	163	8	467
Daniela Paresque	F	Moderado	6	33,1	83	34	69	74	250	10	1587
Theresa Darroz	F	Severo	7	35,2	136	21	43	141	190	28	1224
Joel Nardini	M	Severo	12	44,3	62	47	76	129	154	7	1815
Leontina Martas	F	Moderado	9	31,2	81	20	48	138	243	19	1646
Mara Avelar	F	Leve	13	75,4	156	52	36	156	279	11	2941
Elza Borghi	F	Leve	7	32	133	55	82	155	168	0	1662
Marlene Wandermurem	F	Leve	11	56	130	59	69	196	213	4	2464
Silvio Lougun	M	Leve	13	73,1	103	16	75	148	264	9	3142
Hans Lazarin	M	Leve	6	35,4	74	38	39	149	189	5	1451
Terezinha Marchioro	F	Severo	1	5	132	51	43	144	204	0	230
Alda Scaramussa	F	Leve	7	26,5	99	15	45	142	267	22	1224
Sandra DAltoè	F	Leve	11	54	141	31	60	108	150	11	2646
Vitoria Delarmelina	F	Moderado	15	63,9	92	53	49	117	255	25	2788
Hugo Schbert	M	Leve	3	15,6	151	33	90	153	135	16	842
Marie Belinato	F	Leve	14	48,4	109	14	73	56	166	13	1789
Joam Especemile	M	Moderado	7	42,8	141	12	44	136	154	5	2399
Renato Xavier	M	Moderado	8	49,4	95	52	71	85	217	12	2766
Manoel Tagliaferro	M	Leve	3	9,8	157	56	54	87	214	5	561
Edmar Heckert	M	Moderado	1	4,3	105	12	38	127	276	3	241
Lauro Monteiro	M	Severo	3	12,3	130	30	90	125	159	26	718
Zilda Amorim	F	Moderado	11	62,7	139	23	51	67	258	23	3192
David Lagoeiro	M	Leve	1	7	73	55	47	60	171	25	243
Catterina Fioresi	F	Leve	12	41,2	92	38	64	50	166	19	2216
Ana Ferraco	F	Moderado	13	55,7	76	12	91	111	172	0	2840
Cleonice Martinsen	F	Moderado	13	57,4	86	43	94	54	196	28	2294
Nilson Boldrini	M	Leve	4	24,4	80	18	78	72	270	20	1475
Samuel Ladeira	M	Severo	10	44,4	151	54	73	150	258	29	1662
Eva Braggion	F	Moderado	3	12,5	106	44	56	82	175	3	0
Dionisio Busato	M	Moderado	15	50,8	87	40	86	76	268	17	2149
Lucia Caliari	F	Leve	9	56	135	15	51	79	216	29	3118
Albino Pinheiro	M	Moderado	14	57,9	155	25	66	56	211	20	3110
Clara Malini	F	Moderado	7	38,1	60	60	54	52	241	6	1447
Abel Cozumeco	M	Leve	11	56	151	58	48	101	135	0	0
Carmen Benevides	F	Moderado	11	53	143	52	87	56	171	20	1850
Edson Nassar	M	Moderado	8	36,5	119	21	89	65	184	15	1423
Miguel Bianco	M	Leve	7	33,9	156	27	39	130	190	18	1893
Silvestre Rodriques	M	Severo	8	40,3	140	47	70	52	175	14	2135
Osmar Balarini	M	Moderado	10	47,7	76	57	45	192	178	22	2429
Iolanda Cavallina	F	Moderado	12	71,2	109	40	55	156	210	21	4019
Martha Mascarelo	F	Leve	9	45,4	112	51	79	67	252	10	2812
Manuela Spigolon	F	Moderado	1	6,3	108	52	49	146	208	19	266
Julia Toneto	F	Leve	6	26,7	157	18	91	138	223	13	1361
Braz Bortoloti	M	Moderado	5	16,4	146	16	65	67	138	18	808
Zuanne Thomaz	F	Moderado	10	57,7	120	43	91	126	234	22	1974
Rosalina Cosseti	F	Leve	5	19,8	122	14	47	157	133	9	1267
Afonso Aveleda	M	Moderado	9	53,8	65	22	78	100	241	6	2957
Luigi Carneiro	M	Moderado	4	14	152	23	70	81	198	8	632
Marianna Fittipaldi	F	Leve	9	41,1	150	38	69	95	133	17	1816
Agostinho Vieira	M	Leve	9	51,6	88	12	80	86	235	20	2485
Francisca Lovatel	F	Moderado	3	14,9	153	25	38	145	199	0	670

Fonte: Resultados originais da pesquisa

(conclusão)

Tabela 2. Códigos R

```
library(tidyverse) #pacote para manipulacao de dados
library(cluster) #algoritmo de cluster
library(dendextend) #compara dendogramas
library(factoextra) #algoritmo de cluster e visualizacao
library(fpc) #algoritmo de cluster e visualizacao
library(gridExtra) #para a funcao grid arrange
library(readxl)
library (reshape)
library(dplyr)

DT_PacientesOriginal <- read.table("dados/_DatasetPacientes3.csv", sep = ";", header = T, dec = ",")

rownames(DT_PacientesOriginal) <- DT_PacientesOriginal[,1]
DT_PacientesOriginalTodasVariaveis <- DT_PacientesOriginal
DT_PacientesOriginal <- DT_PacientesOriginal[, -c(1:3)]
DT_PacientesPadronizado <- scale(DT_PacientesOriginal)
d <- dist(DT_PacientesPadronizado, method = "euclidean")

fviz_nbclust(DT_PacientesPadronizado, FUN = hcut, method = "wss")

hc1 <- hclust(d, method = "single" )
hc2 <- hclust(d, method = "complete" )
hc3 <- hclust(d, method = "average" )
hc4 <- hclust(d, method = "ward.D" )

plot(hc2, cex = 0.6, hang = -1)
rect.hclust(cluster.hierarquico, k = 4)

hc2_4grupos <- cutree(hc2, k = 4)
table(hc2_4grupos)

Pacientes_Grupos <- data.frame(hc2_4grupos)
HC_DatasetPacientes_fim <- cbind(DT_PacientesOriginalTodasVariaveis, Pacientes_Grupos)

names(HC_DatasetPacientes_fim)[names(HC_DatasetPacientes_fim)=="hc2_4grupos"] <- "Grupo"

MediaGrupo_Pacientes <- HC_DatasetPacientes_fim %>%
  group_by(Grupo) %>%
  summarise(n = n(),
            Idade_Anos = mean(Idade_Anos),
            Peso_kg = mean(Peso_kg),
            Glicemia_mgdl = mean(Glicemia_mgdl),
            Prolactina_ngmL = mean(Prolactina_ngmL),
            HDLb_mgdl = mean(HDLb_mgdl),
            LDLr_mgdl = mean(LDLr_mgdl),
            Triglicérides_mgdl = mean(Triglicérides_mgdl),
            DoseDiáriaRisperidona_mg = mean(DoseDiáriaRisperidona_mg),
            DoseDiáriaValproato_mg = mean(DoseDiáriaValproato_mg))

#DESVIO PADRÃO
DesvioPadraoGrupo_Pacientes <- HC_DatasetPacientes_fim %>%
  group_by(Grupo) %>%
  summarise(n = n(),
            Idade_Anos = sd(Idade_Anos),
            Peso_kg = sd(Peso_kg),
            Glicemia_mgdl = sd(Glicemia_mgdl),
            Prolactina_ngmL = sd(Prolactina_ngmL),
            HDLb_mgdl = sd(HDLb_mgdl),
            LDLr_mgdl = sd(LDLr_mgdl),
            Triglicérides_mgdl = sd(Triglicérides_mgdl),
            DoseDiáriaRisperidona_mg = sd(DoseDiáriaRisperidona_mg),
            DoseDiáriaValproato_mg = sd(DoseDiáriaValproato_mg))


Grupo1<-subset(HC_DatasetPacientes_fim, Grupo=='1')
write.csv2(Grupo1,"dados/_Grupo1.csv", row.names = TRUE, fileEncoding = "UTF-8")

#Agora vamos rodar de 3 a 6 centros  e visualizar qual a melhor divisao
nhc2 <- kmeans(DT_PacientesPadronizado, centers = 2)
```

```
nhc3 <- kmeans(DT_PacientesPadronizado, centers = 3)
nhc4 <- kmeans(DT_PacientesPadronizado, centers = 4)

G2 <- fviz_cluster(nhc2, geom = "point", data = DT_PacientesPadronizado) + ggtitle("k = 4")
G3 <- fviz_cluster(nhc3, geom = "point", data = DT_PacientesPadronizado) + ggtitle("k = 5")
G4 <- fviz_cluster(nhc4, geom = "point", data = DT_PacientesPadronizado) + ggtitle("k = 6")
grid.arrange(G2, G3, G4, nrow = 2)

NH_Group2 <- data.frame(nhc2$cluster)

NHC_DatasetPacientes_fim <- cbind(DT_PacientesOriginalTodasVariaveis, NH_Group2)
names(NHC_DatasetPacientes_fim)[names(NHC_DatasetPacientes_fim)=="nhc2.cluster"] <- "Grupo"
```

Fonte: Resultados originais da pesquisa