NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Alterações da Glicemia:
Uma análise de *Clusters*

Ana Lúcia de Carvalho Frade Pina

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ALTERAÇÕES DA GLICEMIA: UMA ANÁLISE DE *CLUSTERS*

por

Ana Frade Pina

Proposta de Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Gestão de Informação com especialização em Business Intelligence e Gestão do Conhecimento

Orientador: Professor Doutor Roberto Henriques

Coorientador: Professora Doutora Maria Paula Macedo

Outubro 2019

DEDICATÓRIA

Tempo de Travessia

"Há um tempo em que é preciso
abandonar as roupas usadas
Que já tem a forma do nosso corpo
E esquecer os nossos caminhos que
nos levam sempre aos mesmos lugares
É o tempo da travessia
E se não ousarmos fazê-la
Teremos ficado para sempre
À margem de nós mesmos"

(Fernando Pessoa)

Aos meus pais, por me terem ensinado que o nosso caminho é aquele que nos preenche.

Ao Pedro, pelo apoio incondicional na minha *Travessia.*

AGRADECIMENTOS

A possibilidade de analisar grandes quantidades de dados, de elevada complexidade, tem levado a abordagem personalizada dos indivíduos, transversalmente em vários campos da nossa sociedade. A medicina não é exceção, e assistimos hoje a uma nova mudança de paradigma. Passamos da medicina baseada na evidência à medicina de precisão.

Considero um privilégio assistir a esta época, em que se vislumbra a resposta a problemas de saúde e de evolução complexos e prementes, e que tanto impacto têm na nossa sociedade. É o caso da doença metabólica, cujo estudo tem sido o meu principal interesse nos últimos anos. Assim, ser-me-ia difícil deixar escapar a oportunidade de me formar no sentido de poder contribuir para a busca de tais respostas. Encontrar tais respostas é uma aspiração ou um objetivo que irei perseguir. O fim de um caminho que agora início com este trabalho.

Este trabalho está longe de ser meu. Este trabalho é de todas as pessoas envolvidas no estudo PREVADIAB 2 a quem reconhecidamente agradeço na pessoa do Professor João Filipe Raposo.

Agradeço também ao Professor Roberto Henriques, meu orientador, pelo apoio e paciência incansável, pela liberdade e confiança.

À Professora Maria Paula Macedo, também orientadora neste trabalho, para além de todo o seu envolvimento e trabalho aqui investidos, agradeço-lhe a inspiração, por fomentar a liberdade de ideias, pela disponibilidade constante para as discutir sem preconceito, por liderar pelo exemplo, e por buscar sempre as qualidades de com quem colabora.

RESUMO

A diabetes tipo 2 é considerada a epidemia do século XXI. Os valores de corte de glicemia, utilizados no diagnóstico da diabetes tipo 2 e da pré-diabetes, são estabelecidos por convenção. No caso da pré-diabetes, não existe acordo entre as diferentes sociedades (Organização Mundial de Saúde e American Diabetes Association), no que respeita ao valor de corte da glicemia em jejum e da HbA1c. Também na classe da pré-diabetes, existem indivíduos que já apresentam complicações da diabetes, outros irão progredir, e ainda aqueles que nunca progredirão na doença. Assim, a classificação baseada apenas nos valores de glicemia parece ser insuficiente, não só para o diagnóstico, mas também para identificar, o risco de progressão de cada indivíduo. A diabetes tipo 2 tem etiologia multifatorial, e complexa. Com base nos presentes critérios, podemos estar a agrupar indivíduos com diferentes fenótipos sob o mesmo grupo de diagnóstico. A abordagem igual de fenótipos diferentes pode contribuir para a ineficácia da prevenção, do diagnóstico e da terapêutica, o que se traduz no *fardo* socioeconómico que a diabetes tipo2 representa. A análise de *clusters* tem como objetivo pôr em evidência grupos naturais numa determinada população. Permitindo a análise de dados complexos, revela padrões de características que definem grupos diferentes. Em particular, os *Self-organizing Maps* (*SOM*), são uma metodologia robusta de análise de clusters, que permitem a redução dos dados a uma grelha de 2 dimensões, conservando a topologia dos mesmos. Este trabalho teve como objetivo utilizar a análise de *clusters,* nomeadamente o *SOM,* para revelar grupos, que representem diferentes fenótipos da doença metabólica e que possam ser úteis na compreensão dos mecanismos fisiopatológicos, e na melhoria da prevenção, diagnóstico e tratamento destes doentes.

Aplicámos, em primeiro lugar, um algoritmo de *SOM* a 1010 indivíduos da coorte Prevadiab2, reduzindo assim a dimensionalidade dos dados. Para este algoritmo utilizamos parâmetros (27) antropométricos e bioquímicos, reconhecidamente importantes na fisiopatologia da diabetes tipo 2. De seguida, com base num cluster hierárquico (método Ward), definimos os *clusters* finais.

Identificámos 10 *clusters*, com diferentes perfis antropométricos e metabólicos. Todos os *clusters* apresentam indivíduos com normoglicemia e com hiperglicemia (pré-diabetes e/ou diabetes) em diferentes proporções. Nos 5 *clusters* que contêm pessoas com diabetes, encontramos também indivíduos com pré-diabetes, e surpreendentemente, encontramos ainda pessoas com normoglicémia, embora os últimos estejam presentes em menor proporção.

A aplicação do *SOM*, a uma população, que inclui pessoas com normoglicemia e hiperglicemia, permitiu identificar grupos com diferentes fenótipos antropométricos e metabólicos. Mais relevante, os resultados obtidos levantam várias questões relevantes relativas aos mecanismos fisiopatológicos subjacentes aos diferentes fenótipos. A resposta a estas questões pode ter um impacto determinante na melhoria da prevenção, do diagnóstico e da eficácia terapêutica da diabetes tipo 2.

PALAVRAS-CHAVE

Cluster; Diabetes tipo 2; Normoglicemia; *Self-Organizing Maps;* Pré-diabetes

ABSTRACT

Type 2 diabetes (T2D) is considered the 21st century epidemic. Glycemic cutoff values, for T2D and prediabetes diagnosis, are established by convention. In the case of prediabetes, there is no agreement between the different societies (World Health Organization and American Diabetes Association) regarding the cutoff value of fasting glycaemia and HbA1c. Also, in the prediabetes class, there are individuals who already show diabetes complications, others will progress, and those who will never progress in the disease. Thus, classification based solely on glycemic values seems to be insufficient, not only for diagnosis, but also to identify the risk of progression of each individual. T2D has a multifactorial and complex etiology. Based on actual criteria, we may be grouping individuals with different phenotypes under the same diagnostic group. Applying equal approach to the different phenotypes may contribute to the inefficacy of prevention, diagnosis and therapy, which translates into the socioeconomic burden that T2D represents. Cluster analysis aims to highlight natural groups in a given population. By allowing the analysis of complex data, it reveals patterns of characteristics that define different groups. In particular, Self-Organizing Maps (SOM) is a robust clustering methodology, that reduces the data into a 2-dimension topological greed. The aim of this study was to do a cluster analysis using a SOM, to reveal groups representing different metabolic disease phenotypes, which may be useful in understanding pathophysiological mechanisms and in improving the prevention, diagnosis and treatment of these subjects.

First, we applied a SOM to 1010 individuals from the Prevadiab2 cohort, reducing the data dimensionality. For this algorithm we used anthropometric and biochemical parameters (27), which are recognized as important in T2D pathophysiology. Then, based on a hierarchical cluster (Ward method), we define the final clusters.

We identified 10 clusters, with different anthropometric and metabolic profiles. All clusters have individuals with normoglycemia and hyperglycemia (prediabetes and/or diabetes) in different proportions. In the 5 clusters that contain people with diabetes, we also find individuals with prediabetes, and surprisingly, even normoglycemic subjects, although the latter are present in a smaller proportion.

The application of SOM to a population, including normoglycemic and hyperglycemic people, allowed the identification of groups with different anthropometric and metabolic phenotypes. More relevant, the results obtained raise several important questions regarding the pathophysiological mechanisms underlying the different phenotypes. The answer to these questions can have a decisive impact in improving the prevention, diagnosis and therapeutic efficacy of T2D subjects.

KEYWORDS

Cluster*;* Normoglycemia; Prediabetes; Self-Organizing Maps*;* Type 2 Diabetes

ÍNDICE

1. Introdução 1
2. Análise de *Clusters* aplicada ao estudo da Hiperglicemia 6
   2.1. Classificação e Fisiopatologia da Hiperglicemia 6
      2.1.1. Critérios de diagnóstico da Diabetes e Hiperglicemia intermédia 6
      2.1.2. Fisiopatologia da Diabetes tipo 2 9
   2.2. Análises de Clusters no estudo da Hiperglicemia 11
   2.3. Análise de Clusters – *Self-Organizing Maps* 13
      2.3.1. Parâmetros do SOM 17
      2.3.2. SOM multicamada ou SuperSOM 18
      2.3.3. Seleção das medidas de avaliação do SOM 19
      2.3.4. *Clustering* dos protótipos 20
   2.4. Hipótese Geral e Objetivos do Estudo 21
3. Materiais e Métodos 22
   3.1. O Estudo Prevadiab2 23
   3.2. Variáveis Clínicas e Inclusão de indivíduos 24
   3.3. SuperSOM do Prevadiab 2 24
      3.3.1. Limpeza dos dados 24
      3.3.2. Transformação dos dados 25
      3.3.3. Seleção e extração de variáveis 25
      3.3.4. Análise estatística sumária dos dados selecionados 28
      3.3.5. Modelação - Algoritmo *SUPERSOM* 28
      3.3.6. *Clustering* dos protótipos 31
4. Resultados e Discussão 32
   4.1. Resultados e Análise da grelha final do superSOM 32
   4.2. Resultados e Análise das unidades da grelha final do superSOM 33
   4.3. Agregação das unidades do SOM 39
   4.4. Análise e Discussão dos *Clusters* 41
      4.4.1. Perfil global dos *Clusters* 43
      4.4.2. Perfil individual dos *Clusters* 49
      4.4.3. Discussão 52
5. Conclusões 56
6. Limitações e Recomendações para Trabalhos Futuros 57
7. Bibliografia 58

8. Anexos 61

ÍNDICE DE FIGURAS

FIGURA 1 - ESQUEMA SIMPLIFICADO DO *SELF-ORGANIZING MAP*. 15
FIGURA 2 - REPRESENTAÇÃO DE UMA MATRIZ – U 16
FIGURA 3. ESQUEMA DA METODOLOGIA UTILIZADA 22
FIGURA 4. CLASSIFICAÇÃO DA POPULAÇÃO PELA OMS/IDF (A) E PELA ADA (B) 28
FIGURA 5. MATRIZ U DO *SUPERSOM* SELECIONADO 32
FIGURA 6. QUALIDADE DO MAPEAMENTO. 32
FIGURA 7. NÚMERO DE INDIVÍDUOS MAPEADOS ÀS DIFERENTES UNIDADES 33
FIGURA 8. *COMPONENT PLANES* (VARIÁVEIS NORMALIZADAS POR GRELHA DO *SUPERSOM*) 34
FIGURA 9. DISTRIBUIÇÃO DAS CLASSES DE HIPERGLICEMIA DA OMS/IDF PELOS DIFERENTES UNIDADES DA GRELHA DO *SUPERSOM*: A) NÚMERO DE INDIVÍDUOS EM CADA GRUPO, CLASSIFICADOS POR CLASSE DA IDF; B) PROPORÇÃO DE INDIVÍDUOS DE CADA CLASSE DA IDF, EM CADA GRUPO. 38
FIGURA 10. DISTRIBUIÇÃO DO IMC (A) E PA (B) PELAS DIFERENTES UNIDADES.. 39
FIGURA 11. CÁLCULO DO NÚMERO ÓTIMO DE *CLUSTERS* PELO ÍNDICE DE DAVIES-BOULDIN NO *SUPERSOM* 40
FIGURA 12. DENDOGRAMA DO *CLUSTER* HIERÁRQUICO DOS PROTÓTIPOS DO *SUPERSOM* (MÉTODO WARD). 40
FIGURA 13. MAPEAMENTO DOS INDIVÍDUOS AO *CLUSTER* CORRESPONDENTE, DE ACORDO COM O PONTO DE CORTE DO C*LUSTER* HIERÁRQUICO. 41
FIGURA 14. DISTRIBUIÇÃO DO GÉNERO PELOS *CLUSTERS* 42
FIGURA 15. DISTRIBUIÇÃO DAS CLASSES DE HIPERGLICEMIA DA OMS/IDF PELOS 10 *CLUSTERS*: A) PROPORÇÃO DE INDIVÍDUOS DE CADA CLASSE EM CADA *CLUSTER*; B) NÚMERO DE INDIVÍDUOS EM CADA *CLUSTER*, DISTRIBUÍDOS SEGUNDO A CLASSIFICAÇÃO DA OMS/IDF. 43
FIGURA 16. *HEATMAP* CONSIDERANDO AS VARIÁVEIS MEDIDAS APENAS EM JEJUM (MEDIANA), UTILIZADAS PARA O PERFIL DOS *CLUSTERS* 44
FIGURA 17. *HEATMAP* CONSIDERANDO OS PARÂMETROS REFERENTES À GLICEMIA (MEDIANA), DOS 10 *CLUSTERS*.. 45
FIGURA 18. *HEATMAP* CONSIDERANDO OS PARÂMETROS MEDIDOS DURANTE A PTGO (MEDIANA), DOS 10 *CLUSTERS*. 45
FIGURA 19. PERFIL DA GLICEMIA DURANTE A PTGO DE CADA *CLUSTER*.. 46
FIGURA 20. PERFIL DE ÁCIDOS GORDOS LIVRES DURANTE A PTGO DE CADA *CLUSTER* 47
FIGURA 21. PERFIL DE *CLEARANCE* DE INSULINA DURANTE A PTGO DE CADA *CLUSTER* 48
FIGURA 22. PERFIL DE NÍVEIS DE PEPTÍDEO C NO SANGUE VENOSO DURANTE A PTGO DE CADA *CLUSTER* 48
FIGURA 23. PERFIL DE NÍVEIS DE INSULINEMIA DURANTE A PTGO DE CADA *CLUSTER* 49

ÍNDICE DE TABELAS

TABELA 1. CRITÉRIOS DE DIAGNÓSTICO DA DIABETES E PRÉ-DIABETES ATUALMENTE ACEITES 7
TABELA 2. PARÂMETROS CLÍNICOS E ANALÍTICOS REGISTADOS NA AVALIAÇÃO DO ESTUDO PREVADIAB 2 23
TABELA 3. VARIÁVEIS NORMALIZADAS INCLUÍDAS NO MODELO 26
TABELA 4. VARIÁVEIS UTILIZADAS NO PERFIL DOS *CLUSTERS* 27
TABELA 5. SUMÁRIO DOS PARÂMETROS DO ALGORITMO *SUPERSOM* 31
TABELA 6. DISTRIBUIÇÃO DA POPULAÇÃO PELOS 10 *CLUSTERS* 42
TABELA 7. RESUMO DOS PERFIS DOS *CLUSTERS*, REFERENTES À RESISTÊNCIA À INSULINA *VERSUS* A FUNÇÃO DA CÉLULA ß 50

LISTA DE SIGLAS E ABREVIATURAS

ADA	*American Diabetes Association*
AGJ	Alteração da glicemia em jejum
ALT	Alanina aminotransferase
AST	Aspartato aminotransferase
AUC	Área sob a curva (*Area under the curve*)
BMU	*Best Matching Unit*
GGT	Gama glutamil-transpeptidase
HbA1c	Hemoglobina glicosilada
HDL	Lipoproteínas de alta densidade (*High density lipoprotein)*
IDB	Índice de Davies-Bouldin
IDF	*International Diabetes Federation*
IMC	Índice de massa corporal
IR	Insulinoresistência ou resistência à insulina
LDL	Lipoproteínas de baixa densidade (*Low density lipoprotein*)
NDDG	National Diabetes Data Group
OMS	Organização Mundial de Saúde
OND	Observatório Nacional da Diabetes
PA	Perímetro abdominal
PTGO	Prova de tolerância à glicose oral
QE	*Quantization error*
SOM	*Self-organizing map*
TA	Tensão arterial
TDG	Tolerância diminuída à glicose
TE	*Topographic error*

1. INTRODUÇÃO

A diabetes corresponde a uma alteração do metabolismo da glicose que cursa com elevados níveis de concentração de glicose no sangue (hiperglicemia). Estas alterações surgem quando existe um desequilíbrio, entre a quantidade de insulina produzida pelo pâncreas (hormona responsável pela internalização da glicose nas células do organismo que são insulinodependentes, para produção de energia), e a capacidade que o organismo tem de utilizar a insulina produzida, ou insulinorresistência(IR) (DeFronzo, 2009), esta última traduzida pela quantidade de insulina necessária para internalizar a glicose nas células insulinodependentes.

Se na diabetes tipo 1, o defeito responsável pela hiperglicemia é uma incapacidade da célula ß para produzir insulina, na diabetes tipo 2, considera-se que as alterações da glicemia derivam de uma deficiência relativa de insulina em relação à insulinorresistência presente (DeFronzo, 2009). Embora existam outros tipos mais raros de diabetes, a diabetes tipo1, a diabetes tipo2 e a diabetes gestacional, constituem os três grupos principais de diabetes (IDF, 2017), sendo a diabetes tipo2 a forma mais frequente (World Health Organization, 2016), correspondendo a cerca de 90-95% dos casos de diabetes (Wu, Ding, Tanaka, & Zhang, 2014).

A alteração dos níveis de glicemia é apontada como principal responsável pelo aparecimento das complicações da diabetes. O impacto da diabetes reside mais nas complicações que provoca ao longo do tempo, que são as principais responsáveis pelos elevados níveis de morbilidade (insuficiência renal, cegueira, amputações, eventos cardiovasculares) e mortalidade nestes indivíduos (Fowler, 2011; Observatório Nacional da Diabetes, 2016).

No contexto da diabetes tipo 2, define-se ainda a pré-diabetes ou hiperglicemia intermédia, na qual, os indivíduos apresentam valores de glicemia em jejum e pós-prandiais inferiores aos considerados como valores de corte para o diagnóstico da diabetes, e se considera terem maior risco de progressão para diabetes, maior risco de eventos cardiovasculares, bem como uma incidência maior de complicações da hiperglicemia, relativamente à população normal (Bansal, 2015). A abordagem da hiperglicemia intermédia constitui uma oportunidade de evitar a progressão para a diabetes, bem como o aparecimento de complicações. Neste grupo podemos considerar: as alterações da glicemia em jejum (AGJ), se apenas os valores de glicemia em jejum estão alterados (no intervalo considerado para a hiperglicemia intermédia), permanecendo normais os valores da glicemia pós-prandial, ou 120’ após uma prova de tolerância à glicose oral (PTGO); a tolerância diminuída à glicose (TDG), se apenas os valores de glicemia pós-prandial ou aos 120’ de uma PTGO estão alterados, com valores de glicemia em jejum normais; ou alterações mistas, se ambos os

valores de glicemia em jejum, e aos 120' de uma PTGO, se encontram no intervalo de valores definidos para a hiperglicemia intermédia.

A diabetes tipo 2 é considerada um importante problema de saúde pública pela Organização Mundial de Saúde (OMS), sendo uma das quatro doenças não comunicáveis considerada como prioritária (World Health Organization, 2016) .

A prevalência da diabetes tipo 2 tem vindo a aumentar em todo o mundo, tendo duplicado desde 1980. Estima-se, que em 2014, existiam 422 milhões de adultos com diabetes, e que esta foi causa direta de cerca de 1,5 milhões de mortes (World Health Organization, 2016). A International Diabetes Federation (IDF) estima que, em 2040, o número de indivíduos com diabetes suba a 642 milhões (Ogurtsova et al., 2017). Em Portugal, cerca de 40,7% da população portuguesa tem hiperglicemia intermédia ou diabetes (3,1 milhões de indivíduos). O Relatório Anual do Observatório Nacional da Diabetes de Portugal (OND) estimou a prevalência da diabetes, em mais de 1 milhão de indivíduos em Portugal em 2015, correspondendo a cerca de 13,3% da população, e a um crescimento da prevalência em cerca de 13,5% desde 2009 (Observatório Nacional da Diabetes, 2016). A diabetes pode ser considerada uma doença silenciosa, pois num elevado número de casos, é diagnosticada após anos de evolução, quando surgem as complicações, como as complicações microvasculares, macrovasculares e neuropáticas. No mesmo relatório, o OND apontou para uma prevalência de 7,5% de diabetes não diagnosticada, o que corresponde a cerca de 44% dos indivíduos com diabetes (Observatório Nacional da Diabetes, 2016). Relativamente à hiperglicemia intermédia estimou-se, também em 2015, uma prevalência de 27,4%, correspondendo a 2,1 milhões de indivíduos (10,4% de AGJ, 14,3% de TDG, e 2,7% de AGJ e TDG). Para além do impacto pessoal e familiar, a diabetes tem um elevado impacto socioeconómico nos países (World Health Organization, 2016; Observatório Nacional da Diabetes, 2016). De acordo com a IDF, em 2015, os gastos com saúde devidos à diabetes, em todo o mundo, foi da ordem dos 673 biliões de US dólares (Ogurtsova et al., 2017). Em Portugal, no ano de 2015, o custo com a diabetes representou 8-10% da despesa em saúde (Observatório Nacional da Diabetes, 2016).

O diagnóstico da diabetes tipo 2 é feito, com base nos níveis de glicemia no sangue, medidos em jejum, pós-prandial e aos 120 minutos após uma PTGO, ou pelos valores de hemoglobina glicosilada (HbA1c). A hemoglobina, proteína presente no sangue, cuja função é a de transporte de oxigénio, é glicosilada aumentando os níveis de HbA1c, quando os níveis de glicemia sobem. Apesar de ter algumas limitações, a HbA1c é considerada como um marcador dos níveis de glicemia a que o indivíduo está exposto.

Os valores de corte para o diagnóstico da diabetes foram estabelecidos por convenção, com base na distribuição dos níveis de glicemia aos 0' e 120' de uma PTGO, e na presença de complicações microvasculares da diabetes, nomeadamente da retinopatia diabética. Os valores de corte para a hiperglicemia intermédia foram também estabelecidos por convenção, no entanto não existe concordância para os valores de glicemia em jejum entre as diferentes sociedades (Cefalu & et al., 2017; International Diabetes Federation, 2017; WHO & IDF, 2006). Independentemente dos valores de corte adotados pelas diferentes organizações, sabemos que existem indivíduos considerados normais que podem apresentar maior risco de progressão para a diabetes, risco de eventos cardiovasculares, e complicações da diabetes (microvasculares, macrovasculares e outras), e que no grupo de indivíduos com diagnóstico de hiperglicemia intermédia existem indivíduos que não progridem na doença, não desenvolvem complicações, nem eventos cardiovasculares (Kumar, Nandhini, Sadishkumar, Sahoo, & Vivekanadan, 2016). As glicemias em jejum, pós-prandiais, e HbA1c parecem ser insuficientes para uma identificação mais precisa destes indivíduos.

A diabetes tipo 2 tem uma etiologia complexa, multifatorial, onde se encontram envolvidos fatores genéticos, de estilos de vida adotados e ambientais. Têm sido identificados vários fatores de risco, como por exemplo a idade, género, o sedentarismo, a dieta, a obesidade, a etnia, os antecedentes familiares, a hipertensão arterial, o tabagismo, entre outros (World Health Organization, 2016; Wu et al., 2014; Cefalu & et al., 2017).

Apesar da fisiopatologia da hiperglicemia ser explicada por um desequilíbrio, entre a produção de insulina da célula ß e a insulinoresistência do indivíduo, esta última muito relacionada com a obesidade, mais especificamente ao índice da massa corporal e perímetro abdominal (IMC e PA)(World Health Organization, 2016), hoje sabe-se que para este desequilíbrio concorrem múltiplos órgãos *(ominous octet)* (DeFronzo, 2009). Assim, quer a deficiente produção de insulina, quer a existência e evolução de insulinorresistência, dependem elas próprias de mecanismos múltiplos, e a verdadeira fisiopatologia da diabetes tipo 2 continua a ser desconhecida. Podemos mesmo considerar que estes fatores e mecanismos, se podem associar de modos diferentes, em qualidade e intensidade, dando origem a diferentes fenótipos, que resultam na hiperglicemia, mas que têm diferentes etiologias e mecanismos.

A abordagem terapêutica dos indivíduos com diabetes e hiperglicemia intermédia inclui, para além das alterações dos estilos de vida (dieta, exercício físico, cessação tabágica), intervenção farmacológica e mais recentemente a cirurgia metabólica, considerada como uma opção terapêutica para indivíduos com IMC > 30Kg/m2 (Cefalu & et al., 2017). A intervenção farmacológica é instituída através de um algoritmo de um modo transversal, e apesar de haver recomendações para que a

seleção dos fármacos seja feita de acordo com algumas características dos doentes (impacto no IMC, efeitos secundários, custo, e preferência do doente) (Cefalu & et al., 2017) , esta está longe de ser uma terapêutica de precisão.

A possibilidade de existirem diferentes fenótipos sob o mesmo diagnóstico, e que são abordados de modo semelhante, do ponto de vista terapêutico, pode explicar, em parte, a ineficácia da terapêutica no controlo da doença e na evicção do aparecimento de complicações, que se traduz na manutenção da incidência da morbilidade e mortalidade destes doentes com implicações socioeconómicas significativas (Cali', Bonadonna, Trombetta, Weiss, & Caprio, 2008; Häring, 2016) .

A evolução da tecnologia e ciências de computação tem permitido, não só o armazenamento de grandes quantidades de dados, mas também a sua exploração. Para tal têm sido desenvolvidos múltiplos algoritmos que permitem a análise de bases de dados de grandes dimensões, quer em quantidade de observações, quer em número de variáveis observadas, com o objetivo de extrair informação relevante, aplicando-a na resolução de problemas (R. Wehrens, 2007). Este facto vem ultrapassar a incapacidade que o ser humano tem em fazer esta análise, permitindo extrair padrões que assim, são descobertos e compreensíveis. Estes algoritmos têm ajudado à resolução de problemas extremamente complexos (Koh & Tan, 2005).

As técnicas de *data-mining* e, em especial, a análise de *clusters* (análise não supervisionada) têm sido utilizadas em vários campos, e também na medicina(Koh & Tan, 2005; Jothi, Aini, Rashid, & Husain, 2015; Țăranu, 2015), com o objetivo de, a partir da análise de dados multidimensionais complexos, revelar grupos com padrões diferentes nas suas características. Espera-se que a análise destes *clusters* ponha em evidência estruturas relevantes, que tenham significado e possam ser úteis na compreensão dessa agregação.

Apesar da variedade de algoritmos de *clustering* existentes, com diferentes implementações, eles têm um objetivo comum. O objetivo é agrupar os indivíduos, de modo a que os mais semelhantes se encontrem no mesmo grupo, e que os diferentes grupos tenham a maior dissemelhança possível.

Neste trabalho propomo-nos a utilizar a análise de *clusters* numa população, em particular o algoritmo superSOM implementado em R (M. R. Wehrens & Kruisselbrink, 2018), aplicado a uma base de dados (Prevadiab2) com elevado número de variáveis (antropométricas e parâmetros analíticos metabólicos), com o objetivo de encontrar grupos e revelar padrões, que traduzam diferentes fenótipos da hiperglicemia. Estes poderão ajudar a esclarecer diferentes mecanismos fisiopatológicos, responsáveis pelas alterações da glicemia, e podem servir de base a uma melhor

classificação destes indivíduos, dando um contributo para a personalização da terapêutica dos mesmos, dentro do novo paradigma da medicina de precisão.

No capítulo que se segue (capítulo 2) faz-se uma breve revisão da literatura da classificação das diferentes classes de hiperglicemia e da sua fisiopatologia, das aplicações do *data-mining* e em particular da análise de *clusters* ao estudo da diabetes, do SOM, especificamente focamos o superSOM, e a sua parametrização, e ainda, dos algoritmos de clustering das unidades geradas por estes algoritmos. Este capítulo termina com a definição da hipótese, bem como dos objetivos do trabalho.

No capítulo 3. descrevemos a população utilizada e a metodologia em que assenta este trabalho, nomeadamente as várias fases do processo de análise de *clusters* e parametrização do algoritmo superSOM em R. No capítulo 4. mostram-se e discutem-se os resultados obtidos. No capítulo 5 retiram-se as conclusões e no capítulo 6 são apontadas, por um lado as limitações ao trabalho e, por outro, futuras linhas de trabalho que derivam dos nossos resultados.

2. ANÁLISE DE *CLUSTERS* APLICADA AO ESTUDO DA HIPERGLICEMIA

As alterações da glicemia surgem como um contínuo, e não como um fenómeno de "tudo ou nada". Deste modo, a identificação de um ponto de corte exato, que separe os indivíduos com normoglicemia, pré-diabetes e diabetes torna-se difícil. Apesar dos critérios de diagnóstico de pré-diabetes e diabetes, terem sido várias vezes revistos, desde a sua primeira publicação pela OMS em 1965, estes continuam a considerar apenas os valores da glicemia, com pontos de corte definidos pela distribuição da glicemia nas populações. Assim, estes critérios não capturam, de todo, o conjunto de alterações metabólicas e outras que possam estar presentes, e que se encontram na base da hiperglicemia. No entanto, o número de moléculas, partículas e órgãos identificados como fazendo parte dos mecanismos fisiopatológicos da diabetes tipo2 cresce diariamente. Dada a multiplicidade de agentes e mecanismos potencialmente envolvidos, definir uma classificação que os consiga refletir é uma tarefa árdua. Alguns trabalhos têm aplicado algoritmos de data-mining à medicina e mais concretamente à diabetes. Recentemente foram publicados alguns trabalhos, que aplicam algoritmos de *clustering*, nomeadamente *k-means* e *Self-Organizing maps (SOM)*, a diferentes vertentes do estudo da diabetes tipo 1 e tipo 2. Neste contexto, a análise de *clusters* surge como uma potencial e atrativa solução, na exploração das causas e mecanismos da hiperglicemia, no sentido de apoiar a definição de uma classificação mais completa e clinicamente aplicável, em particular o (*SOM)*, pelas suas características, à frente revistas.

2.1. CLASSIFICAÇÃO E FISIOPATOLOGIA DA HIPERGLICEMIA

2.1.1. Critérios de diagnóstico da diabetes e hiperglicemia intermédia

Os critérios de diagnóstico da diabetes tipo 2 e da hiperglicemia intermédia baseiam-se em valores de glicemia, medidos em jejum, no pós-prandial, às 2h de uma PTGO, e pela HbA1c. A Tabela 1 apresenta os critérios de diagnóstico atualmente definidos pela OMS/IDF e pela ADA. Se estas entidades se encontram de acordo quanto aos valores de corte de diagnóstico da diabetes tipo2, o mesmo não acontece com os valores da hiperglicemia intermédia (Kumar et al., 2016; Cefalu & et al., 2017).

Diabetes tipo 2

Em 1965 a OMS fez as primeiras recomendações quanto aos critérios de diagnóstico da diabetes tipo2, na tentativa de os uniformizar, com base em valores de glicemia às 2h de uma PTGO (World Health Organization, 1965). Desde então os valores de corte modificaram-se, e outros valores

laboratoriais foram adicionados. Em 1979, o *National Diabetes Data Group* (NDDG), nos EUA, faz uma revisão dos critérios com base na distribuição bimodal das glicemias às 2h de uma PTGO, de várias populações. Verificou-se que um grupo se distribuía abaixo de 200mg/dl, enquanto outro tinha uma distribuição acima 240mg/dl e que a retinopatia diabética, embora não exclusiva, tinha uma prevalência superior no último grupo. Foi também proposto um valor de corte para a glicemia medida em jejum, dada uma distribuição bimodal semelhante, com um valor de corte de 140 mg/dl (National Diabetes Data Group, 1979). Em 1980, a OMS considera o valor da glicemia em jejum de 145mg/dl para o diagnóstico da diabetes e de 200mg/dl para a glicemia às 2h de uma PTGO (World Health Organization, 1980), e em 1985 baixam o valor de corte da glicemia em jejum para 140mg/dl (World Health Organization, 1985).

Tabela 1. Critérios de diagnóstico da diabetes e pré-diabetes atualmente aceites (OMS/IDF e ADA)

	Glicemia em jejum mg/dl (mmol/L)			Glicemia às 2h da PTGO mg/dl (mmol/L)			HbA1c %	
	IDF/OMS	ADA		IDF/OMS	ADA		IDF/OMS	ADA
Diabetes	>=126 (7)	>=126 (7)	ou	>=200 (11,1)	>=200 (11,1)	ou	>= 6,5	>=6,5
Pré-diabetes AGJ isolada	110-125 (6,1-6,9)	100-125 (5,5-6,9)	e	<140 (7,8)	<140 (7,8)	ou	----------	5,7 – 6,4
Pré-diabetes TDG isolada	<110 (<6,1)	<100 (<5,5)	e	140-199 (7,8-11,0)	140-199 (7,8-11,0)	ou		
Pré-diabetes AGJ e TDG	110-125 (6,1-6,9)	100-125 (5,5-6,9)	e	140-199 (7,8-11,0)	140-199 (7,8-11,0)	ou		

Em 1997 a ADA baixa o limiar da glicemia em jejum para 110 mg/dl, critério que a OMS adota em 1999 (The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, 1997). Estes são os valores de corte da glicemia em jejum e da glicemia às 2h de uma PTGO atualmente aceites, pelas principais sociedades.

Apesar de todas as limitações da HbA1c, desde 2010 que esta é reconhecida nos critérios atuais de diagnóstico da diabetes tipo 2 (International Expert Committee, 2009), sendo atualmente reconhecida por todas as organizações.

Pré-diabetes ou hiperglicemia intermédia.

No caso da hiperglicemia intermédia, a definição dos valores de corte não é tão clara, pelo que tem gerado controvérsia e desacordo entre as organizações (Kumar et al., 2016; Cefalu & et al., 2017).

Podemos considerar a hiperglicemia intermédia como um estado em que já há alterações do metabolismo da glicemia, com valores superiores ao normal, mas inferiores aos diagnósticos da diabetes. Este estado representa um risco elevado de progressão para diabetes tipo 2 e doença cardiovascular, e já se encontra associada a formas precoces de complicações microvasculares (retinopatia diabética, nefropatia, neuropatia), (Guariguata et al., 2014). O desacordo entre as organizações e a dificuldade em encontrar um valor de corte ótimo, parece residir no facto de que, quer o risco de progressão para diabetes, quer o risco de aparecimento de doença macro ou microvascular não parecerem associar-se a um limiar de glicemia, mas aumentarem de modo contínuo com o aumento dos valores da glicemia (Tabák & et al., 2012), já existindo, embora de modo muito mais fruste, em pessoas consideradas normais, mesmo tendo por base os critérios mais estritos (Kumar et al., 2016).

Em 1965, a OMS, embora não reconheça a hiperglicemia intermédia tal como considerada hoje, descreve um grupo de pessoas a que chama *borderline*, com valores de glicemia às 2h de uma PTGO no limiar do normal, e alerta para o aumento de risco neste grupo (World Health Organization, 1965).

Em 1979, o NDDG (EUA) define o grupo – *Impaired Glucose Tolerance*. Considera que estes indivíduos, com glicemias às 2h de uma PTGO entre os valores considerados normais e de diabetes tipo 2, têm um risco aumentado de progressão para a diabetes tipo 2 (1-5% por ano), para além de um aumento de prevalência de doença aterosclerótica, de alterações eletrocardiográficas e de morte (National Diabetes Data Group, 1979).

Em 1980 a OMS considera o grupo *Impaired Glucose Tolerance,* com base nas glicemias às 2h de uma PTGO, desde que a glicemia em jejum fosse normal, e em 1985 mantém este grupo, revendo no entanto os seus valores de corte (World Health Organization, 1985).

Em 1997 a ADA considera o grupo dos AGJ, com glicemias em jejum entre 110-126 mg/dl e com glicemias às 2h da PTGO inferiores as 200 mg/dl (The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, 1997). A OMS adota os mesmos critérios em 1999.

Em 2003 a ADA volta a baixar o limiar da glicemia em jejum para 100mg/dl para o diagnóstico de AGJ, mas desta vez a OMS e a IDF mantêm o critério anterior, que ainda vigora atualmente. Outro ponto de não concordância é a HbA1c. Este parâmetro é utilizado nos critérios da ADA, como diagnóstico da hiperglicemia intermédia, o mesmo não acontecendo nos critérios da IDF e da OMS (Kumar et al., 2016; Cefalu & et al., 2017).

Esta discrepância de olhares, entre os valores de hiperglicemia intermedia, não é apenas teórica, tendo importantes consequências práticas. Por um lado, leva ao aumento do número de indivíduos diagnosticados com pré-diabetes nos países que adotam os critérios da ADA, contribuindo ainda mais para a já pesada questão de saúde pública. Por outro lado, faz com que, nestes países, se tratem indivíduos com pré-diabetes, que não seriam tratados, se residissem em países que adotam os critérios da OMS/IDF, como é o caso de Portugal, e vice-versa.

2.1.2. Fisiopatologia da diabetes tipo 2

A intolerância à glicose, num olhar simplista, deriva de um desequilíbrio entre a insulinoresistência e a produção de insulina pelo pâncreas. Na história natural da doença, Tabák & et al. identificam várias fases. Na primeira fase, a insulinorresistência é compensada pela maior produção de insulina. Na segunda fase, as células ß começam a não conseguir compensar a IR, com uma subida ligeira nos valores de glicemia em jejum e pós-prandial. Esta duas fases existem, fundamentalmente, antes do aparecimento da hiperglicemia intermédia. A fase seguinte corresponde a um período de descompensação, em que a insuficiência da célula ß em compensar a insulinorresistência é mais marcada, com elevação rápida dos níveis de glicemia. Surge então a hiperglicemia intermédia e a diabetes (Tabák & et al., 2017). Apesar de a insulinoresistência, hepática e muscular, estarem presentes precocemente na história natural da doença, a doença não surge sem uma falência, relativa ou absoluta, da célula ß. Assim, o declínio na tolerância à glicose associa-se a uma marcada diminuição na secreção de insulina, sem melhoria da sensibilidade à insulina. Embora a IR e a falência da célula ß possam dar contributos diferentes ao desenvolvimento de diabetes tipo 2, o início e progressão da diabetes é determinada pela taxa de progressão da falência da célula ß (DeFronzo, 2009).

A insulinorresistência hepática manifesta-se por um excesso de produção de glicose no fígado em jejum, apesar de poder coexistir com hiperinsulinemia, e por uma deficiente inibição da produção da glicose após a refeição. No músculo, a IR traduz-se numa incapacidade para captar a glicose da circulação após uma refeição (Tabák & et al., 2017). Para além das diferenças nos valores

de glicemia em jejum e pós-prandial, os indivíduos com AGJ e TDG isoladas têm diferenças na curva da glicemia durante uma PTGO. Ainda que ambos apresentem um estado de IR elevada, nos AGJ a IR é tipicamente hepática, com valores de resistência no músculo quase normal, e os TDG apresentam maior IR a nível do músculo, com pequenas elevações na insulinorresistência hepática (Cali' et al., 2008; Tabák & et al., 2017).

A falência da célula ß não deve ser avaliada pela quantidade de insulina que secreta em termos absolutos, mas pela insulina que secreta relativamente a uma variação de glicemia, ajustada à insulinorresistência. Esta falência surge precocemente no desenvolvimento da diabetes (Tabák & et al., 2017). Múltiplos fatores concorrem para o declínio da função da célula beta, nomeadamente idade, genética, insulinorresistência, diminuição da secreção e resistência às incretinas. Estima-se que, na pré-diabetes, os indivíduos perderam já cerca de 50-80% da função da célula beta (DeFronzo, 2009).

Mais recentemente foram considerados outros atores na fisiopatologia da diabetes tipo 2, nomeadamente: o adipócito, através da resistência à insulina nestas células, mantendo-se a lipólise com consequente lipotoxicidade; o trato gastrointestinal, pela secreção de incretinas, estando estas diminuídas ou existindo uma resistência à ação das mesmas na diabetes tipo 2; a célula alfa do pâncreas que produz glucagina, cujos níveis se correlacionam com a produção de glicose hepática; o rim, estando a sua capacidade de absorção de glicose aumentada nos indivíduos com diabetes; o cérebro, controlando através do hipotálamo o metabolismo periférico, também parece tornar-se resistente à insulina (DeFronzo, 2009).

Considerando a multiplicidade e a complexidade de critérios de diagnóstico, mecanismos, e de possíveis combinações entre os mesmos, não é de admirar, que a compreensão atual de todos estes processos seja limitada. A hiperglicemia tem sido estudada, principalmente, através de metodologias estatísticas clássicas, e tem-se de algum modo generalizado as observações à globalidade de indivíduos. No entanto, o crescente aumento da prevalência da diabetes, e a ineficácia da aplicação transversal das alternativas terapêuticas, refletem a necessidade e urgência de compreender melhor esta condição. As metodologias de *data-mining* surgem como uma promissora opção, capaz de começar a deslindar este puzzle, e explorar esta condição numa perspetiva mais individualizada e simultaneamente mais integrada.

2.2. ANÁLISES DE CLUSTERS NO ESTUDO DA HIPERGLICEMIA

O sector da saúde gera quantidades enormes de dados, incluindo registos clínicos eletrónicos, relatórios clínicos, dados administrativos, entre outros. Estas complexas bases de dados estão ainda subutilizadas, apesar da crescente aplicação de técnicas de data-mining. Neste sector, as técnicas de data-mining podem ser aplicadas numa perspetiva de gestão administrativa, de gestão clínica, e ainda de investigação clínica. O estudo e exploração destas bases é um desafio, pois tratam-se de dados fundamentalmente não estruturados, complexos e diversos na qualidade e forma (Nithya, Duraiswamy, & Gomathy, 2013). Por outro lado, existem dados incompletos, erróneos e em falta, que podem levar a conclusões erradas (Househ & Aldosari, 2017),(Koh & Tan, 2005). No entanto, estas técnicas são fundamentais e promissoras (Koh & Tan, 2005), principalmente no contexto do novo paradigma da medicina de precisão (Collins & Varmus, 2015).

Os modelos preditivos (supervisionados) de *data-mining* têm sido os mais frequentemente aplicados à clínica, quer na predição da doença, quer assistindo na decisão clínica (Jothi et al., 2015). No entanto, os modelos de descrição e visualização podem ser muito úteis na compreensão dos dados e na descoberta de padrões, principalmente quando são complexos e contêm interações não lineares (Koh & Tan, 2005). A análise de *clusters*, mais raramente aplicada na clínica, tem sido utilizada, por exemplo, para estudar a readmissão de doentes em unidades de cuidados intensivos (Veloso et al., 2014), o tempo de internamento (Belciug, 2009), dados genéticos (Chipman & Tibshirani, 2006), deteção de recorrência do cancro da mama (Nithya et al., 2013), e encontrar diferentes fenótipos de algumas patologias, como por exemplo da asma brônquica (Schatz et al., 2014).

Os métodos de data-mining têm também sido utilizados na investigação clínica e básica na área da diabetes. Em 2011, Marinov *et al* fez uma revisão sistemática acerca deste assunto (Marinov, Mosa, Yoo, & Boren, 2011), na qual selecionaram 17 artigos, de 1999 a 2010 e agruparam-nos por tópico de investigação: interpretação e predição dos níveis de glicemia, seleção de características, análise genómica e outros. Nesta altura, apesar dos algoritmos de classificação serem os mais utilizados, existiam já alguns trabalhos de *clustering*. Estes foram principalmente relacionados com a identificação de genes comuns à diabetes tipo 2 e a outras doenças, como é o caso da periodontite e da sinusite.

No estudo de doentes com diabetes tipo 1, Tirunagari S. utilizou um algoritmo de Self-Organizing Maps, para identificar grupos de indivíduos com perfis semelhantes de comportamento de auto-cuidado ou estilos de vida, nomeadamente na gestão dos níveis de glicemia, toma de insulina, alimentação e exercício físico. A identificação destes grupos, permite fazer um ajuste mais

personalizado, de acordo com os comportamentos que têm em cada grupo (Tirunagari, Poh, Hu, & Windridge, 2015). Com o objetivo de criar uma nova classificação fenotípica de doentes com diabetes tipo1, e encontrar marcadores que diferenciem o fenótipo, através da avaliação genética dos indivíduos, Topila I *et al* treinaram um Self-Organizing Map multicamada (superSOM), reduzindo assim a dimensionalidade dos dados, e seguidamente fizeram um *cluster* das unidades com um algoritmo hierárquico (Toppila, 2016). Para além de terem identificados grupos com diferente risco de desenvolver complicações na diabetes tipo 1, este trabalho estabelece também uma estrutura metodológica, interessante, que pode ser aplicada no estudo de outras patologias.

Outro tipo de abordagem foi utilizada por Li L *et al,* que fazendo uso de um algoritmo de *clustering*, com base em análise topológica de redes, identificaram 3 subgrupos de doentes com diabetes tipo 2: subtipo 1, caracterizado por indivíduos com diabetes tipo2 com retinopatia e nefropatia diabética; o subgrupo 2, com indivíduos com elevada incidência de cancro e doença cardiovascular; e o subgrupo 3, mais associado a doenças cardiovasculares, doenças neurológicas, alergias e a infeção por HIV. No caso da pré-diabetes, e fazendo também uso de algoritmos de *clustering,* o grupo de Kim utilizou esta abordagem para encontrar grupos de indivíduos, com diferentes níveis de risco de progredir para diabetes tipo 2 (Kim et al., 2014). No entanto, neste trabalho são classificados com hiperglicemia intermédia apenas os indivíduos com alterações da glicemia em jejum, sendo os indivíduos com potencial tolerância diminuída à glicose, incluídos entre a população definida como tendo normoglicemia, já que utilizam apenas a glicemia em jejum para esta classificação.

Mais recentemente, Ahlqvist E. *et al* utilizaram o k-means e o cluster hierárquico para definir uma nova classificação para a diabetes tipo 2. Utilizaram um coorte de indivíduos com diabetes diagnosticada de novo, com base na presença de anticorpos-GAD, idade, IMC HbA1c, HOMA-IR e HOMA-B encontraram 5 *clusters*: cluster 1, Diabetes Autoimune Severa (*SAID*), caracterizado pelo início precoce, BMI baixo, fraco controlo metabólico, deficiência de insulina e presença de anticorpos GADA; cluster 2, Diabetes com Deficiência de Insulina Severa (*SIDD*), de início precoce, IMC relativamente baixo, baixa secreção de insulina, e fraco controlo metabólico; cluster 3, Diabetes com Insulinoresistência Severa (SIRD), caracterizada por elevada insulinoresistência e IMC elevado; cluster 4, Diabetes com obesidade moderada (*MOD*); e o cluster 5, em tudo semelhante ao cluster 4 mas com indivíduos mais idosos, que denominaram de Diabetes relacionada com a meia idade (MARD). Ainda, estes autores encontraram diferentes relações de complicações a estes grupos, nomeadamente o grupo SIRD relaciona-se mais com insuficiência renal. Estes resultados foram replicados em 3 coorte diferentes, embora da mesma zona geográfica. Assim, os autores propõem uma nova classificação com base nestes grupos, que permitem uma abordagem terapêutica

diferente. No entanto, dado que foram utlizados para o cluster, apenas indivíduos com diabetes de acordo com a classificação atual, esta nova classificação baseia-se na anterior, não se distanciando da mesma, e não permitindo perceber a relação com a pré-diabetes ou com a normalidade.

É expectável que a publicação de trabalhos, aplicando as metodologias do data-mining à diabetes, cresça consideravelmente. No entanto, sem esquecer a importância do conhecimento atual, é importante que estes sejam aplicados de novo, à população no seu todo, de modo a que possam validar os grupos já existentes, ou pelo contrário, apontar para a necessidade de uma abordagem diferente, eventualmente até de uma nova classificação. É na diferenciação dos indivíduos, da sua globalidade, que reside o grande potencial destas técnicas, não sendo possível atingi-lo se forem aplicadas, a classes formadas com base na estatística clássica, que não permite avaliar as potenciais diferenças entre as pessoas, como são as classes de diagnóstico atualmente aceites.

2.3. ANÁLISE DE CLUSTERS – SELF-ORGANIZING MAPS

Uma das principais tarefas do *data-mining* consiste na descoberta de padrões, informação útil e conhecimento, que não se revelam naturalmente, pela aplicação de processos computacionais de sistemas de informação a bases de dados complexas e de grande dimensão (Jothi et al., 2015).

A análise de *clusters* é uma técnica não supervisionada, que agrupa os indivíduos por semelhança. Os algoritmos de *clustering* criam grupos ou *clusters*, maximizando a semelhança dentro dos grupos e a dissemelhança entre os grupos, de modo a que os indivíduos em determinado grupo são mais semelhantes entre si, do que os indivíduos em *clusters* diferentes (Jain, Murty, & Flynn, 1999). Diz-se não supervisionado, pois a classificação num grupo é feita, sem conhecimento *à priori* da classe a que pertencem os dados de treino. Dada a quantidade de dados existentes atualmente, e a dificuldade que o Homem tem em processá-los e compreendê-los, este método torna-se uma ferramenta extremamente importante, porque permite explorar e revelar padrões naturais escondidos nos dados, identificando grupos ou subgrupos que têm caraterísticas semelhantes e ajudando, também, à sua visualização, tornando possível a análise destas bases de dados multidimensionais complexas (Koh & Tan, 2005).

A análise de *clusters* tem vindo a ser cada vez mais utilizada no reconhecimento de padrões, no processamento de imagem, na pesquisa de informação, e aplicada em diversos campos como por exemplo à biologia, à química, ao marketing, à geografia, e, também à medicina, sendo um método muito utilizado por várias comunidades de investigação (Jain et al., 1999).

Existem múltiplos algoritmos para a análise de *clusters* e podem ser divididos em hierárquicos e de partição. Os algoritmos hierárquicos encontram *clusters* que se podem agrupar de modo hierárquico. Os métodos de partição como é o caso do *k-means*, constroem partições de uma base com N objetos, em *k* clusters. Dada a vasta variedade de algoritmos existentes, a escolha do que melhor se adapta ao problema que se pretende resolver, nem sempre é fácil e deve ter em consideração as características, as vantagens e desvantagens dos diferentes algoritmos (Jain et al., 1999).

O Self-Organizing Map (SOM), descrito por Kohonen nos anos 80 (Kohonen, 1990), é um algoritmo não supervisionado, de aprendizagem competitiva, baseado em redes neuronais artificiais (algoritmos inspirados na arquitetura do Sistema Nervoso). Ao agrupar os objetos pela sua semelhança, este algoritmo mapeia-os numa rede de duas dimensões, que preserva as relações topográficas entre os mesmos. O SOM permite explorar, resumir e visualizar dados de elevada dimensionalidade (Kesumawati & Setianingsih, 2016). Por fazer um resumo dos dados em protótipos numa rede bidimensional, permitindo a sua visualização e a sua caracterização, o SOM é um algoritmo popular e tem sido muito utilizado na exploração de dados multidimensionais em diferentes áreas (Tirunagari et al., 2015).

A rede SOM Kohonen é uma rede neuronal de duas camadas, com uma camada de entrada e uma camada de saída (Figura 1). A camada de entrada é constituída pelos vetores de entrada, que correspondem aos objetos que se pretendem agrupar dispostos no espaço multidimensional. A camada de saída é representada por uma grelha bidimensional de unidades, que constituem protótipos dos vetores de entrada, agrupados pelo algoritmo. A avaliação da semelhança dos vetores de entrada às unidades protótipo é feita por uma medida de distância (distância Euclidiana, ou outra), dependendo do problema em questão. Em cada iteração do algoritmo, os vetores de entrada são agrupados à unidade da rede cuja distância é menor, sendo essa a unidade ganhadora (*Best Matching Unit - BMU*).

O algoritmo inicia-se quando um vetor de entrada é apresentado à rede. Os vetores de entrada podem ser escolhidos aleatoriamente, e podem ser apresentados sequencialmente à rede (algoritmo sequencial), ou podem ser apresentados simultaneamente em cada iteração (algoritmo *batch*) (Fort, Cottrell, & Letremy, 2001). Para cada vetor de entrada é calculada a distância a todas as unidades da rede, e selecionada aquela que o melhor representa como unidade ganhadora (*BMU)*. Esta unidade vai-se aproximar do vetor de entrada, de acordo com uma taxa de aprendizagem definida. As unidades vizinhas da unidade ganhadora podem-se adaptar também, de acordo com uma função de vizinhança, sendo a abrangência da influência determinada pelo raio de

vizinhança. Assim, em cada iteração a rede adapta-se aos dados de entrada através da aproximação das unidades ganhadoras, mas também da adaptação das unidades vizinhas. No algoritmo sequencial (*on-line*), em cada iteração é apresentado um vetor de entrada, e ao conjunto de iterações que apresenta todos os vetores existentes denomina-se época. O algoritmo é treinado em várias épocas, e em cada época diminui-se a taxa de aprendizagem e o raio de vizinhança. Nas primeiras épocas, quando a taxa de aprendizagem e o raio de vizinhança são máximos, a rede modifica-se mais, pois as unidades que a constituem são mais fortemente atraídas pelos vetores de entrada. A rede é desdobrada quando as unidades se dispõem, de modo a melhor representar os objetos que lhe são próximos. À medida que as épocas se sucedem, a taxa de aprendizagem e o raio de vizinhança vão diminuindo, as unidades começam a estabilizar, sendo nas últimas épocas que se fazem os ajustes mais finos dos protótipos. No final, a rede de saída constitui uma representação bidimensional dos vetores de entrada no espaço multidimensional, mantendo a sua relação topográfica, na qual cada unidade constitui um protótipo, que representa e resume os objetos que nela se agrupam (Yin, 2008). A rede treinada permite, para além da visualização e compreensão dos dados, a classificação de novos dados. Para classificar um novo objeto, calcula-se a distância do mesmo a todas as unidades da rede, e atribui-se à unidade que lhe é mais próxima (*BMU*).

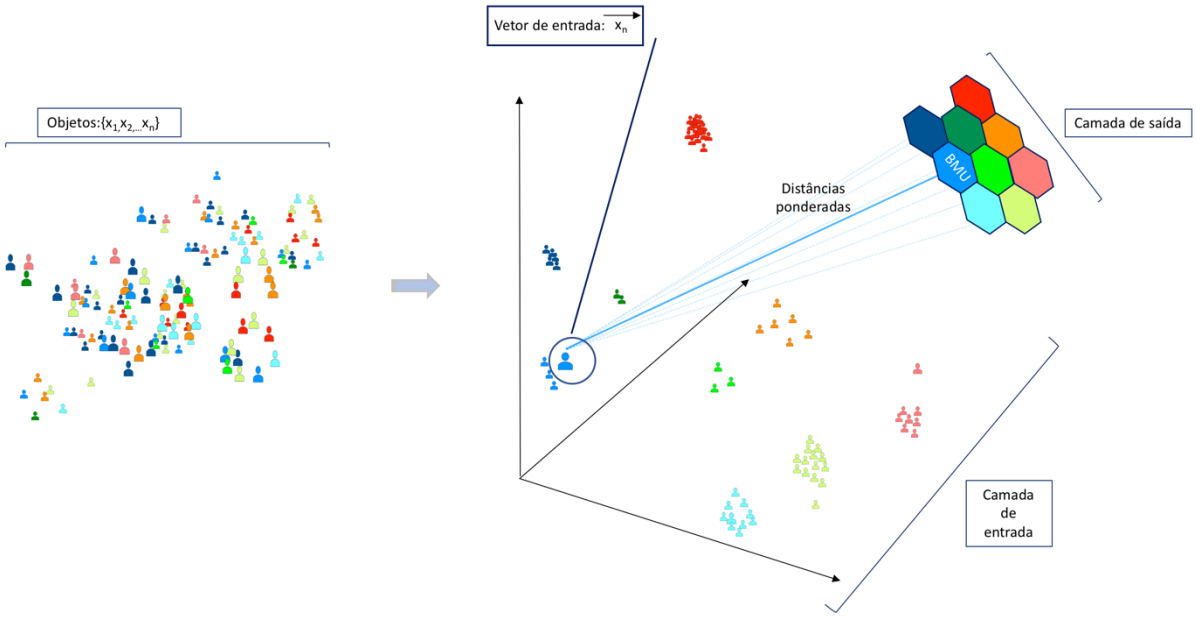


Figura 1. Esquema simplificado do *Self-Organizing Map*: Os vetores de entrada correspondem aos objetos da população, dispostos no espaço multidimensional (camada de entrada). Em cada iteração, são calculadas as distâncias ponderadas de cada vetor de entrada ($\overrightarrow{x_n}$) a todas as unidades do *SOM*, e a unidade que apresenta menor distância para cada $\overrightarrow{x_n}$ move-se para junto do mesmo (*BMU*). A camada de saída é uma rede bidimensional, constituída por unidades, que representam o centroide do grupo de dados que contêm, e que mantém a topologia dos dados.

Para a visualização e compreensão dos protótipos na rede podemos utilizar a matriz-U. A matriz-U é uma representação do SOM, na qual se representa, entre as unidades da rede, a distância entre as mesmas numa escala de cor. Na Figura 2., por exemplo, a cor mais escura entre as unidades adjacentes corresponde a uma maior distância entre as mesmas, e traduz, portanto, protótipos que diferem mais entre si, do que os que se encontram separados por uma cor mais clara, que representa uma menor distâncias dos pesos de cada unidade ou neurónio. As áreas mais claras representam *clusters* mais próximos, separados por áreas mais escuras, de maior distância (Ultsch, 2003). Esta ferramenta é também muito útil, quando se desconhece o número de *clusters*, permitindo identificá-lo, por aproximação, e de um modo intuitivo. Os vários parâmetros, que constituem os vetores de entrada, também podem ser analisados visualmente através dos *component planes*. Nesta representação utiliza-se uma grelha de unidades preenchidas com um gradiente de cor, fazendo-se corresponder os extremos deste gradiente aos extremos dos valores da variável a analisar. Deste modo, a representação da distribuição de cada variável na grelha pode ser observado, de um modo muito mais intuitivo, através de uma distribuição de cor (Kohonen, 2001). Em resumo, na matriz U podemos detetar unidades que se podem agrupar, por serem muito próximas e portanto semelhantes, e os *component planes*, ajudam a revelar que características ou dimensões dos objetos são mais determinantes nas suas diferenças (Toppila, 2016).

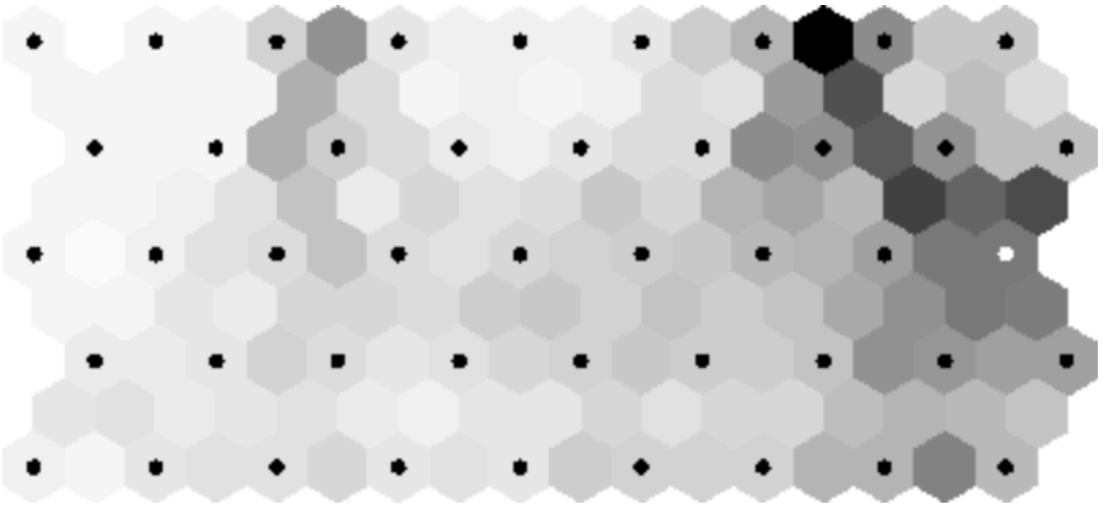


Figura 2. Representação de uma Matriz – U (Ultsch, 2003)

O SOM tem várias características que o torna apelativo para aplicação na área clínica. Apesar de ter semelhanças com algoritmos padrão como o *k*-means, apresenta vantagens, como poder não

ser necessário fixar o número de *clusters à priori* (SOM emergente), lidar com bases de dados grandes e complexas e encontrar cluster com vários formatos. Bação *et al* sugerem, inclusivamente, que o SOM é mais robusto que o *k-means*, e que faz uma pesquisa mais exaustiva do espaço de soluções, sendo menos propenso em parar em ótimos locais, desde que parametrizado corretamente (Bação, Lobo, & Painho, 2005).

2.3.1. Parâmetros do SOM

Na implementação do algoritmo de SOM , a parametrização correta é fundamental para a qualidade dos resultados (Bação et al., 2005). Várias decisões têm que ser tomadas e incluem: a forma das unidades da rede; o tamanho (número de unidades) e a topologia da rede; a inicialização do algoritmo, o raio e a função de vizinhança, a taxa de aprendizagem e a forma (função) como a taxa de aprendizagem e o raio de vizinhança decrescem (Kohonen, 2013); e ainda o número de épocas ou iterações, e o modo como em cada época os vetores são apresentados à rede (algoritmo *batch* ou algoritmo sequencial).

A forma das unidades da rede pode ser arbitrária, tomando normalmente uma de duas formas: quadrada, em que cada nó tem quatro vizinhos diretos, ou hexagonal, tendo cada nó seis vizinhos diretos. A rede hexagonal torna mais fácil a visualização.

A rede pode apresentar diversas topologias. Nas redes planas, que se apresentam como se de uma folha de papel se tratasse, existem um conjunto de nós periféricos, que têm menor número de vizinhos, e que sofrem influência de menos unidades. Os valores mais extremos têm tendência para se agrupar nestas unidades, nos cantos e lados da rede. Um modo de ultrapassar este problema é ligar os lados e cantos da rede, dando origem a outros formatos como é o caso do cilindro, do *toróide* ou outros. A rede treinada nestes formatos, não tendo unidades com menor vizinhança, é depois avaliada num plano bidimensional. Neste caso, na interpretação da matriz U e dos *component planes* deve ser tido em conta, que, aquando do treino, as unidades de um dos lados da grelha estavam ligadas às unidades do lado oposto.

O número de unidades, e o modo como se dispõem na grelha, pode interferir também na performance do SOM. Este algoritmo é muitas vezes utilizado como um método de redução de dimensionalidade, determinando-se o número de *clusters* por visualização ou aplicando um algoritmo para o agrupamento dos protótipos da rede. Neste caso, utiliza-se uma rede com maior número de unidades. Um caso particular deste tipo de visualização é o SOM emergente, em que o algoritmo é treinado com um número elevado de unidades, de modo a deixar bem demarcado os

intervalos entre os *clusters* (Ultsch & Mörchen, 2005). Embora a seleção do número de unidades possa ser feita arbitrariamente, ou por um processo recursivo com avaliação de resultados, existem algumas regras que devem ser consideradas Um dos métodos de determinar o tamanho da rede é pela regra prática m=5√n, em que *m* é o número de *clusters*, e *n* é o número de objetos de treino. A relação entre os lados da rede é a relação entre os 2 *eigenvalues* mais elevados de uma matriz de covariância dos dados de treino. Mas o SOM pode ser utilizado como algoritmo de *clustering* diretamente. Neste caso, o número de unidades deve ser previamente conhecido, e limita e corresponde ao número de *clusters* encontrados por este algoritmo. Um dos métodos de calcular previamente o número de *clusters,* é através da construção de um gráfico, entre o número sucessivo de *clusters* (*k*=2,3,4,5.....20) e o valor de uma métrica de qualidade, que traduz a distância *intra-cluster* versus a distância *inter-cluster*. Determina-se, então, o ponto onde se desenha um "*cotovelo*", correspondendo normalmente ao número ótimo de *clusters* (Zhao, 2012).

A posição inicial das unidades no espaço multidimensional pode ser determinada de modo aleatório, ou ser definida de acordo com o conhecimento dos dados. Nenhum destes métodos parece ser superior (Toppila, 2016).

A taxa de aprendizagem inicial, o raio e a função de vizinhança, e a forma como estes decaem, são também fundamentais nos resultados obtidos com o algoritmo. Embora possam ser ajustados recursivamente, a taxa de aprendizagem é tipicamente elevada e próxima de 1 no início, e o raio de vizinhança é também abrangente nesta fase, de modo a influenciar maior número de unidades. Pode-se optar por fazer decair estes parâmetros, à medida que as épocas se sucedem, de modo inversamente proporcional, linear, ou exponencial (Kohonen, 2013).

O número de épocas deve ser suficientemente grande para permitir uma boa acuidade e que o algoritmo tenha uma boa aprendizagem, sem gastar capacidade computacional e tempo desnecessários.

2.3.2. SOM multicamada ou SuperSOM

Na exploração e resolução de problemas mais complexos, existe uma variante do SOM que permite o treino de várias grelhas de modo simultâneo, utilizando diferentes conjuntos de caraterísticas dos mesmos objetos em cada uma das grelhas (R. Wehrens, 2015), denominado aqui SuperSOM. Dado que o treino é feito de modo independente, cada grelha pode conter um número diferente de variáveis dos vetores de entrada. Neste algoritmo, a unidade vencedora é calculada, através do somatório ponderado das distâncias do vetor de entrada às unidades, que se

correspondem através dos espaços multidimensionais das várias grelhas. Deste modo, a unidade selecionada para cada vetor de entrada será a que melhor o representa em média, através de todas as grelhas treinadas. As grelhas nas diferentes camadas são semelhantes, de modo que há correspondência das unidades através das grelhas. Em cada época, para cada vetor de entrada, a unidade vencedora é a mesma em todas as redes, e esta move-se para junto do vetor que representa, do mesmo modo que se processa num algoritmo SOM com uma única camada.

Os pesos das redes podem ser ajustados de modo a dar importâncias diferentes a cada uma delas. Isto é feito, tendo em conta o número de variáveis que existem em cada rede, dado que uma variável numa rede que inclui mais dimensões, terá menor peso que outra, que se encontra numa rede treinada com menor número de variáveis. Ainda, o conhecimento do problema em resolução pode ajudar também no ajuste deste parâmetro, pois algumas características das observações podem ser reconhecidamente mais importantes relativamente a outras.

2.3.3. Seleção das medidas de avaliação do SOM

A análise de *clusters* é uma técnica não supervisionada. Não existindo nenhuma classe conhecida *à priori* dos dados de treino, com a qual se possa comparar os resultados, a avaliação dos resultados torna-se mais difícil do que nos casos de classificação (ou supervisionados). Como o agrupamento em *clusters* não tem soluções certas ou erradas, dado que os dados estão a ser agregados por semelhança, sem conhecimento das classes que podem estar representados nos mesmos, a avaliação é feita através de índices de validação, ou seja, medidas que traduzem a qualidade dos *clusters*, avaliando em que medida, os *clusters* representam os vetores de entrada que lhe deram origem e os representam. Os índices de validação dos *clusters* permitem, para além da avaliação da qualidade dos resultados da aplicação de um algoritmo, comparar resultados na aplicação do mesmo algoritmo, ajudando à sua parametrização, bem como comparar algoritmos (Zhao, 2012).

Os índices de validação de *clusters*, podem ser divididos em índices de validação externa e interna. Enquanto que os índices de validação externa pressupõem um conhecimento à priori dos dados, os índices de validação interna utilizam informação intrínseca aos dados (Zhao, 2012).

Os índices de validação interna baseiam-se normalmente em dois critérios: compacidade, que mede a proximidade dos objetos num grupo; e a separação, que mede a dissemelhança entre grupos. Exemplos destas medidas são os índices de Dunn, silhueta, Davies-Bouldin (Zhao, 2012). No índice de Davies-Bouldin (DB) calculam-se as semelhanças de cada cluster *C* a todos os restantes, e é

atribuído o valor de maior semelhança. O índice de DB, que se pretende minimizar, é obtido pela média do máximo da razão entre a semelhança *intra-cluster* e a separação inter-*cluster*, para cada *cluster* presentes.

Uma das medidas mais utilizadas para avaliar a qualidade dos *clusters* na aplicação do SOM, e que foi proposta por Kohonen, é o *quantization error* (QE) (Uri & Breard, 2017). Esta medida é também útil na afinação dos parâmetros do SOM, na qual se fazem pequenos ajustes de modo a minimizá-la. O QE mede a semelhança dos protótipos aos vetores de entrada, e é definido como a distância média entre os vetores de entrada e a sua BMU, sendo esta a unidade da rede que os representa no espaço multidimensional. Se os vetores de entrada são muito distantes da sua BMU, o QE vai aumentar, traduzindo uma fraca representação por parte das BMU. Pelo contrário, se o QE é baixo, é porque a distância média dos vetores de entrada à sua BMU é baixa, estando por isso mais próximos, indicando que o protótipo constitui uma boa representação dos vetores de entrada. Um dos objetivos do SOM é manutenção da topologia dos dados na rede. Outra medida, o *topographic error* (TE), mede a qualidade do mapa na modelação dos vetores de entrada. Calculam-se a primeira e a segunda unidades da rede mais próximas a cada vetor de entrada, e avalia-se as suas posições. Considera-se que a topologia da rede está conservada, se estas unidades são vizinhas. O TE é calculado pelo quociente entre o número total de erros e o número total de vetores de entrada (Uri & Breard, 2017).

2.3.4. *Clustering* dos protótipos

O superSOM pode ser utilizado para criar *clusters,* sem necessidade de aplicar outros algoritmos posteriormente. Para tal, determina-se inicialmente o número de *clusters* a formar, e faz-se corresponder este valor ao número de unidades da grelha do SOM. No entanto, esta algoritmo é também muito útil na exploração e visualização dos dados, podendo ele próprio ser utilizado na determinação do número de *clusters*. Assim, o SOM (ou o superSOM), pode ser utilizado com o objetivo de reduzir a dimensionalidade dos dados, sem conhecimento prévio do número de *clusters* a constituir. Neste caso, escolhe-se normalmente uma grelha com um número elevado de unidades (SOM emergente), e em seguida aplica-se outro algoritmo de cluster (k-means, algoritmos de cluster hierárquicos) para agrupar os protótipos da rede (J. Vesanto & Alhoniemi, 2000). O algoritmo hierárquico é frequentemente utilizado como segundo algoritmo, pois permite ir agrupando ou desagrupando as unidades do SOM, de um modo hierárquico, à medida que se exploram os dados (Juha Vesanto, Alhoniemi, & Member, 2000). Para tal constrói-se um dendograma, diagrama que representa a hierarquia, que se corta ao nível que se pretende analisar. Podemos então agregar e

desagregar as unidades, de acordo com o ganho de informação que se consegue a cada nível. Topilla *et al*, testaram vários métodos de agregação dos protótipos do SOM (*spectral clustering, affinity propagation*, cluster hierárquico pelo método Ward), concluindo que o cluster hierárquico, pelo método Ward, é o mais robusto. Ainda, ao possibilitar a criação de um dendograma, apoia a decisão quanto ao número de *cluster* a identificar.

2.4. HIPÓTESE GERAL E OBJETIVOS DO ESTUDO

A hipótese deste trabalho é que a aplicação do algoritmo superSOM (algoritmo de clustering), a uma base de dados de prevalência da Diabetes (Prevadiab2), origina *clusters*, associados a novos fenótipos, contidos nas classes de hiperglicemia, definidas pelos critérios de diagnóstico atualmente utilizados. A identificação de diferentes fenótipos, utilizando características antropométricas e valores analíticos considerados importantes na fisiopatologia da hiperglicemia, pode dar um contributo à compreensão dos mecanismos da doença, bem como à definição de uma classificação mais adequada, a uma abordagem terapêutica de precisão, que se espera mais eficaz.

Objetivos específicos:

- Avaliar se a análise de *clusters,* utilizando o algoritmo *superSOM,* identifica diferentes fenótipos, nas classes de hiperglicemia atualmente definidas

Objetivo secundário:

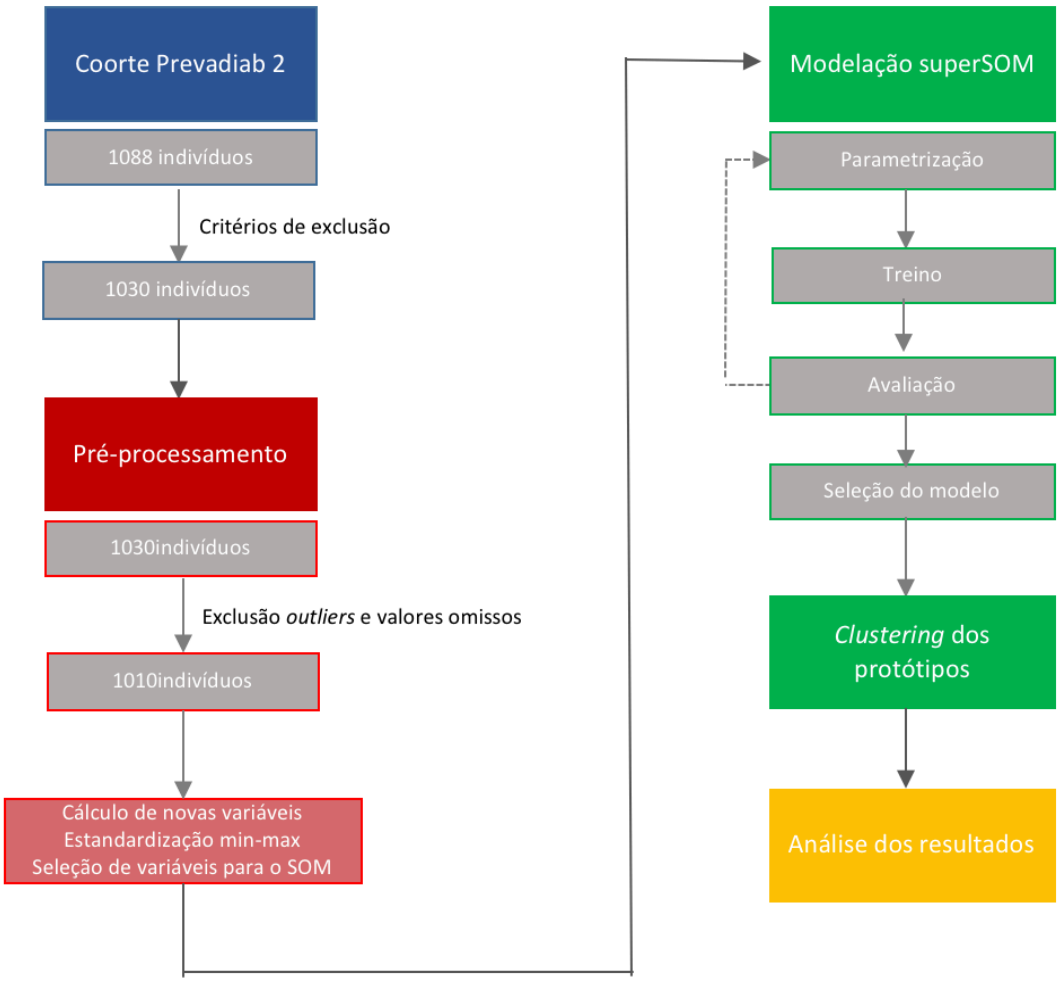
- Avaliar e caracterizar os *clusters*, encontrados na aplicação do algoritmo *superSOM,* no que respeita às características antropométricas e de parâmetros analíticos metabólicos.

3. MATERIAIS E MÉTODOS

Neste trabalho utiliza-se a base Prevadiab2, descrita em 3.1. Após a preparação e pré-processamento dos dados (3.2.) descreve-se a aplicação do algoritmo *SuperSOM*, implementado em R, e os parâmetros utilizados. No capítulo 3.4. explica-se como o modelo foi avaliado, como foi feita a afinação dos parâmetros, e selecionado o modelo final a avaliar. A Figura 3. representa, a metodologia utilizada neste trabalho, esquematicamente e de modo sumário.

Neste trabalho foram utilizados os softwares: EXCEL® e R®. Em particular, neste último, foram utilizadas para além das funções base, os *packages* *kohonen* (M. R. Wehrens & Kruisselbrink, 2018), *stats* (R Core Team R Foundation for & Computing, 2018), e *ggplot2* (Wickham, 2016).

Figura 3. Esquema da metodologia utilizada



3.1. O ESTUDO PREVADIAB2

Utilizámos os dados provenientes do estudo PREVADIAB2.

O estudo PREVADIAB é um estudo de prevalência da diabetes em Portugal. Iniciou-se em 2008/2009 com o estudo PREVADIAB 1, cujo protocolo já se encontra publicado *(Gardete-Correia et al.*, 2010). Resumidamente, foram incluídas 5167 pessoas entre os 20 e 79 anos, selecionadas de 122 localizações de Portugal, representativas da distribuição da população portuguesa. A estes indivíduos, após o jejum noturno, foi colhida uma amostra de sangue e realizada uma PTGO (75g de glicose) (Hoff, Fish, & Service, 2001). Foram ainda registados dados antropométricos e outras análises bioquímicas.

O estudo PREVADIAB 2 é um *follow-up* do PREVADIAB 1, realizado entre Dezembro de 2011 e Junho de 2014, que teve como objetivo primário avaliar a progressão para pré-diabetes e diabetes. Neste estudo foram contactadas para participação, as pessoas sem diabetes que tinham sido avaliadas no PREVADIAB 1. Das pessoas contactadas, responderam e foram incluídos 1088 indivíduos, entre os 22 e 86 anos, distribuídas por 55 centros de saúde de Portugal. A cada pessoa foi solicitado um jejum noturno de pelo menos 12h. No dia da visita à clínica, foi feito um questionário clínico, foram recolhidos dados biométricos e clínicos, foi recolhida uma amostra de sangue em jejum, na qual foram avaliados parâmetros bioquímicos. Em sequência, foi realizada uma prova oral de tolerância à glucose (PTGO) (Hoff et al., 2001), com colheitas adicionais de sangue aos 30' e 120', avaliando-se para além da glicemia a insulina, o peptídeo C, os ácidos gordos livres (Tabela 2). Quando houve alguma dúvida numa medição, esta foi repetida.

Tabela 2. Parâmetros clínicos e analíticos registados na avaliação do estudo PREVADIAB 2

Variáveis Gerais: Idade, Género, Escolaridade, Atividade Física, Consumo de frutas e vegetais, Consumo de anti-hipertensores, Tabagismo, Antecedentes familiares de Diabetes tipo 2.
Variáveis Antropométricas/Clínicas: IMC, PA, Bioimpedância, Tensão arterial sistólica, Tensão arterial diastólica
Variáveis Laboratoriais em jejum, aos 0' da PTGO: Plaquetas, Albumina, Creatinina, AST, ALT, GGT, Colesterol Total, HDL, LDL, Triglicéridos, Ácidos gordos livres, Glicose, Insulina, Peptídeo C.
Variáveis Laboratoriais aos 30' da PTGO: Glicose, Insulina, Peptídeo C e Ácidos gordos livres
Variáveis Laboratoriais aos 120' da PTGO: Glicose, Insulina, Peptídeo C e Ácidos gordos livres

3.2. VARIÁVEIS CLÍNICAS E INCLUSÃO DE INDIVÍDUOS

A população incluída neste trabalho é de 1030 indivíduos. Dos 1088 indivíduos avaliados no PREVADIAB 2 excluímos 58, por terem diagnóstico de diabetes tipo 2 e estarem a fazer medicação antidiabética no momento da avaliação.

Das variáveis recolhidas e medidas no estudo PREVADIAB 2, considerámos as que têm maior objetividade na sua análise, como é o caso dos dados antropométricos e dos valores analíticos. Assim foram selecionadas 27 variáveis, cujas estatísticas se encontram no Anexo A:

- Idade e Género
- IMC, PA e Bioimpedância
- Plaquetas, Albumina, Creatinina, AST, ALT, GGT, Colesterol Total, HDL, LDL, Triglicéridos, Ácidos gordos livres (0', 30' e 120' da PTGO), Glicose (0', 30' e 120' da PTGO), Insulina (0',30' e 120' da PTGO), Peptídeo C (0', 30' e 120' da PTGO).

3.3. SUPERSOM DO PREVADIAB 2

3.3.1. Limpeza dos dados

Sendo uma base de um estudo epidemiológico, esta base contém dados de elevada qualidade, ao contrário do que é habitual nas bases de dados clínicos.

Não encontrámos valores que possamos identificar como erros. Quanto aos *outliers*, para além dos valores terem sido confirmados, estes são os que provavelmente nos interessam mais, dado que estamos a estudar casos de doença. Tratando-se de um algoritmo de *clustering,* é ainda expectável que sejam formados grupos com estes indivíduos, que nos interessam estudar. Assim, analisámos os *outliers* por género e excluímos apenas *outliers* extremos, que correspondem a 18 observações que apresentavam valores superiores a 5 *SD*.

Foram excluídos dois registos que apresentavam elevada percentagem de valores omissos, nomeadamente de variáveis que os classificam quanto à glicemia.

Foram excluídas variáveis com percentagem de valores omissos superior a 10% (Anexo A): bioimpedância, plaquetas e albumina. Embora seja uma percentagem relativamente elevada, optamos por este valor, tendo em conta que a implementação do algoritmo *superSOM* em R que vamos utilizar, lida com os valores omissos, deixando um número de observações suficiente para prosseguir com a análise.

Relativamente aos valores omissos nas variáveis que iremos selecionar, optámos por não os simular com um dos múltiplos métodos existentes. A implementação em R do algoritmo *superSOM* consegue lidar com eles e, deste modo, evitamos um viés.

3.3.2. Transformação dos dados

As variáveis a ser utilizadas têm escalas diferentes, pelo que é necessária a sua normalização. Optámos por utilizar o método min-max (-1,1), e normalizámos todas as variáveis. No entanto, corrigimos as variáveis para o género antes da sua normalização. Variáveis como o IMC e o PA, por exemplo, têm significados diferentes em ambos os géneros, e a sua diferença poderia escamotear padrões inerentes aos dados. Por exemplo, o perímetro abdominal de risco de uma mulher é inferior ao do homem.

Testámos a normalidade da distribuição das variáveis com o teste Mann-Whitney U. A maior parte das variáveis não têm uma distribuição normal. Optámos por deixá-las com a sua distribuição original. Cientes de que pode interferir na performance do algoritmo, esta não é estritamente necessária e traz-nos alguns benefícios na interpretação dos resultados.

3.3.3. Seleção e extração de variáveis

A seleção e extração de variáveis, para as diferentes grelhas do *superSOM*, foi feita equilibrando dois pontos essenciais: variáveis o mais objetivas possível, e aquelas, que derivadas de cálculos, são consideradas importantes do ponto de vista científico, na exploração dos diferentes mecanismos fisiopatológicos da diabetes tipo 2. Assim tivemos a necessidade de extrair algumas variáveis, que embora não sejam originais consideramos fundamentais nesta análise.

Algumas das análises laboratoriais clínicas foram medidas durante uma PTGO aos 0', 30' e 120'. Isto permite-nos ter um perfil temporal destes valores para cada indivíduo, e parece-nos poder ter importância na diferenciação dos diferentes fenótipos, pelo que selecionamos os diferentes valores temporais de glicemia, de insulina, de peptídeo C, de *clearance* de insulina, e de ácidos gordos livres. A *clearance* de insulina, dada pela relação entre o peptídeo C e a insulina, traduz se o nível de insulina se deve à secreção pela célula ß do pâncreas, ou à maior ou menor eliminação da mesma da corrente sanguínea. Extraímos então, as diferenças dos valores temporais para cada uma delas, de modo a valorizar também o modo como se modificam no tempo, e não apenas o valor absoluto.

Consideramos também outros valores laboratoriais, como os dos lípidos (ex. LDL e HDL), dado que a dislipidemia tem sido implicada nos mecanismos que estão associados à génese da diabetes tipo 2. Selecionámos ainda variáveis antropométricas, reconhecidamente implicadas na fisiopatologia da diabetes, como o perímetro abdominal e o Índice de Massa Corporal. Por fim, e considerando a importância da insulinoresistência e da função da célula ß, e dado que a IR apresentava uma elevada correlação com a insulinemia, extraímos o HOMA-B e o *Disposition Index* (DI). O DI é o quociente entre o HOMA-B e o HOMA-IR, e como nem o HOMA-B, nem o DI apresenta uma elevada correlação com as outras variáveis selecionadas, consideramos assim a insulinoresistência de modo indireto. Quanto à medida de tensão arterial (sistólica e diastólica), apesar de se considerar uma comorbilidade importante, como foi avaliada apenas numa medição e alguns indivíduos estavam a fazer antihipertensivo, foram excluídas da modelação para evitar um viés.

As variáveis selecionadas para as diferentes grelhas foram testadas quanto à sua correlação e foram excluídas quando o coeficiente de correlação foi superior a 0,85 (Anexo B). Todas as variáveis utilizadas em cada grelha têm valores abaixo deste limiar.

As Tabelas 3 e 4 mostram as variáveis extraídas e selecionadas, para o modelo e para avaliação posterior dos *clusters*.

Tabela 3. Variáveis normalizadas incluídas no modelo

Variável	Descrição
Minmax_BMI	Índice de Massa Corporal
Minmax_PA	Perímetro abdominal
Minmax_TAG	Triglicéridos
Minmax_LDL	Lipoproteínas de baixa densidade
Minmax_HDL	Lipoproteínas de alta densidade
Minmax_HOMAB	Função da célula ß calculado pelo modelo HOMA1
Minmax_DI	*Disposition índex* calculado pelo: HOMA_B/HOMA_IR
Minmax_FFA_0	Ácidos gordos livres aos 0' de uma PTGO
Minmax_D_FFA_30_0	Variação de ácidos gordos livres entre os 0' e os 30' de uma PTGO
Minmax_D_FFA_120_30	Variação de ácidos gordos livres entre os 30' e os 120' de uma PTGO
Minmax_D_FFA_0_120	Variação de ácidos gordos livres entre os 120' e os 0' de uma PTGO
Minmax_Glucose_0	Glicemia aos 0' de uma PTGO
Minmax_D_Gluc_30_0	Variação de glicemia entre os 0' e os 30' de uma PTGO
Minmax_D_Gluc_120_30	Variação de glicemia entre os 30' e os 120' de uma PTGO
Minmax_D_Gluc_0_120	Variação de glicemia entre os 30' e os 120' de uma PTGO
Minmax_C_Peptide_0	Concentração de Peptídeo C no sangue aos 0' de uma PTGO
Minmax_D_Cpep_30_0	Variação de concentração de Peptídeo C no sangue entre os 0' e os 30' de uma PTGO
Minmax_D_Cpep_120_30	Variação de concentração de Peptídeo C no sangue entre os 30' e os 120' de uma PTGO
Minmax_D_Cpep_0_120	Variação de concentração de Peptídeo C no sangue entre os 120' e os 0' de uma PTGO
Minmax_Insulina_0	Concentração de Insulina no sangue aos 0' de uma PTGO
Minmax_D_Insulin_30_0	Variação de concentração de insulinemia entre os 0' e os 30' de uma PTGO
Minmax_D_Insulin_120_30	Variação de concentração de insulinemia entre os 30' e os 120' de uma PTGO
Minmax_D_Insulin_0_120	Variação de concentração de insulinemia entre os 120' e os 0' de uma PTGO
Minmax_Clearance_0	*Clearance* de insulina aos 0' de uma PTGO
Minmax_D_Clearance_30_0	Variação da *clearance* de insulina entre os 0' e os 30' (AUC)

Minmax_D_Clearance_120_30	Variação da *clearance* de insulina entre os 30’ e os 120’ (AUC)
Minmax_D_Clearance_0_120	Variação da *clearance* de insulina entre os 120’ e os 0’ (AUC)

Tabela 4. Variáveis utilizadas no perfil dos *clusters*

Variável	Descrição
IMC	Índice de massa corporal = Peso/altura^2
PA	Perímetro abdominal
TAG	Triglicéridos
LDL	Lipoproteínas de baixa densidade
HDL	Lipoproteínas de alta densidade
HOMA_B	Função da célula ß calculado pelo modelo HOMA1
HOMA_IR	Insulinoresistência calculada pelo modelo HOMA1
FFA_0	Ácidos gordos livres aos 0’ de uma PTGO
FFA_30	Ácidos gordos livres aos 30’ de uma PTGO
FFA_120	Ácidos gordos livres aos 120’ de uma PTGO
Glucose_0	Glicemia aos 0’ de uma PTGO
Glucose_30	Glicemia aos 30’ de uma PTGO
Glucose_120	Glicemia aos 120’ de uma PTGO
C_Peptide_0	Concentração de Peptídeo C no sangue aos 0’ de uma PTGO
C_Peptide_30	Concentração de Peptídeo C no sangue aos 30’ de uma PTGO
C_Peptide_120	Concentração de Peptídeo C no sangue aos 0’ de uma PTGO
Insulina_0	Concentração de Insulina no sangue aos 0’ de uma PTGO
Insulina_30	Concentração de Insulina no sangue aos 30’ de uma PTGO
Insulina_120	Concentração de Insulina no sangue aos 120’ de uma PTGO
Clearance_0	*Clearance* de insulina em jejum (aos 0’ de uma PTGO) = C_Peptide_0/ Insulina_0
Clearance_AUC_0_30	*Clearance* de insulina dos 0’aos 30’ de uma PTGO = C_Peptide_AUC_0_30/ Insulina_AUC_0_30
Clearance_AUC_30_120	*Clearance* de insulina dos 30’aos 120’ de uma PTGO = C_Peptide_AUC_30_120/ Insulina_AUC_30_120
Idade	Idade
Género	Género

3.3.4. Análise estatística sumária dos dados selecionados

Os 1010 indivíduos analisados têm uma idade média de 60+/-13 anos e 60% são mulheres.

A Figura 4, mostra a distribuição por classes de hiperglicemia de acordo com a classificação da IDF e da ADA. Como seria expectável, a distribuição, por diagnóstico da pré-diabetes e diabetes, difere de acordo com os critérios utilizados. No Anexo C podemos observar a distribuição das variáveis selecionadas para a modelação do *superSOM*.

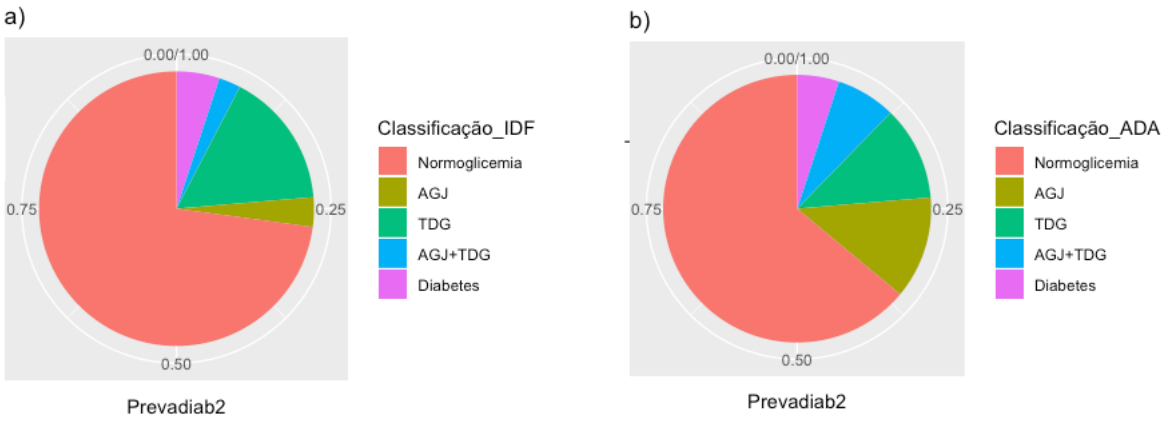


Figura 4. Classificação da população pela OMS/IDF (a) e pela ADA (b)

3.3.5. Modelação - Algoritmo *SUPERSOM*

Para o treino do *superSOM* utilizamos a implementação do algoritmo *superSOM* em R, do *package* Kohonen versão 3.0.5. (M. R. Wehrens & Kruisselbrink, 2018).

Este algoritmo permite a utilização de várias grelhas com as mesmas características, nas quais se agrupam diferentes variáveis. As unidades correspondem-se nas diferentes grelhas, e as unidades na grelha final representam os protótipos dos vetores de entrada que agrupam, resultando da média das distâncias destes às unidades de cada grelha.

3.3.5.1. Agrupamento de variáveis em cada grelha

Para o treino do *superSOM* utilizámos 8 grelhas. Deste modo podemos agrupar as variáveis de um modo cientificamente lógico, controlando o peso que pretendemos que cada grupo de variáveis tenha.

Assim agrupamos as variáveis normalizadas nas grelhas do seguinte modo:

#1ª grelha – Glicemia em jejum, diferença de glicemia dos 0’ aos 30’, diferença de glicemia dos 30’ aos 120’, diferença de glicemia dos 120’ aos 0’ de uma PTGO.

#2ª grelha – Concentração de peptídeo C no sangue venoso em jejum, diferença de concentração peptídeo C no sangue venoso entre os 0’ e os 30’, diferença de concentração de peptídeo C no sangue venoso entre os 30’ e os 120’, diferença de concentração de peptídeo C no sangue venoso entre os 120’ aos 0’ de uma PTGO.

#3ª grelha - Insulinemia em jejum, diferença de insulinemia dos 0’ aos 30’, diferença de insulinemia dos 30’ aos 120’, diferença de insulinemia dos 120’ aos 0’ de uma PTGO.

#4ª grelha – *Clearance* de insulina em jejum, diferença de *clearance* de insulina dos 0’ aos 30’ (a *clearance* aos 30´ é calculada pela AUC dos 0’ aos 30’), diferença de *clearance* de insulina dos 30’ aos 120’ (a *clearance* aos 120´ é calculada pela AUC dos 30’ aos 120’), diferença *clearance* de insulina dos 120’ aos 0’ de uma PTGO.

#5ª grelha – Ácidos gordos livres em jejum, diferença de concentração dos ácidos gordos livres no sangue venoso entre os 0’ aos 30’, diferença de concentração dos ácidos gordos livres no sangue venoso entre os 30’ aos 120’, diferença de concentração dos ácidos gordos livres no sangue venoso entre os 120’ aos 0’ de uma PTGO.

#6ª grelha – Concentração de colesterol LDL, colesterol HDL e triglicéridos no sangue venoso periférico.

#7ª grelha – Índice de massa corporal e perímetro abdominal.

#8º grelha - HOMAB, e *Disposition Index.*

Nas primeiras cinco grelhas consideramos as variáveis em jejum e o modo como evoluem ao longo da PTGO, e, portanto, não só os valores absolutos, mas também o perfil das mesmas no tempo.

A 6ª grelha diz respeito ao perfil lipídico de colesterol e triglicéridos.

A 7ª grelha considera os dados antropométricos do indivíduo.

Na última grelha consideramos a função da célula ß e a insulinorresistência. Optámos por considerar a insulinorresistência de modo indireto, através do quociente entre o HOMA-B e o HOMA-IR, dado a elevada correlação que o HOMA-IR tem com a insulinemia (>0.85).

3.3.5.2. Seleção dos parâmetros

Para a seleção dos parâmetros do algoritmo tivemos em consideração os parâmetros sugeridos por Toppila *et al*, no trabalho em que testa os parâmetros a utilizar na implementação do *supersom* do *package kohonen* no R, avaliando diferentes fenótipos na diabetes tipo1 no que respeita às complicações, constituindo uma estrutura metodológica para trabalhos deste género (Toppila, 2016). Depois de testados, foi feita uma afinação dos parâmetros, tendo em consideração a minimização do QE, e os parâmetros finais estão resumidos na Tabela 5.

Utilizamos grelhas hexagonais, toróides. Assim, cada unidade tem 6 unidades vizinhas, não havendo unidades com menos vizinhos.

O número de unidades escolhidas foi de 27, com uma base racional. Dado que existem na classificação da pré-diabetes e diabetes 3 classes de glicemias aos 0' e 120' de uma PTGO, considerando que neste estudo é avaliada também a glicemia aos 30' desta prova, e que ela possa ter também 3 classes, então, 27 é o número de combinações possíveis entre as três classes das três variáveis (3*3*3).

O tamanho e a altura da grelha derivam do número de unidades (9*3). Assim a altura é de 9 unidades com uma largura de 3 unidades.

Utilizamos o algoritmo *online*, utilizando a *sum-of-squares* como medida de distância. A inicialização do algoritmo, e o posicionamento das unidades é feito de modo aleatório, assim escolhemos a inicialização aleatória que está implementada como padrão no *package kohonen* do R A função de vizinhança é a gaussiana e a taxa de aprendizagem escolhida decresce de 0,05 a 0,01, ambas também implementadas como padrão. O tamanho da vizinhança inicia-se contendo 2/3 das unidades e decai de um modo linear. O número de épocas após os testes avaliando a progressão do QE foi definida em 500.

Relativamente aos pesos de cada grelha, considerando o número de variáveis que continham e a importância, foram definidos da seguinte forma de um modo sequencial da 1ª à 8ª grelhas: 0.18,0.18,0.18,0.18,0.03,0.03,0.03,0.18. Optámos assim, por valorizar mais os perfis da glicemia, da insulinemia, dos níveis de peptídeo C e *clearance* de insulina, em detrimento de variáveis que representam comorbilidades. As comorbilidades, embora sejam classicamente descritas com a diabetes tipo 2, nem sempre estão presentes. Deste modo, se encontrarmos diferentes fenótipos no que respeita à distribuição dos perfis, e se as comorbilidades forem importantes na sua individualização, é expectável que elas se diferenciem em cada *cluster*, ainda que os seus pesos nas grelhas não sejam tão significativos.

Este algoritmo consegue lidar com valores omissos. No entanto, é possível definir a percentagem de valores omissos a partir do qual um objeto é desconsiderado pelo algoritmo. Esta foi definia aos 50% (0.5).

Com estes parâmetros foram treinados 500 *superSOM,* e foi escolhido o que tinha menor QE. As unidades desta grelha foram então agregadas e analisadas, conforme se descreve no capítulo seguinte.

Tabela 5. Sumário dos parâmetros do algoritmo *superSOM*

Grelha		**Unidades**	
Topologia	Toróide	Topologia	Hexagonal
Altura / Largura	3 / 9	Número	27
Número	8		
Algoritmo			
Tipo	Sequencial	Vizinhança	
Inicialização	Aleatória	Função	
Medida de distância	*Sum-of-squares*	Inicial	0.05-0.01
Taxa de aprendizagem		Decaimento	Linear
Variação	0.05-0.01		
Decaimento	Linear		

3.3.6. *Clustering* dos protótipos

Para o *clustering* dos protótipos, utilizamos um dendograma, resultante da aplicação de um algoritmo de *cluster* hierárquico, utilizando o método Ward (*hclust* do *package stats*). O ponto de corte foi definido tendo em conta, por um lado o índice de Davies-Bouldin, numa agregação feita pelo *superSOM*, com uma parametrização idêntica, e por outro avaliando o benefício e utilidade da informação real ganha, ao descer a linha de corte no dendograma. Deste modo podemos ajustar o agrupamento, utilizando os mesmos protótipos. Os *clusters* finais foram então avaliados quanto à classificação em hiperglicemia de acordo com a OMS/IDF, bem como quanto ao perfil das variáveis que deram origem aos *clusters*.

4. RESULTADOS E DISCUSSÃO

Nesta secção descrevem-se e avaliam-se os resultados do *superSOM* selecionado (4.1.) e das suas unidades (4.2.). Faz-se, seguidamente, uma agregação por semelhança, ou quando a dissemelhança não o justifica do ponto de vista científico, com base num algoritmo de *clustering* hierárquico (4.3.). Avaliam-se os *clusters* encontrados, e faze-se uma discussão dos mesmos (4.4.).

4.1. RESULTADOS E ANÁLISE DA GRELHA FINAL DO SUPERSOM

O QE dos 500 *superSOM* treinados variou de 0,064 e 0,105. Escolheu-se o modelo que apresentou o menor QE (0,064).

Dado que neste trabalho utilizamos dois algoritmos de *clustering,* para clareza de descrição, chamaremos unidades, aos *clusters* derivados do *superSOM*, e *clusters* ou grupos, aos que derivam da agregação hierárquica das unidades do *superSOM*.

As 27 unidades, presentes na grelha final, fazem um resumo dos dados em protótipos (Figura. 5). As unidades apresentam distâncias diferentes entre si, sendo que existe um maior número de unidades mais próximas, e algumas que apresentam maior dissemelhança, parecendo existirem cerca de 9 *clusters* pela avaliação da matriz U.

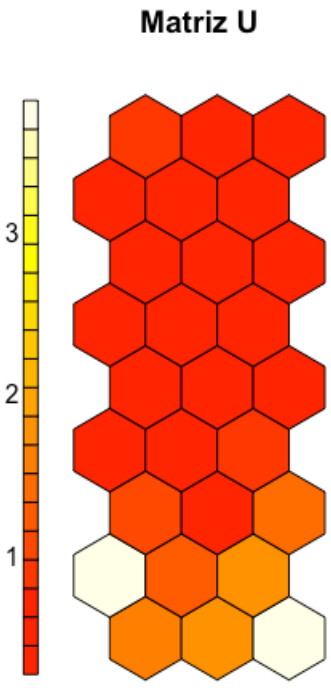


Figura 5. Matriz U do *superSOM* selecionado

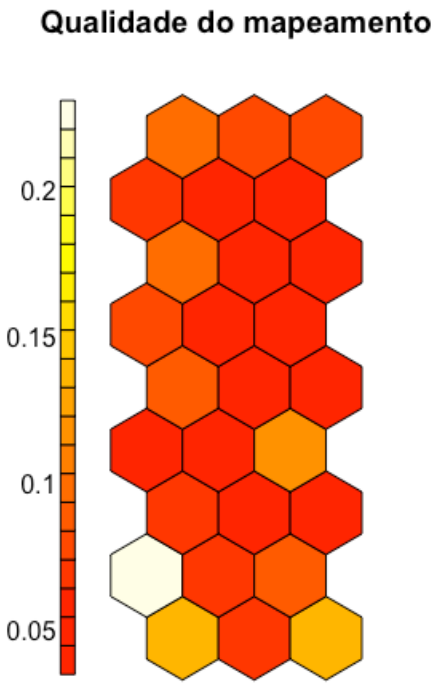


Figura 6. Qualidade do mapeamento: A escala de cor corresponde à distância média dos vetores de entrada ao centroide da unidade à qual são mapeados.

A Figura 6. representa a distância média dos indivíduos agrupados em cada unidade ao protótipo da mesma, dando uma perspetiva da qualidade do mapeamento. Observámos que, na maior parte das unidades, essa distância se encontra no limiar inferior da escala, havendo apenas uma unidade no limite superior, e cerca de 10 unidades com valores médios, sendo relativamente homogénea.

Dos 1010 indivíduos, 47 indivíduos não foram agrupados, por apresentarem um valor de variáveis omissas superior ao definido como aceitável no algoritmo. Assim as 27 unidades representam 969 indivíduos (Anexo D). As unidades diferem no número de indivíduos que agrupam, sendo expectável que as unidades que agrupam maior número de pessoas, agrupem os indivíduos com normoglicemia, o que será explorado em capítulo posterior. Sublinhamos ainda que a unidade que agrupa menor número de indivíduos tem também maior valor médio da distância dos indivíduos à mesma, e, portanto, menor qualidade no mapeamento.

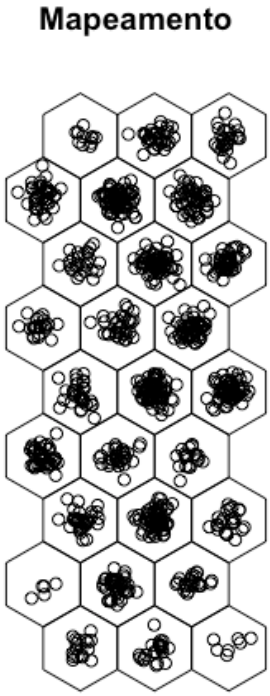
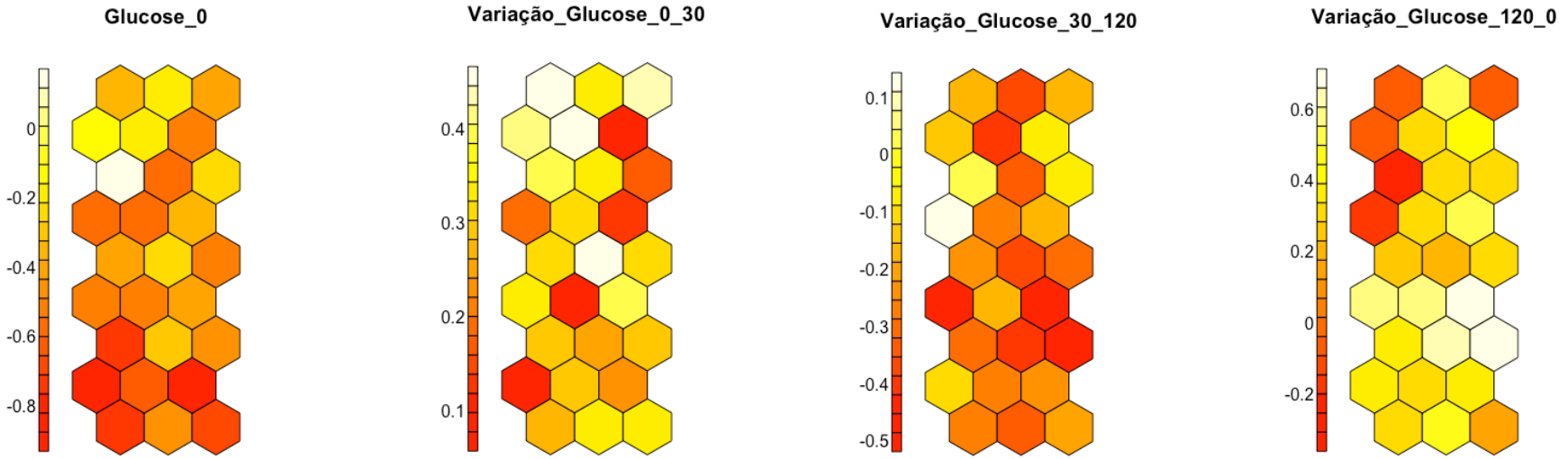


Figura 7. Número de indivíduos mapeados às diferentes unidades

4.2. RESULTADOS E ANÁLISE DAS UNIDADES DA GRELHA FINAL DO SUPERSOM

Apesar das variáveis se distribuírem de um modo contínuo, sendo difícil encontrar pontos de corte, quando avaliadas em conjunto, numa perspetiva multidimensional, são encontrados grupos. As diferenças dos valores médios, das variáveis utilizadas na modelação, podem ser avaliadas em *component planes* (Figura. 8) dando uma perspetiva das diferenças entre os grupos. Da análise dos gráficos anteriores respeitantes aos perfis de glicemia, insulinemia, *clearance* de insulina e níveis de Peptídeo C, podemos verificar que existem perfis semelhantes em mais do que uma unidade.

Grelha 1



Grelha 2

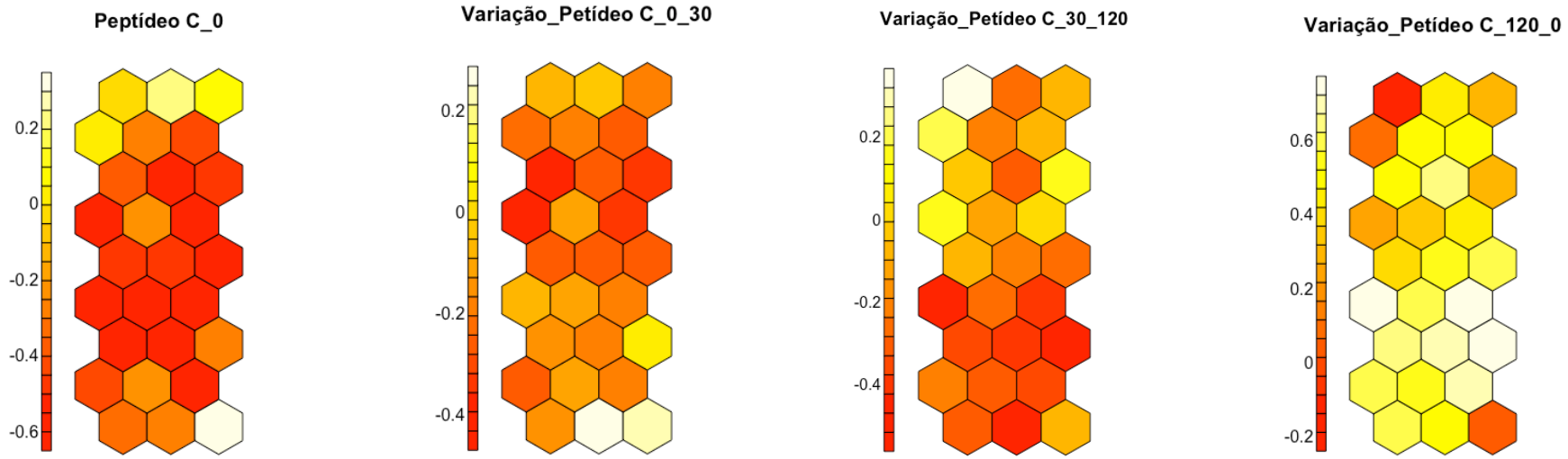
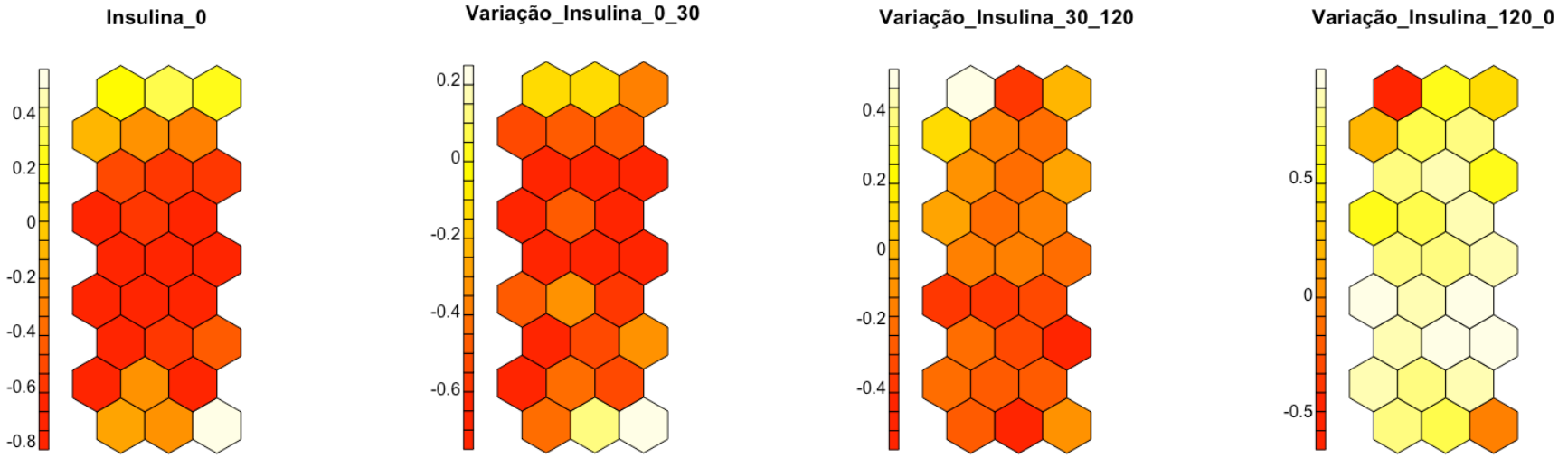


Figura 8. *Component planes* (variáveis normalizadas por grelha do *superSOM*): Grelha 1 – Glicémia; Grelha 2 – Peptídeo C.

Grelha 3



Grelha 4

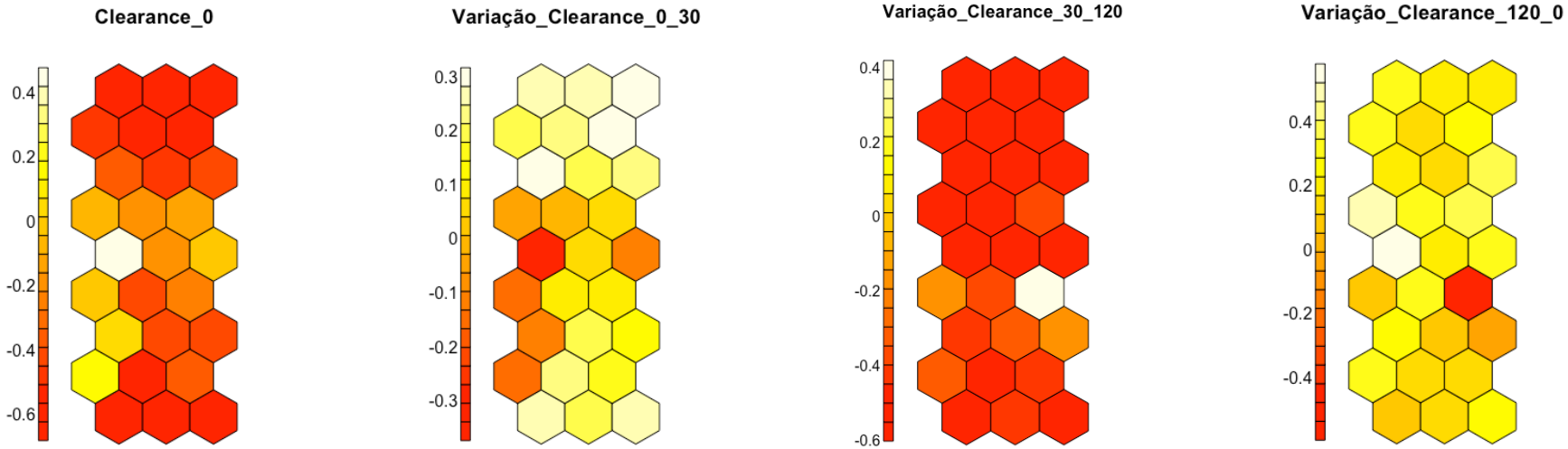
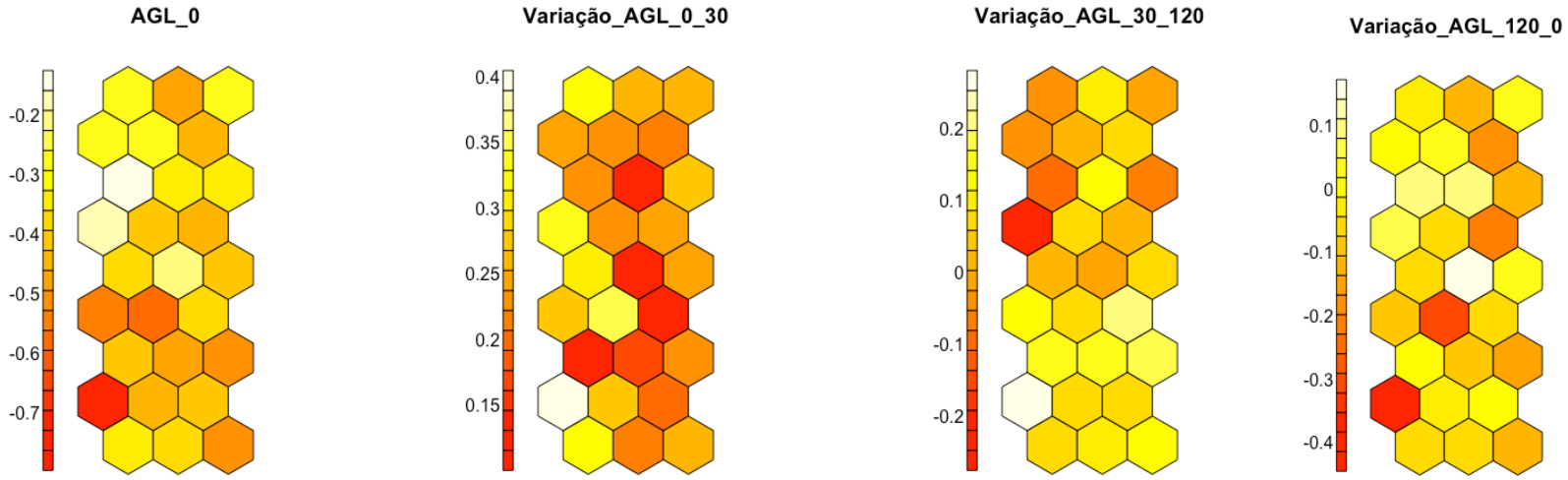


Figura 8. (cont.). *Component planes* (variáveis normalizadas por grelha do *superSOM*): Grelha 3 – Insulinemia; Grelha 4 – *Clearance* de insulina.

Grelha 5



Grelha 6

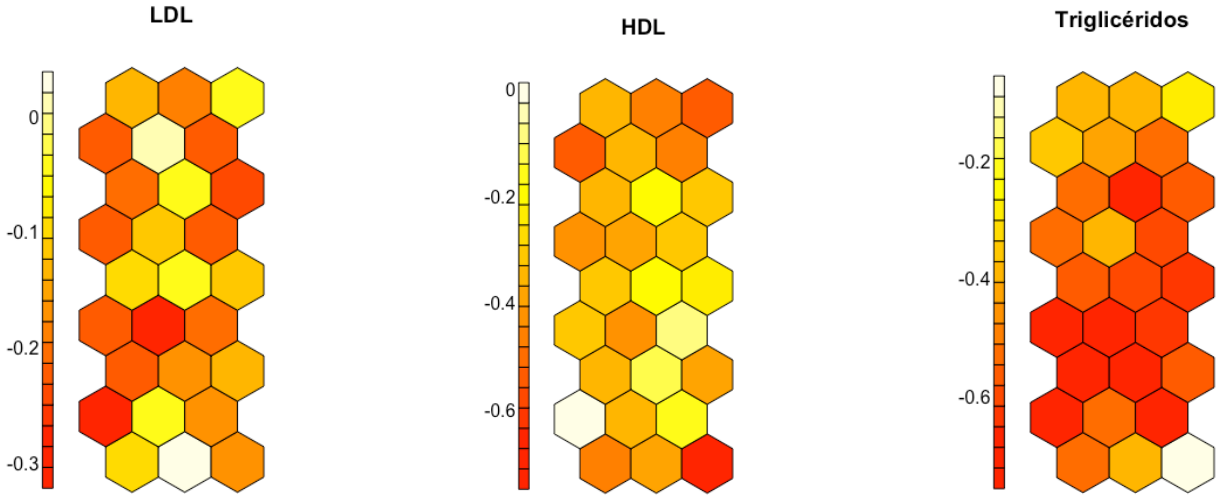
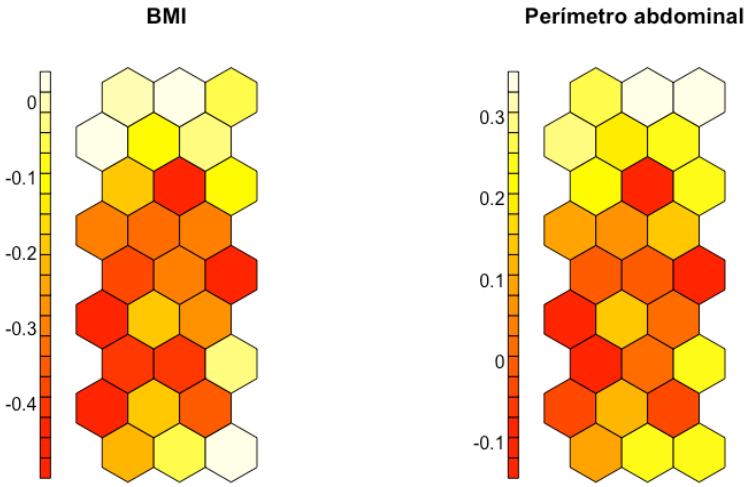


Figura 8. (cont.). *Component planes* (variáveis normalizadas por grelha do *superSOM*): Grelha 5 – Ácidos gordos livres (AGL); Grelha 6 – Outros lípidos.

Grelha 7



Grelha 8

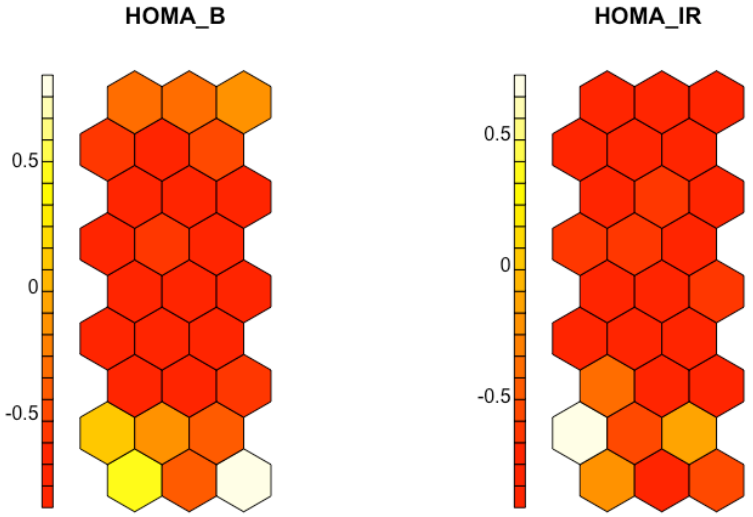


Figura 8. (cont.). *Component planes* (variáveis normalizadas por grelha do *superSOM*): Grelha 7 – Constituição corporal; Grelha 8 – Célula ß e *Disposition Index*

No que respeita à distribuição das diferentes classes de hiperglicemia (OMS/IDF) nas diferentes unidades, apesar de existirem alguns grupos puros (em que sencontra apenas uma classe), a maior parte das unidades agrupa indíviduos com diferentes classes de glicemia (Figura 9.). É de realçar, que embora em proporções diferentes, todas as unidades apresentam indivíduos normais. Dado que estes indíviduos são considerados semelhantes aos que com eles se agrupam, é interessante perceber que há indivíduos considerados como tendo normoglicemia, segundo a classificação atual, que são semelhantes a indivíduos considerados como tendo alterações da glicemia.

a) b)

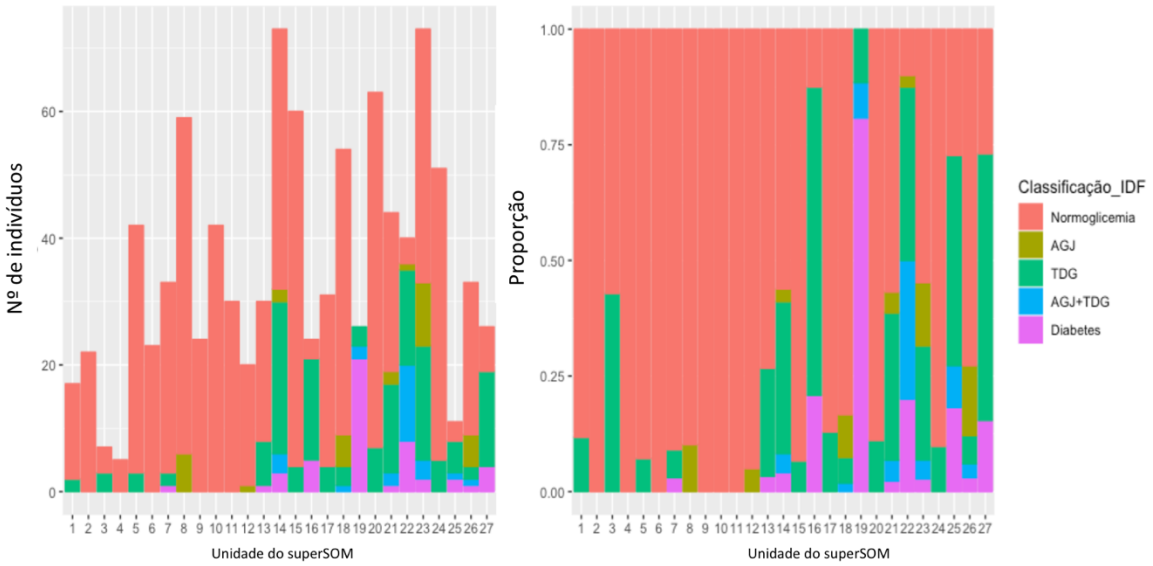


Figura 9. Distribuição das classes de hiperglicemia da OMS/IDF pelas diferentes unidades da grelha do *superSOM*: a) Número de indivíduos em cada grupo, classificados por classe da IDF; b) Proporção de indivíduos de cada classe da IDF, em cada grupo.

Estes resultados mostram, ainda, que grupos com maior proporção de indivíduos com normoglicemia, podem apresentar valores semelhantes, aos indivíduos com alterações da glicemia, em variáveis que representam fatores de risco e comorbilidades, como por exemplo o IMC , o perímetro abdominal e o colesterol total (Figura 10.). Assim, por exemplo, a unidade 11, que agrupa exclusivamente indivíduos normais, apresenta valores semelhantes em IMC e PA à unidade 16, que agrupa maior proporção de indivíduos com alterações da glicemia. Por outro lado, dois grupos com maior proporção de alterações da glicemia, podem mostrar diferentes valores dos mesmo parâmetros, com é o caso das unidades 16 e 22. Apesar da forte associação de determinados parâmetros às alterações da glicemia, estes não parecem ser nem necessários, nem suficientes às mesmas, e parecem caracterizar diferente fenótipos.

(a) (b)

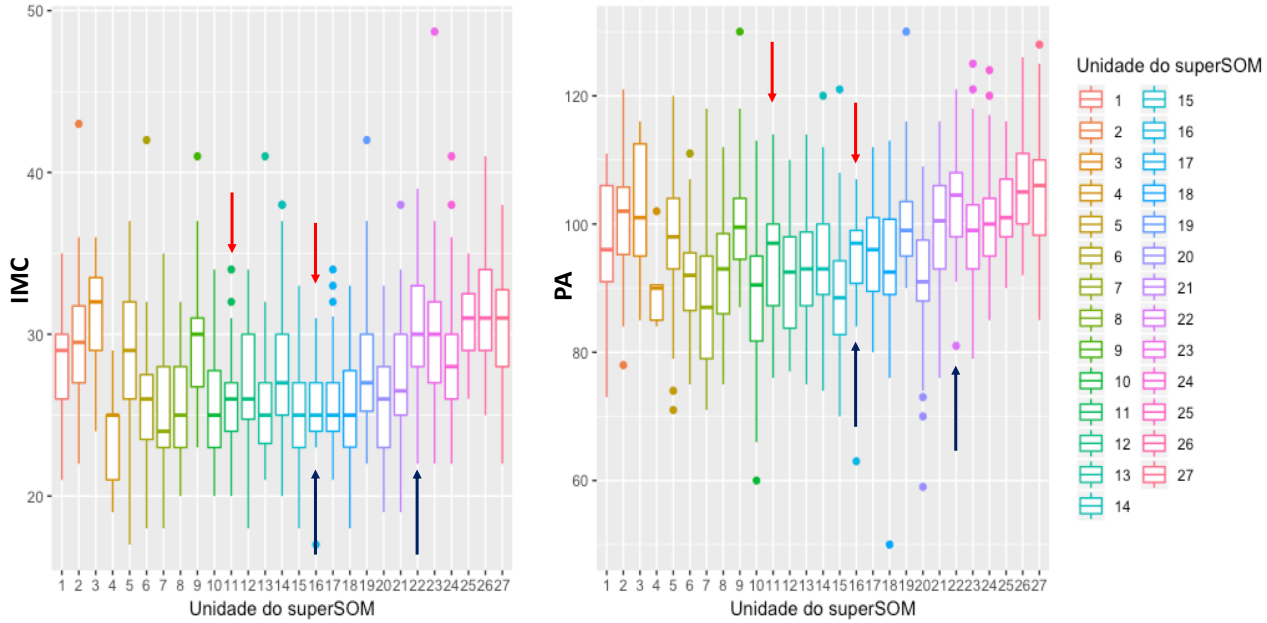


Figura 10. Distribuição do IMC (a) e PA (b) pelas diferentes unidades. Seta vermelha - unidades 11 e 16: apesar de a unidade 16 ter uma maior proporção de indivíduos com alteração da glicemia, comparativamente com a unidade 11, estes dois grupos são semelhantes no que se refere ao IMC e ao PA. Seta azul: apesar das unidades 16 e 22 terem uma preponderância de indivíduos com alterações da glicémia, têm valores de BMI e PA diferentes, apontando para diferentes fenótipos.

4.3. AGREGAÇÃO DAS UNIDADES DO SOM

A análise da matriz U, parece apontar para a existência de 9 *clusters*.

O gráfico da Figura 11. mostra o índice de DB para os diferentes números de *clusters,* quando aplicado o algoritmo *superSOM*, mostrando mínimos em 7 e 10 *clusters*. Este método dá uma indicação do número ótimo de *clusters*. No entanto, o cluster hierárquico permite-nos avaliar os resultados em diferentes pontos de corte, e optarmos pelo que faz mais sentido de acordo com a informação que é retirada.

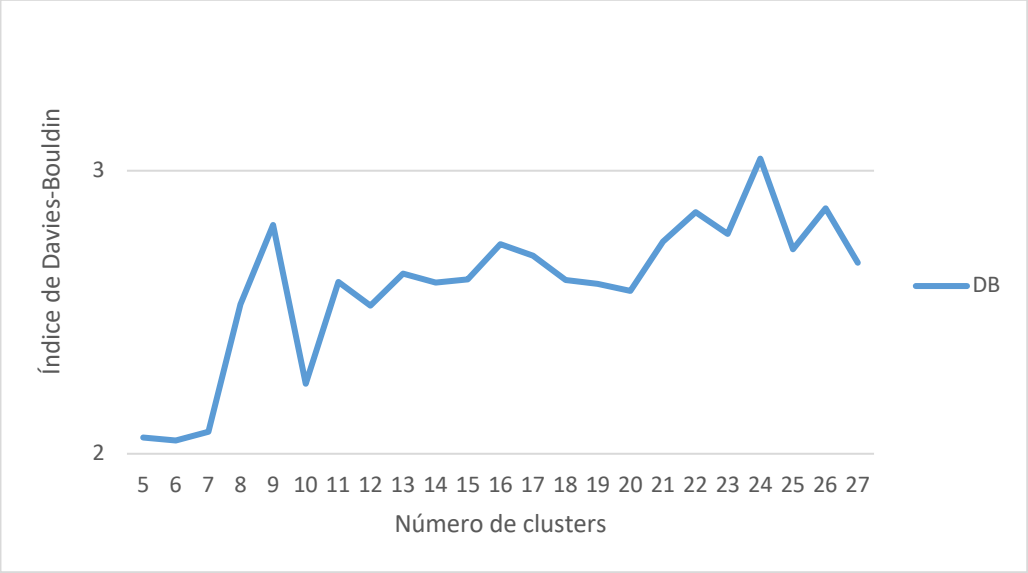


Figura 11. Cálculo do número ótimo de *clusters* pelo índice de Davies-Bouldin no *superSOM*

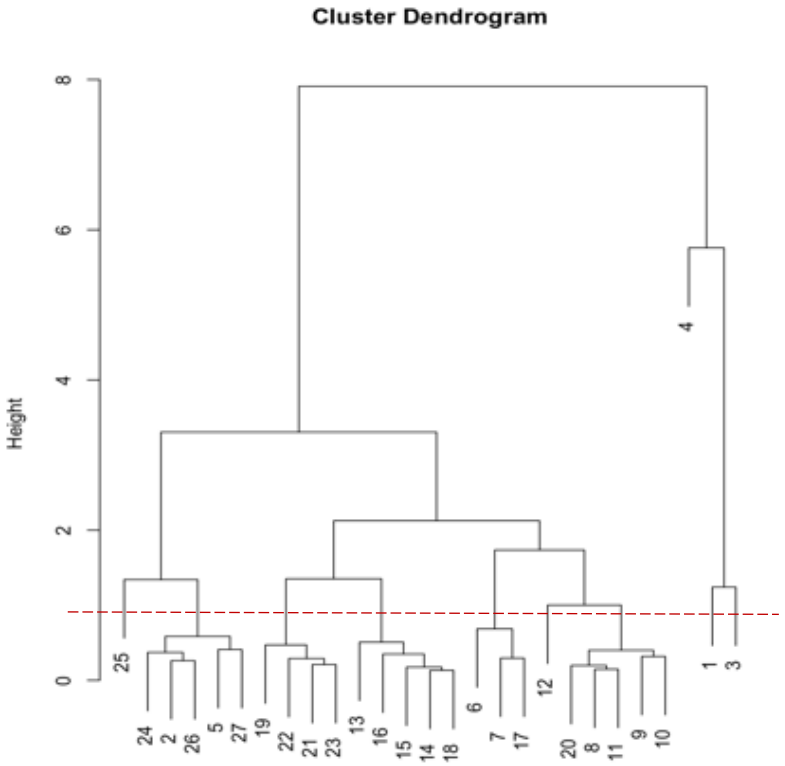


Figura 12. Dendograma do *cluster* hierárquico dos protótipos do *superSOM* (método Ward). Linha vermelha tracejada – ponto de corte considerado na formação dos *clusters.*

Iniciámos a análise com 7 *clusters*, de acordo com o método do "cotovelo" na análise do gráfico da Figura 12, e progredimos até a divisão num maior número de *clusters* não nos trazer

informação adicional relevante, relativamente aos perfis dos mesmos, o que se verificou na passagem de 10 para 11 *clusters*, pelo que o número escolhido foi 10, correspondendo ao ponto de corte do dendograma representado na Figura 12 (linha vermelha tracejada). Este corresponde ao 2º valor mais baixo no índice de DB quando calculado para o número de *clusters* sucessivos. O mapeamento dos indivíduos aos 10 *clusters* assim formados está representado na Figura 13.

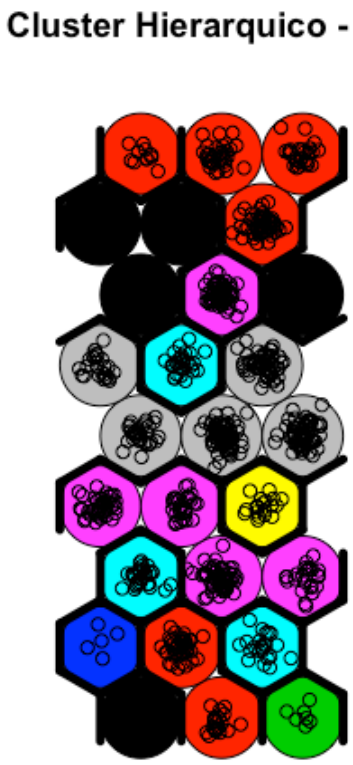


Figura 13. Mapeamento dos indivíduos ao *cluster* correspondente, de acordo com o ponto de corte do c*luster* hierárquico.

4.4. ANÁLISE E DISCUSSÃO DOS CLUSTERS

Os 10 *cluster* formados pela aplicação do método hierárquico (Ward) às unidades do *superSOM*, agrupam um total de 963 pessoas (580 mulheres), sublinhando-se o facto, de que não existem *clusters* puros para o género, embora esta distribuição seja heterogénea (Figura 14.) A distribuição do número de indivíduos também é heterogénea. Assim, os *clusters* 3, 7, 9 e 10 agrupam mais de 150 indivíduos em cada um, enquanto que os *clusters* 1, 2, 4, 5 e 8 agrupam menos de 25 pessoas (Tabela 6.).

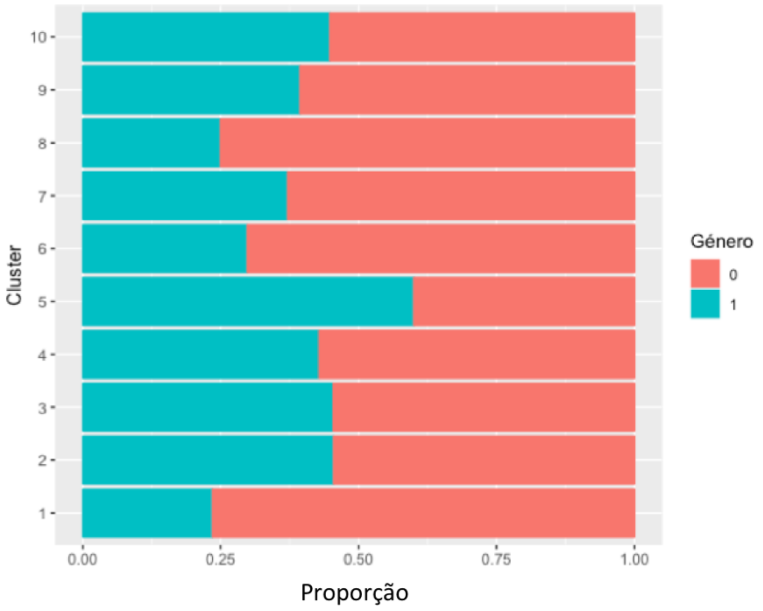


Figura 14. Distribuição do Género pelos *clusters*: 0 – feminino, 1 – masculino.

Tabela 6. Distribuição da população pelos 10 *clusters*

Cluster	Número de indivíduos	*Cluster*	Número de indivíduos
1	17	**6**	87
2	11	**7**	218
3	174	**8**	20
4	7	**9**	241
5	5	**10**	183

Da população avaliada, 73,2% têm normoglicémia (705 pessoas), 3,3% têm alteração da glicemia em jejum (32 pessoas), 15,8% têm tolerância diminuída à glucose (152 pessoas), 2,6% têm hiperglicemia intermedia mista (25 pessoas) e 5,1% têm diabetes (49 pessoas).

Tal como acontecia nas unidades do SOM, estes 10 *clusters* têm diferentes distribuições no que respeita às classes de hiperglicémia, defendidas pelas OMS/IDF (Figura 15.). Apesar de encontrarmos indivíduos com alterações da glicemia (de todas as classes) na maioria dos *clusters*, indicando que possa haver diferentes fenótipos das alterações da glicemia, no que respeita à distribuição das variáveis utilizadas no algoritmo, os *clusters* 9 e 10 agrupam, no conjunto, cerca de 73% destes indivíduos com hiperglicémia (hiperglicemia intermédia e diabetes) e 33% de indivíduos com normoglicémia. Curiosamente, os indivíduos com diabetes não formam um *cluster* único, mas

encontram-se distribuídos por 5 *clusters*. No entanto, cerca de 65% dos indivíduos com diabetes encontram-se no *cluster* 10, enquanto os restantes estão distribuídos por 4 *clusters*. O *cluster* 2, embora tenha apenas 11 indivíduos, apresenta apenas 3 com normoglicemia.

a) b)

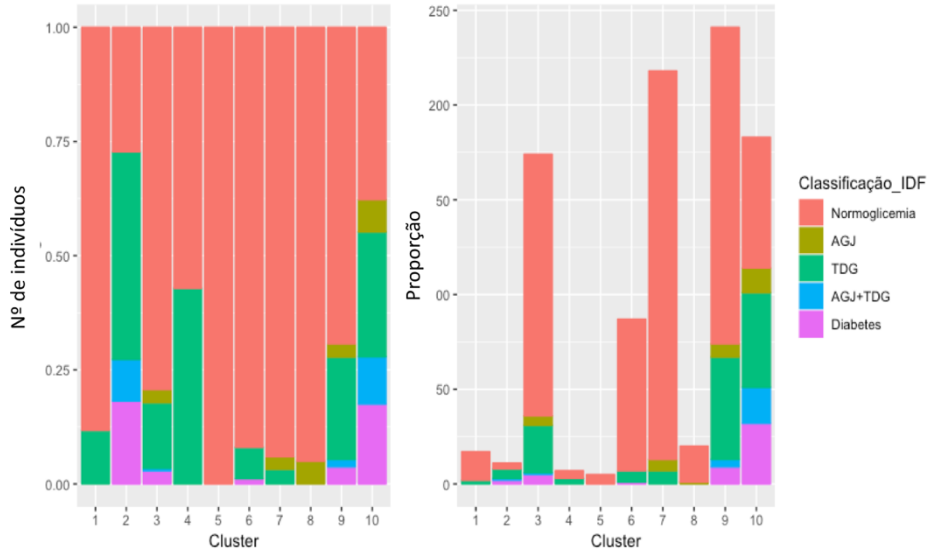


Figura 15. Distribuição das classes de hiperglicemia da OMS/IDF pelos 10 *clusters*: a) Proporção de indivíduos de cada classe em cada *cluster*; b) Número de indivíduos em cada *cluster*, distribuídos segundo a classificação da OMS/IDF.

Sublinha-se também, que de um modo geral, as diferentes classes de hiperglicemia intermédia se misturam, pelos diferentes grupos em que surgem, não parecendo haver uma diferenciação das mesmas, pela aplicação desta metodologia.

4.4.1. Perfil global dos *clusters*

De seguida, analisamos o perfil dos *clusters*, considerando as variáveis que lhes deram origem, de um modo global.

Da análise da Figura 16, que representa a mediana das variáveis medidas apenas em jejum nos diferentes *clusters*, parecem existir dois grandes grupos no que se refere aos índices antropométricos (PA e IMC) e à resistência à insulina. O grupo que engloba os *clusters* 1, 2, 3, 4 e 10, apresenta valores mais elevados destes parâmetros, do que o grupo constituído pelos *clusters* 5, 6, 7, 8 e 9. Observamos também, que a insulinoresistência se correlaciona com o IMC, e seguidamente com o nível de triglicéridos. Os *clusters* com valores mais elevados de mediana de insulinoresistência, IMC e PA contêm uma maior proporção de pessoas com alterações da glicemia (41,6% vs. 16,7%), o que está de acordo com o que tem sido exaustivamente descrito. No entanto ambos os grupos, com maior ou

menores valores de IMC e PA, bem como de IR, contêm indivíduos com normoglicémia e alterações da glicemia, pelo que estes parâmetros não são nem necessários, nem suficientes para o aparecimento de alterações da glicémia. Deste modo, os indivíduos com baixos valores de IMC, perímetro abdominal e resistência à insulina, não estão completamente protegidos destas alterações. De facto, dentro de cada *cluster*, não parece existir uma tendência para que os indivíduos com alterações da glicémia, tenham valores mais elevados dos mesmos parâmetros (Anexo E (a), (b) e (c)).

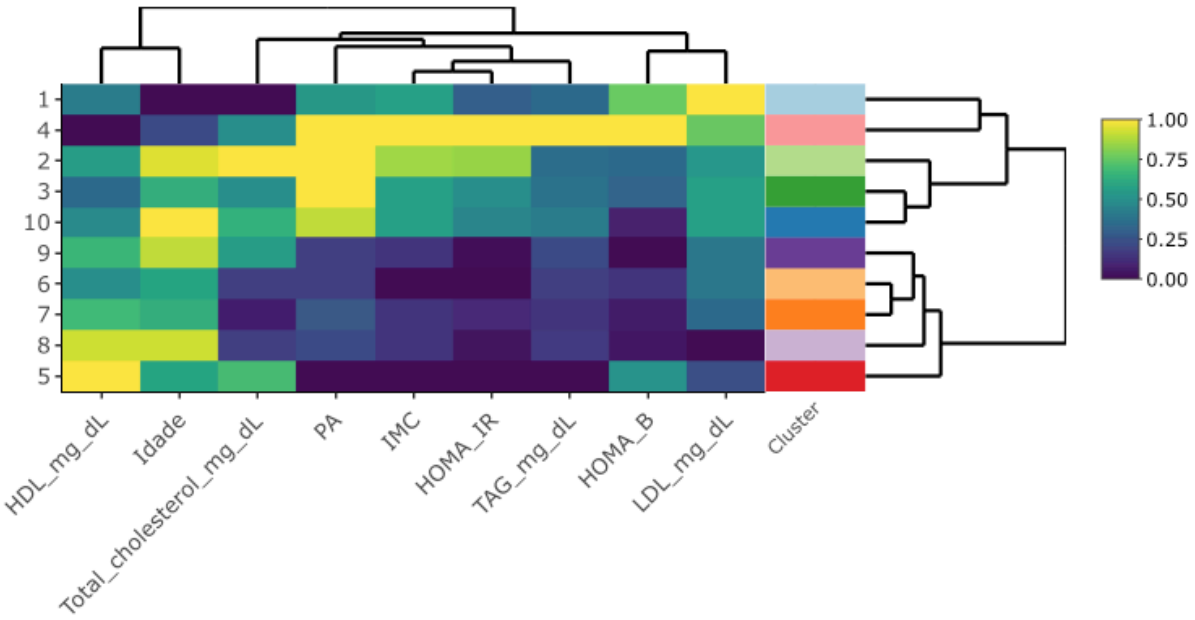


Figura 16. *Heatmap* considerando as variáveis medidas apenas em jejum (mediana), utilizadas para o perfil dos *clusters*

Os *clusters* com maior insulinoresistência, têm tendencialmente valores mais elevados de HOMA B. Como seria de esperar, nos *clusters* com menor resistência à insulina, apenas o que apresenta o valor mais baixo de HOMA B, agrupa indivíduos com diabetes. No entanto, à exceção do *cluster* 5, todos eles agrupam indivíduos classificados como tendo alterações da glicémia. Relativamente ao restante perfil lipídico, ambos os grupos apresentam *clusters* com valores mais ou menos elevados de HDL, sendo que as lipoproteínas de baixa densidade e o colesterol total parecem ter uma distribuição mais uniforme em todos os *clusters* (Anexo E (g) e (h)).

Ao avaliarmos os parâmetros que traduzem, quer os valores de glicemia medidos durante a PTGO, quer os valores globais de glicemia (AUC), verificamos que os *clusters* se podem distinguir pela distribuição de ambos. Os *clusters* 2 e 10, por exemplo, apresentam os valores mais elevados (Figura 17.), embora o *cluster* 2 apresente um valor menor da glicemia em jejum e maior aos 30' e 120', relativamente ao *cluster* 10. Curiosamente, o *cluster* 6, que agrupa um indivíduo com diabetes, surge entre os *clusters* com menores valor de glicemia. Este facto pode ser devido a tratar-se de um *outlier,* ou estarmos perante um indivíduo com outro tipo de diabetes, não identificado neste estudo. O *cluster* 5 parece ter menores valores de glicemia aos 0 e 30' da PTGO. Verificamos também, que

dentro dos *clusters* com maiores valores de glicemia (2, 10, 9 e 3), podemos encontrar grupos com diferentes fenótipos no que respeita aos parâmetros antropométricos, perfil lipídico e de resistência à insulina.

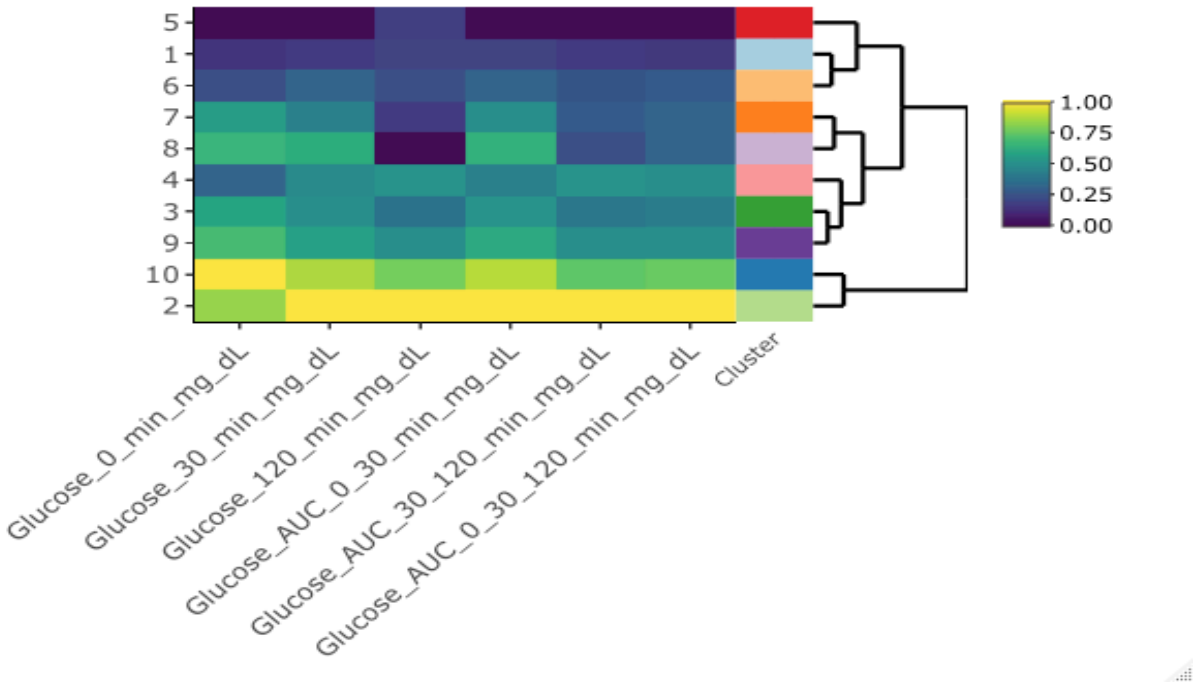


Figura 17. *Heatmap* considerando os parâmetros referentes à glicemia (mediana), dos 10 *clusters*. São considerados, para além dos valores da glicemia dos 0', 30' e 120' da PTGO, as áreas sob a curva (AUC) dos diferentes intervalos.

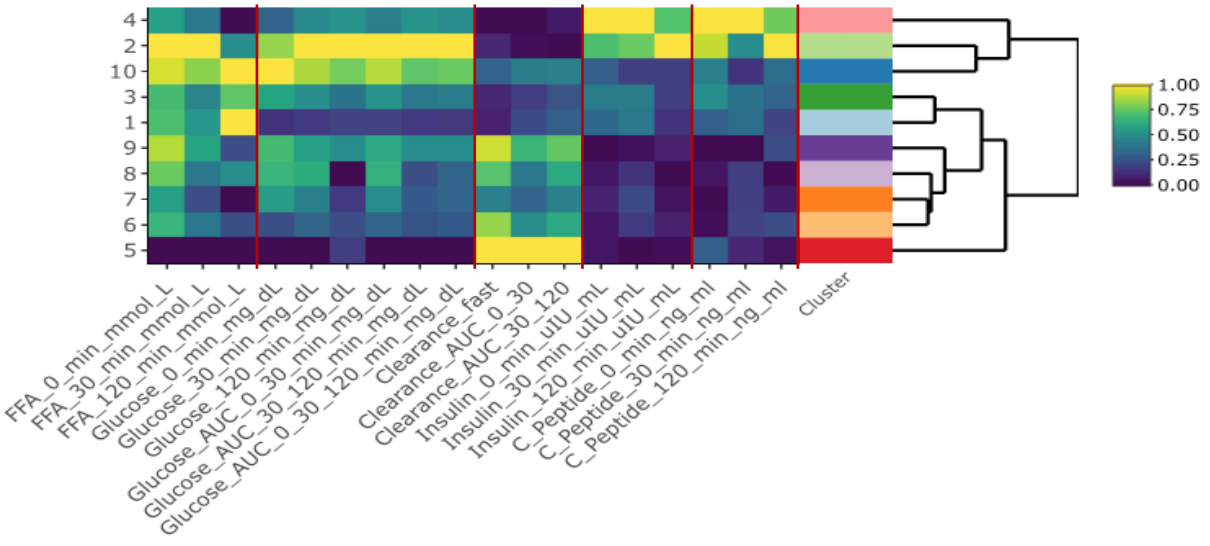


Figura 18. *Heatmap* considerando os parâmetros medidos durante a PTGO (mediana), dos 10 *clusters*.

A Figura. 18 é um *heatmap* da mediana das variáveis medidas durante a PTGO, dos 10 *clusters* encontrados pelo algoritmo. Mais uma vez, podemos verificar diferentes perfis, de ácidos gordos livres, de *clearance* de insulina, de insulina e peptídeo C. Tanto nos *clusters* com menores valores de glicemia (exemplo 1 e 7), como nos *clusters* com maiores valores de glicemia (exemplo 9 e 10), podemos verificar diferentes perfis de *clearance* de insulina, de ácidos gordos livres, bem como de insulina e peptídeo C. Novamente, parecem existir diferentes fenótipos quer dos indivíduos com alterações da glicemia, quer dos indivíduos com normoglicemia.

Para além dos níveis absolutos de ácidos gordos livres, *clearance* de insulina, insulinemia e níveis de peptídeo C no sangue venoso durante a PTGO, podemos verificar que existem diferentes perfis, no que respeita à evolução destes parâmetros nos *clusters* avaliados. O mesmo acontece com os valores da glicémia.

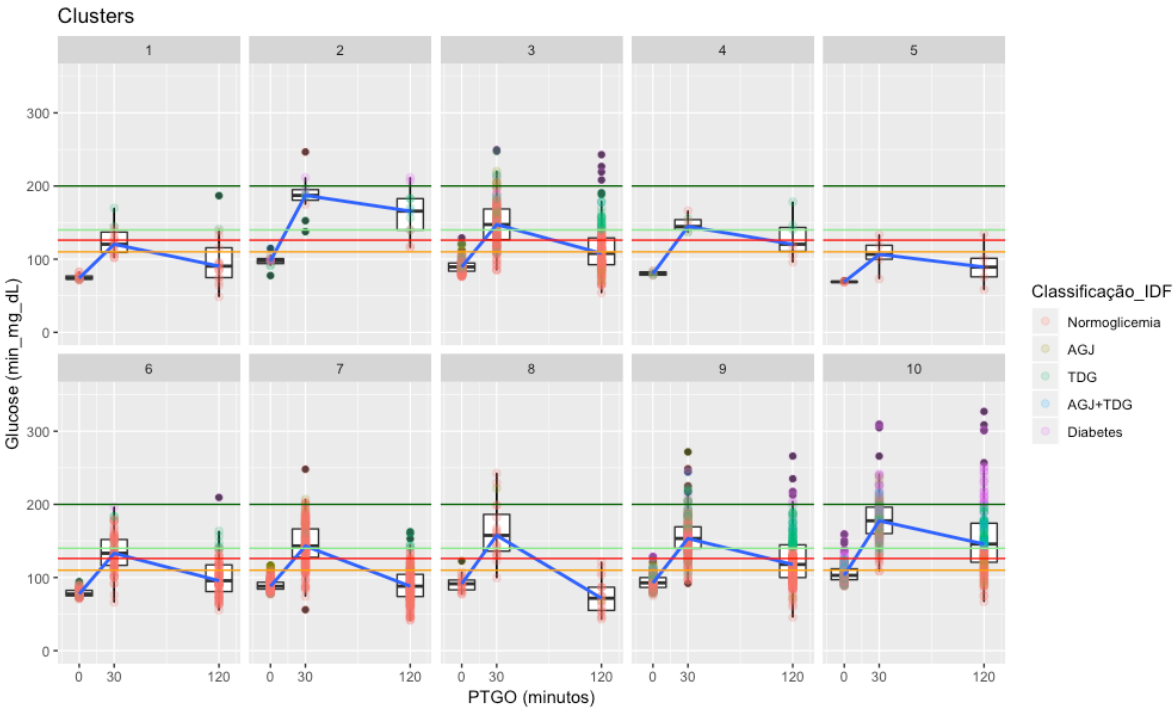


Figura 19. Perfil da glicémia durante a PTGO de cada *cluster*. As linhas horizontais indicam os limites de pré-diabetes e diabetes para a glicemia em jejum (amarelo e vermelho) e aos 120' da PTGO (verde claro e verde escuro), respetivamente.

No que respeita à evolução dos valores de glicémia durante a PTGO (Figura 19.), parecem existir, essencialmente três perfis. A glicémia em jejum é recuperada aos 120', depois da subida aos 30'. A glicémia aos 120' é mais baixa que a glicémia em jejum, sendo este perfil mais raro. E por último, a glicémia aos 120' não diminui, permanecendo em valores próximos da glicémia aos 30'.

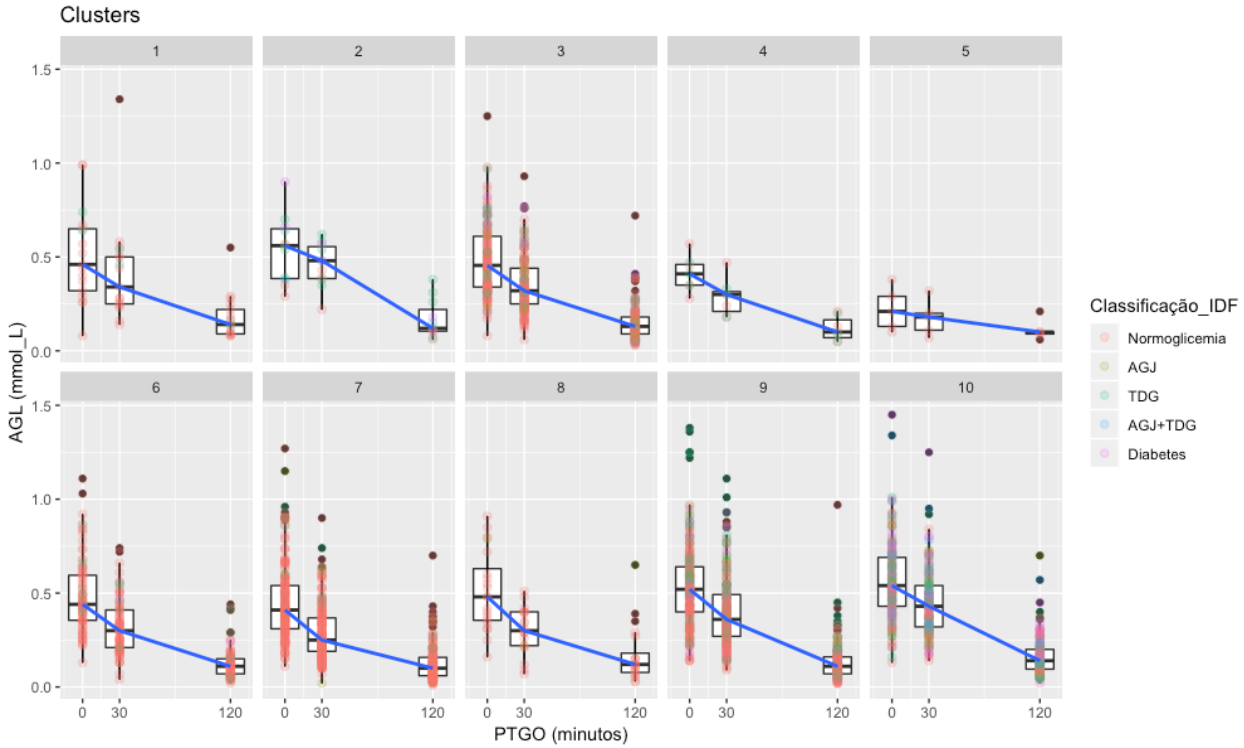


Figura 20. Perfil de ácidos gordos livres durante a PTGO de cada *cluster*

Os níveis de ácidos gordos livres (Figura 20), podem diminuir de um modo mais ou menos acentuado dos 0' aos 30', e dos 30' aos 120' da PTGO, sendo que o declive desta última relação, parece dever-se mais aos níveis dos ácidos gordos aos 0' e 30', já que aos 120' eles são relativamente semelhantes em todos os grupos.

Quanto ao perfil da *clearance* da insulina durante a PTGO (Figura 21.), parecem existir duas diferenças mais evidentes. O declive da diminuição da *clearance* dos 0' para os 30', que pode ser mais ou menos acentuado, e poder haver dos 30' para os 120', uma diminuição adicional da *clearance*, ou esta manter-se constante.

Na Figura 22., que representa a evolução dos níveis de peptídeo C durante a PTGO, ressaltam três perfis, respeitantes à diferença de valores dos 30' aos 120'. Em alguns *clusters* os níveis de peptídeo C tendem a manter-se iguais ou tendencialmente menores, enquanto que em outros tendem a manter-se semelhantes ou tendencialmente mais elevados, e ainda noutros eles elevam-se claramente. Como seria expectável, os *cluster*s em que tendencialmente o peptídeo C diminui, agrupa principalmente pessoas com normoglicemia. Dado que a glicemia aos 30' é de um modo geral mais baixa nestes *clusters*, e que diminui aos 120' para valores considerados de normoglicemia, a secreção de insulina baixa.

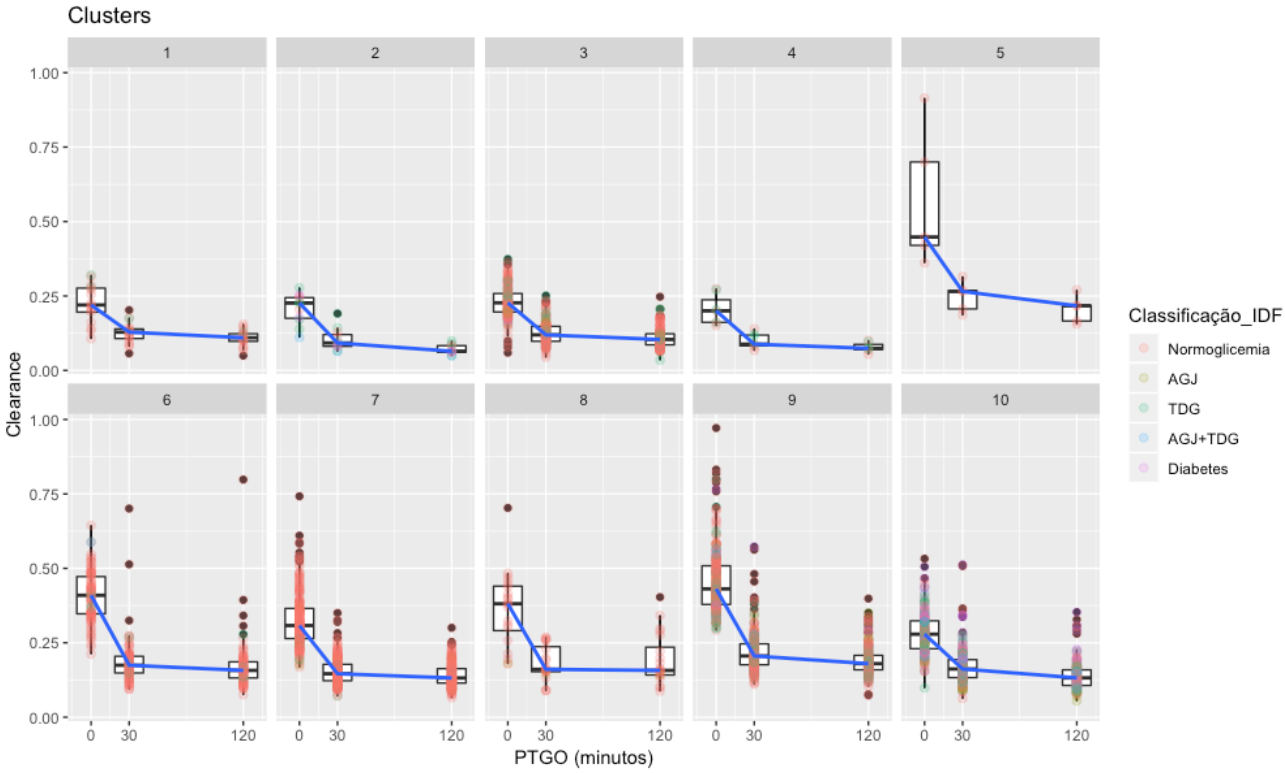


Figura 21. Perfil de *clearance* de insulina durante a PTGO de cada *cluster*

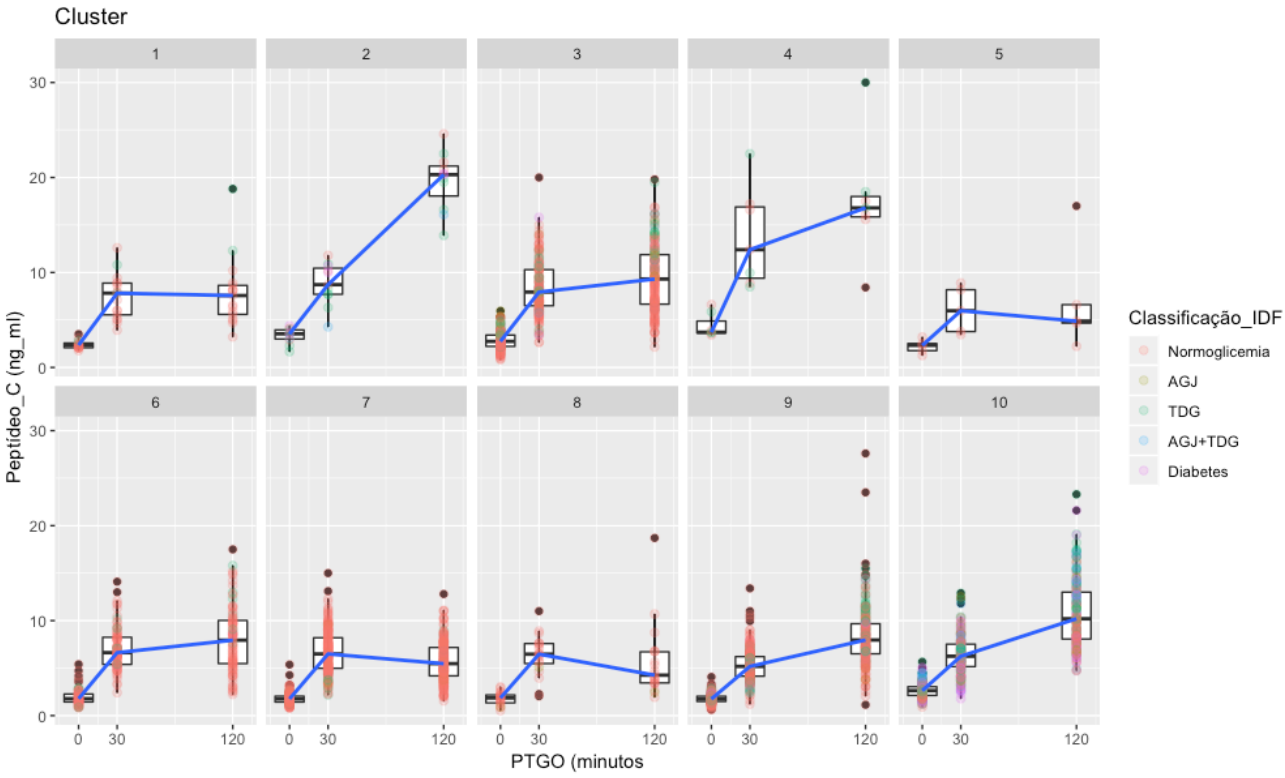


Figura 22. Perfil de níveis de peptídeo C no sangue venoso durante a PTGO de cada *cluster*

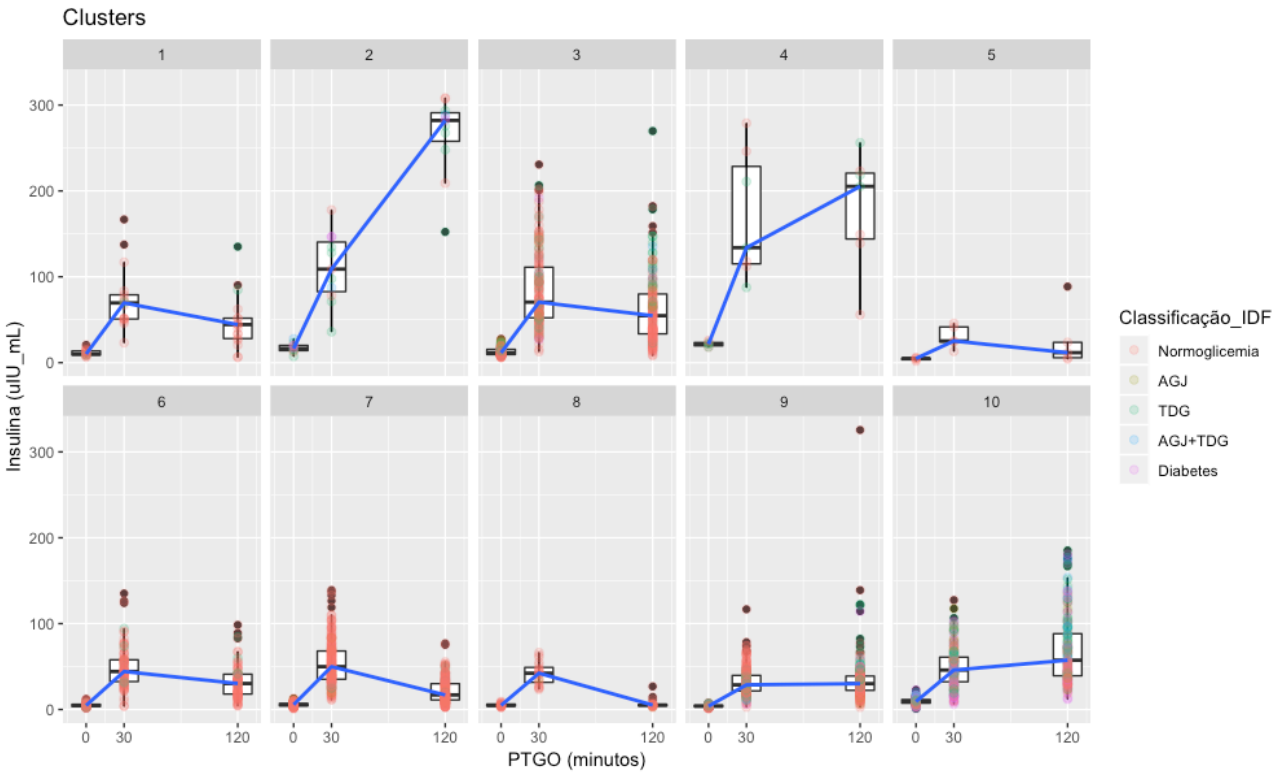


Figura 23. Perfil de níveis de insulinemia durante a PTGO de cada *cluster*

Finalmente, e no que se refere à evolução da insulinemia durante a PTGO, existem grupos com uma subida mais marcada dos 0' aos 30' do que outros. Também, identificamos três perfis, relativamente à evolução destes níveis dos 30' aos 120'. Em alguns *clusters* estes níveis tendem a diminuir, em outros mantêm-se relativamente constantes, e ainda em outros estes níveis sobem francamente.

Pela multiplicidade das curvas dos parâmetros avaliados, e das suas conjugações, podemos perceber, que de facto, as alterações da glicémia estão longe de derivar de alterações fisiológicas monótonas.

4.4.2. Perfil individual dos *clusters*

Analisaremos de seguida, cada *cluster* individualmente. Para clareza da descrição, iremos descrever cada grupo pela ordem apresentada no heatmap referente às glicemias, dos valores mais baixos para os mais altos (Figura. 17). Os nomes dos *clusters* são dados com base nas tendências dos seus perfis, referentes à resistência à insulina e à função da célula ß, e estão resumidos na tabela 7.

Tabela 7. Resumo dos perfis dos *clusters*, referentes à resistência à insulina versus a função da célula b. Vermelho: *Clusters* que contêm indivíduos classificados como tendo diabetes segundo a IDF.

Resistência à insulina (HOMA_IR) / Função da célula b (HOMA_B)	Baixa	Moderada	Alta
Alta	5	1	4
Moderada	6	3	2
Baixa	7,8,9	10	

Cluster 5 – Resistência à insulina baixa, elevada função da célula ß.

Agrupa apenas 5 indvíduos, todos com normoglicemia. Tem baixos valores de medidas antropométricas, resistência à insulina, triglicérideos e ácidos gordos livres, e mais elevado nível de HDL. Têm um elevado valor de *clearance* de insulina. Os níves de glucose são baixos durante toda a PTGO, o mesmo acontecendo com os níveis de insulina e peptídeo C, que ao contrário do que acontece na maioria dos *clusters*, baixa ligeiramente dos 30' aos 120'.

Cluster 1 – Resistência à insulina moderada, elevada função da célula ß.

O *cluster* 1 agrupa 17 indivíduos, dos quais 15 têm normoglicémia e 2 dois têm TDG. Apesar dos valores de glicémia serem ligeiramente mais elevados do que no *cluster* 5, aos 0' e 30' da PTGO, este grupo é o 2º com glicémias mais baixas. No entanto, ao contrário do *cluster* 5, pertence ao grupo com valor mediano de medidas antropométricas, resistência à insulina e triglicéridos mais elevados. Também os valores de ácidos gordos livres são mais altos do que no *cluster* descrito anteriormente. Apresenta baixos valores de *clerance* de insulina e valores de HDL intermédios. Quer os níveis de petpídeo C, quer os níveis de insulina sobem dos 0' aos 30'. No entanto o nível de petídeo C mantém-se constante dos 30 aos 120', enquanto que o nível de insulina baixa neste período.

Cluster 6 – Resistência à insulina baixa, moderada função da célula ß.

O *cluster* 6 contém 87 indivíduos, 7 com alterações da glicémia (seis classificados como TDG, e um classificado com diabetes, também devido ao valor da glicémia aos 120'). Apesar de conter um indivíduo classificado com diabetes, este é dos grupos com glicémia mais baixas, quer durante a PTGO, quer nos valores globais (AUC). Este grupo, tal como o *cluster* 5, pertence ao grupo com baixas

medidas antropométricas, resistência à insulina e triglicéridos. No entanto a função da célula ß é inferior ao *cluster* 5. Tem valores de ácidos gordos livres elevados em jejum, mas diminuem aos 30'. Têm uma *clearance* de insulina elevada em jejum, e que diminui de forma evidente aos 30'. O peptídeo C dos 30' aos 120' sobe tendencialmente, enquanto que a insulina tendencialmente desce.

Cluster 7 – Resistência à insulina baixa, baixa função da célula ß.

O *cluster* 7 apresenta maiores valores de glicémia em jejum e aos 30' da PTGO, do que os descritos anteriormente, embora aos 120' os níveis de glicemia sejam semelhantes. Agrupa 218 pessoas, 13 com hiperglicemia intermedia isolada (6%). Este *cluster* tem baixas medidas antropométricas, resistência à insulina e triglicéridos. Semelhante ao *cluster* 6 em vários aspectos, diferencia-se deste, de modo mais evidente, por ter um valor de HOMA B mais elevado, uma *clearance* de insulina em jejum mais baixa, com menor diminuição dos 0' aos 30'.

Cluster 8 – Resistência à insulina baixa, baixa função da célula ß. Glicémia aos 120' menor que em jejum.

Este grupo representa um perfil mais raro, agrupando 20 indivíduos, dos quais 95% têm normoglicémia. Este é o único grupo em que a glicemia aos 120', tendem a ser mais baixas que a glicémia em jejum, embora os valores em jejum e aos 30' da PTGO, sejam semelhantes ao *cluster* 7. Pertence também ao grupo com parâmetros antropométricos, resistência à insulina e triglicéridos mais baixos com HDL mais elevado que a maioria dos outros grupos. Apesar destes valores de glicemia aos 120', o *cluster* 8 apresenta um HOMA IR tendencialmente maior que o *cluster* 5, enquanto que o HOMA B é francamente menor. Com um perfil de insulinemia semelhante ao do *cluster* 7 (embora com valores absolutos tendencilamente menores), bem como de peptídeo C, diferencia-se deste, por apresentar um perfil de *clearance* semelhante ao do *cluster* 6, com maior *clearance* de insulina em jejum e aos 30' da PTGO, apesar de diminuir de modo evidente aos 30'.

Cluster 4 – Resistência à insulina alta, elevada função da célula ß.

Apresenta apenas 7 indíviduos, dos quais 4 tem normoglicémia e 3 TDG. Este *cluster* surge com os extremos de valores de HOMA B. Pode dever-se a erro de medição, ou representar o extremo de função da célula beta. Apesar dos valores extremos de HOMA B, este *cluster* é o que apresenta valores de HOMA IR mais elevados, juntamente com o *cluster* 2 (descrito mais á frente). Assim, apesar de não conter indivíduos com diabetes, e todos terem glicémia normal em jejum, cerca de metade têm TDG. Este *cluster* pertence ao grupo de elevadas medidas antropométricas, bem como apresenta o mais elevado valor de mediana de triglicéridos. Tem uma *clearance* de insulina baixa em jejum, e que diminui fracamente aos 30', comparativamente com outros grupos. No que respeita à insulinémia e peptídeo C, verifica-se que o seu perfil é semelhante ao do *cluster* 2, em que os níveis de ambas as proteínas sobem de modo marcado dos 30' aos 120'. Ao contrário do *cluster* 2, os indíviduos são mais jovens.

Cluster 3 – Resistência à insulina moderada, moderada função da célula ß.

O *cluster* 3, com glicémias mais elevadas durante toda a PTGO, agrupa 174 indivíduos, 36 com alterações da glicémia (21%), 5 com diabetes. Com valores de parâmetros antropométricos e triglicéridos elevados, apresenta um HOMA IR que é, embora também elevado, mais baixo que os

clusters 2 e 4. Ainda, os perfis de petídeo C e de insulina são em tudo diferentes do *cluster* 4, dado que estes valores mantém-se relativamente constantes dos 30' aos 120'.

Cluster 9 – Resistência à insulina baixa, baixa função da célula ß.

Agrupa 241 pessoas, das quais 74 têm alterações da glicémia (30,7%), 9 com diabetes. O *cluster* 9 apresenta um perfil de glicémias semelhante ao *cluster* 3, tendo no entanto, menores valores de mediana de medidas antropométricas, bem como de resistência à insulina e de triglicéridos, valores estes que se aproximam mais dos *clusters* 5, 6, 7 e 8. O mesmo acontece com o valor de HOMA B. Os valores de petídeo C são semelhantes ao *cluster* 3 em jejum, subindo menos aos 30' e mais aos 120'. Os valores de *clearance* são mais elevados, e os valores de insulina menores que o *cluster* 3.

Cluster 10 – Resistência à insulina moderada, baixa função da célula ß.

O *cluster* 10 tem elevados valores de glicémia em toda a PTGO, sendo o que apresenta valores mais elevados da mediana da glicemia em jejum. Agrupa 183 pessoas, das quais 114 (62%) têm alterações da glicémia, 32 com diabetes (17,4%). Apesar de pertencer ao grupo com parâmetros antropométricos, triglicéridos mais elevados, semelhantes aos *clusters* 1 e 3, os valores de HOMA B são baixos. Apresenta uma evolução de *clearance* de insulina durante a PTGO semelhante ao *cluster* 9, embora com valores absolutos mais baixos, no entanto o perfil de petpídeo C e de insulina não se diferenciem de modo evidente do mesmo.

Cluster 2 – Resistência à insulina alta, com moderada função de célula ß.

O *cluster* 2 tem também, à semelhança do *cluster* 10, elevados níveis de glicémia, durante toda a PTGO, mais pronunciados aos 30' e 120'. Agrupa apenas 11 pessoas, no entanto 8 têm alterações da glicémia, duas com diabetes. É semelhante ao *cluster* 10, no que respeita aos parâmetros antropométricos e de níveis de triglicéridos. Apesar de ter um HOMA IR mais elevado, o HOMA B é também mais elevado que neste *cluster*. Os perfis de *clearance* de insulina, peptídeo C e de insulinemia assemelham-se aos do *cluster* 4, nomeadamente no que respeita ao aumento evidente do peptídeo C e de insulina dos 30' aos 120' da PTGO.

4.4.3. Discussão

A utilização do algoritmo *superSOM* permitiu formar *clusters* com um elevado número de variáveis incluídas na modelação (27). Mais, permitiu agrupar as variáveis de acordo com a sua natureza, criando grelhas com diferentes significados, do ponto de vista clínico. Isto não seria possível, se se aplicasse o algoritmo SOM, apenas com uma grelha, que consideraria todas as variáveis simultaneamente, sem olhar ao seu significado. O peso destas variáveis pode ser controlado de acordo com o desejado, possibilitando uma exploração mais dirigida de acordo com o conhecimento atual.

Apesar dos objetos incluídos numa das unidades, mostrarem maiores distâncias dos mesmos ao seu centro, a distribuição destas distâncias é relativamente uniforme nas restantes unidades.

Podia ter sido utilizado um algoritmo *superSOM* com 10 unidades unicamente, o que evitaria a necessidade de aplicar o algoritmo hierárquico, em segundo lugar. Dada a natureza exploratória deste trabalho, do ponto de vista do conhecimento científico, a metodologia aqui utilizada apoiou e facilitou

a avaliação dos dados no que respeita à definição do número de *clusters*. No entanto a primeira estratégia será certamente considerada em trabalhos futuros.

Nesta análise, consideramos indivíduos com normoglicemia juntamente com os que têm hiperglicemia, ao contrário do trabalho publicado por Ahlqvist *et al*. Assim, apesar de avaliarmos como se agrupam os indivíduos com as diferentes classes de hiperglicemia consideradas atualmente, estas, em nada interferiram na modelação.

Os 10 *clusters* identificados neste trabalho mostram uma distribuição heterogénea do número de indivíduos. Esta heterogeneidade poderá corresponder à maior prevalência de determinados perfis (*clusters* 3,6,7,9 e 10), e à raridade de outros (*clusters* 1,2,4,5 e 8).

Estando agrupados por semelhança de vários parâmetros, considerados importantes na fisiopatologia da hiperglicemia, seria de esperar *cluster* puros para indivíduos com normoglicemia, *cluster* puros para indivíduos com disglicemia, e uma eventual classe impura, de transição entre as anteriores. No entanto, isto não acontece. O único *cluster* completamente puro, agrupa apenas 5 indivíduos com normoglicemia (*cluster* 5). Todos os outros grupos, contêm pessoas com alterações da glicemia, das diferentes classes, em diferentes proporções. Por exemplo, os *clusters* 2 e 10, agrupam maior proporção de indivíduos com alterações da glicemia, relativamente aos outros, contendo no seu conjunto 69% dos indivíduos com diabetes. Por outro lado, encontramos indivíduos com normoglicemia em todos os *clusters*, também em diferentes proporções.

Estes resultados sugerem que as alterações da glicemia podem surgir nos diferentes *clusters*. Apesar de haver uma maior propensão de uns grupos relativamente a outros, o facto de poderem surgir em qualquer um deles, levanta a questão de haver um fator pessoal (informação genética), que pode torná-los mais ou menos suscetíveis. Tal, poderia explicar, o aumento da suscetibilidade, dando origem ao aparecimento de pessoas com alterações, em grupos com preponderância de indivíduos com normoglicemia, bem como a proteção, explicando que existam indivíduos com normoglicemia em *clusters*, que agrupam maioritariamente indivíduos com alterações da glicemia.

Outra questão que pode ser colocada, é se os indivíduos com alterações da glicemia que se encontram agrupados em *clusters* com maior número de indivíduos com normoglicemia, se encontram realmente nas classes de pré-doença, ou se se podem tratar de indivíduos que, apesar das glicemias, não irão progredir no espectro da doença, ou desenvolver complicações. Do mesmo modo, terão os indivíduos com normoglicemia, agrupados nos *clusters* 2 e 10, maior risco de progressão na doença?

Da análise das características dos *clusters*, no que respeita aos diferentes parâmetros avaliados, percebemos que existem diferentes fenótipos, quer para os indivíduos com alterações da glicemia, quer para os indivíduos com normoglicemia. Estes fenótipos não parecem ser suficientes para o aparecimento da doença, apesar do considerável aumento de risco que possam imprimir. Apesar de, por exemplo, o elevado IMC, de perímetro abdominal e de resistência à insulina, aumentar o risco de existir alteração da glicemia, existem nestes *clusters* indivíduos com normoglicemia, bem como existem indivíduos com alterações, em *clusters* em que estes parâmetros são mais baixos. No entanto, surgem outras questões: Estes fenótipos são próprios de cada indivíduo, ou são pontuais, devendo-se, por exemplo, ao estilo de vida? Cada indivíduo evolui dentro do seu *cluster,* à medida que envelhece ou aumenta o IMC, ou ao contrário, salta de *cluster* em *cluster*? Apesar de haver diferenças

de idade, bem como de IMC entre os *clusters*, estas diferenças não são francas na maioria dos *clusters*, e os seus limites são bastante latos, em quase todos os grupos.

Nos grupos com glicemias mais elevadas (4,3,9,10 e 2), percebemos que os níveis de peptídeo C durante a PTGO, sobe dos 30' aos 120', como seria expectável, ou em alguns grupos se mantém constante (*cluster* 3). O *cluster* 3, difere por exemplo do *cluster* 1, em que os níveis de peptídeo C se mantêm constantes também, sem haver glicemias elevadas. Assim, enquanto que em alguns *clusters* a resistência à insulina poder ser fundamental no processo fisiopatológico (*cluster* 2), noutros, a insulinodeficiência parece prevalecer (*cluster* 9). A deficiência de insulina encontrada no *cluster* 9, caracterizado por uma baixa resistência à insulina, poderá estar presente desde cedo.

Observamos que os indivíduos, que produzem mais peptídeo C durante a PTGO, conseguem maiores níveis de insulina, quando a sua *clearance* é mais baixa. Apesar dos níveis baixos de *clearance* de insulina hepática serem relacionados com o dismetabolismo, a sua diminuição durante a PTGO pode ser fundamental para atingir os níveis de insulina necessários ao controlo da glicemia. Por outro lado, o *cluster* 9, que agrupa indivíduos com elevada *clearance*, apresenta uma considerável proporção de pessoas com alterações da glicemia.

Ainda, os *clusters* encontrados não mostraram, na sua globalidade, diferenças evidentes, no que se refere ao perfil em lípidos. Ao darmos pesos menores às variáveis que representam o perfil lipídico na grelha do algoritmo, previligiámos os perfil da glicémia, insulina, peptídeo C e clerance, bem como a resistência à insulina e o *disposition index*. Os resultados sugerem, que as diferenças do perfil lípidico não derivam dos perfis aqui privilegiados, podendo estar alterados em indivíduos com hiperglicemia, mas também em pessoas sem alterações da glicemia. É possível, que progredindo na desagregação, se encontrem *clusters* com maiores diferenças no perfil lipídico. Esta análise exige uma maior quantidade de dados e será alvo de estudos futuros.

Por último os resultados deste trabalho mostram uma oportunidade relevante para diferenciar o tratamento de acordo com o perfil dos *clusters*. Por exemplo, tratar um indivíduo pertencente ao *cluster* 9 com um fármaco da classe das Tiazolidenidionas, medicamento que tem como principal objetivo diminuir a resistência à insulina, pode ser ineficaz, enquanto que utilizar um medicamento que tenha como efeito principal aumentar a secreção de insulina, como é o caso do grupo das Sulfonilureias, poderá ser uma melhor opção. Curiosamente, os resultados vão mais além, mostrando também possíveis caminhos no que respeita, à não menos importante, questão da profilaxia primária. Prevenir o aumento da epidemia da diabetes, passa por identificar melhor as pessoas em risco. É possível, que os indiíduos com hiperglicemia intermedia, e até normoglicemia, pertencentes por exemplo ao *cluster* 10, tenham um risco maior de progressão no espectro da doença, ou da presença de complicações da diabetes, comparativamente com os que se encontram no *cluster* 7. A confirmação destas hipóteses, no que respeita à terapêutica e à prevenção, é premente, dado seu impacto na saúde pública caso se confirmem.

Tratando-se de um algoritmo de *clustering*, é difícil avaliar a solução encontrada quanto à sua qualidade, sendo obviamente possível encontrar outras soluções. Todavia, o objetivo desta metodologia é a exploração dos dados, e de descoberta de padrões e informação, que possa ser relevante, mais profundamente analisada, e posteriormente confirmada. Assumindo a sua correção, estes resultados levantam várias questões, cuja resposta ultrapassa o âmbito deste trabalho, mas que interessam responder no futuro. A heterogeneidade encontrada é grande, e torna árdua a tarefa de

interpretação. Esta heterogeneidade pode explicar, no entanto, as dificuldades no sucesso terapêutico dos indivíduos com diabetes, sublinhando a importância e necessidade de estratégias "talhadas à medida" de cada um – medicina de precisão.

5. CONCLUSÕES

Neste trabalho aplicamos o algoritmo *superSOM* implementado em R, com o objetivo de explorar e estratificar uma população (Prevadiab2) que inclui quer indivíduos com normoglicemia, quer com hiperglicemia (hiperglicemia intermédia e diabetes), com o objetivo de encontrar diferentes fenótipos das alterações da glicemia. Assim, pela primeira vez, faz-se a avaliação de uma população, não baseada na classificação atual da diabetes tipo 2, e que explora a população no seu todo. Esta abordagem permitiu levantar novas questões bastante relevantes, que não teriam surgido de outro modo.

O número de ótimo de *clusters* foi definido por uma abordagem mista. Após a identificação do número de *clusters* pelo método do "cotovelo", calculado com o índice de Davies-Bouldin, utilizamos um algoritmo hierárquico para agregação das unidades do *SOM*. Progredindo na desagregação das unidades de acordo com a relevância da informação nelas contida, esta abordagem permitiu-nos uma avaliação gradual, antes da decisão final.

Não existem medidas absolutas de avaliação da qualidade dos *clusters.* No entanto, apesar do elevado número de variáveis utilizadas e do baixo número de dados, a abordagem utilizada, e em particular o algoritmo do *SOM* utilizado, permitiu identificar grupos de pessoas com diferentes fenótipos, no que respeita aos mecanismos metabólicos. Em 5 destes grupos estão presentes pessoas com diabetes, com hiperglicémia intermédia, e mais interessante, englobam também igual ou menor quantidade de pessoas com normoglicemia. Deste modo, foram identificadas pessoas com hiperglicemia intermedia e com diabetes com diferentes fenótipos.

Para além de terem sido encontrados diferentes fenótipos de hiperglicemia intermédia e de diabetes, estes resultados levantam uma série de questões relevantes e que urgem responder. Serão todos os indivíduos com normoglicémia realmente saudáveis no que respeita ao espectro da doença que culmina na diabetes tipo 2? Poderão estes fenótipos explicar as diferenças que se conhecem no aparecimento de complicações relacionadas com a diabetes tipo 2? Tem sido sugerido por alguns autores que os indivíduos com hiperglicemia intermédia que já apresentam complicações, deveriam ser classificados como já tendo a doença. Estarão estes indivíduos agrupados preferencialmente em alguns dos *clusters* com maior número de doentes? Os indivíduos com hiperglicemia intermédia que estão nos diferentes grupos, têm diferença de risco de progressão na doença? A resposta a qualquer uma destas perguntas é urgente pelo potencial impacto clínico marcante. No entanto, ultrapassa o âmbito deste trabalho, e será alvo de trabalhos futuros.

Os resultados obtidos permitem antever a necessidade de diferentes abordagens de terapêutica, mas também de prevenção, dirigindo-se para a medicina de precisão, apontando já alguns caminhos. Esta mudança de paradigma é a "luz ao fundo do túnel" para a reversão dos maus resultados da terapêutica instituída de modo transversal a todos, e que se traduz no enorme peso socioeconómico, por todo o mundo, que é conhecido.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

A análise de *clusters*, como a maior parte das técnicas utilizadas em data-mining necessita de grandes quantidades de dados. Neste trabalho utilizamos a base PREVADIAB2, e após a limpeza e transformação de dados, incluímos apenas 1010 indivíduos. Considerando o número de variáveis introduzidas, e a baixa proporção de algumas alterações da glicemia, este número pode limitar a qualidade dos resultados. Tendo em conta os resultados atingidos, e a sua potencial relevância, parece-nos importante que sejam replicados e avaliados em bases com maior número de vetores de entrada.

Ao optarmos por não incluir alguns fatores de risco, nomeadamente a tensão arterial, por esta ter resultado de apenas uma medição, e alguns indivíduos se encontrarem a tomar anti-hipertensores, podemos estar a escamotear algum padrão, que se possa relacionar com este parâmetro.

Apesar das limitações enunciadas, os resultados deste trabalho mostram padrões de perfis destes indivíduos, que constituem diferentes fenótipos. Mais, levanta várias questões e hipóteses relevantes, cuja resposta pode ter uma aplicação clínica fundamental, na abordagem desta doença, e que urgem confirmar em trabalhos futuros.

7. BIBLIOGRAFIA

Ahlqvist E., Storm P., Karajamaki A., et al. (2018) Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol 6:361–369.

Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing Maps as Substitutes for K-Means Clustering. *LNCS*, *3516*, 476–483.

Bansal, N. (2015). Prediabetes diagnosis and treatment: A review. *World Journal of Diabetes*, *6*(2), 296–303.

Belciug, S. (2009). Patients length of stay grouping using the hierarchical clustering algorithm. Retrieved from *Annals Math. Comp. Sci. Ser*, *36*(2), 79–84.

Cali', A. M. G., Bonadonna, R. C., Trombetta, M., Weiss, R., & Caprio, S. (2008). Metabolic abnormalities underlying the different prediabetic phenotypes in obese adolescents. *Journal of Clinical Endocrinology and Metabolism*, *93*(5), 1767–1773.

Cefalu, W., & et al. (2017). Standards of Medical Care in Diabetes — 2017. *Diabetes Care*, *40*(January), 142.

Chipman, H., & Tibshirani, R. (2006). Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, *7*(2), 286–301.

Collins, F. S., & Varmus, H. (2015). A New Initiative on Precision Medicine. *New England Journal of Medicine*, *372*(9), 793–795.

DeFronzo, R. A. (2009). From the triumvirate to the ominous octet: A new paradigm for the treatment of type 2 diabetes mellitus. *Diabetes*, *58*(4), 773–795.

Fort, J. C., Cottrell, M., & Letremy, P. (2001). Stochastic on-line algorithm versus batch algorithm for quantization and Self-Organizing maps. *Neural Networks for Signal Processing XI: Proceedings of the 2001 IEEE Signal Processing Society Workshop. IEEE, Piscataway, NJ, USA*, 43–52.

Fowler, M. J. (2011). Microvascular and Macrovascular Complications of Diabetes. *Clinical Diabetes*, *29*(3), 116–122.

Gardete-Correia, L., Boavida, J. M., Raposo, J. F., Mesquita, A. C., Fona, C., Carvalho, R., & Massano-Cardoso, S. (2010). Original Article: Epidemiology First diabetes prevalence study in Portugal: PREVADIAB study. *Diabet. Med*, *27*, 879–881.

Guariguata, L., Whiting, D. R., Hambleton, I., Beagley, J., Linnenkamp, U., & Shaw, J. E. (2014). Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Research and Clinical Practice*, *103*(2), 137–149.

Häring, H. U. (2016). Novel phenotypes of prediabetes? *Diabetologia*.

Hoff, M. H., Fish, U. S., & Service, W. (2001). Standard Operating Procedures. *Validation of Computerized Analytical and Networked Systems*, (2010), 181–200.

Househ, M., & Aldosari, B. (2017). The Hazards of Data Mining in Healthcare. *Studies in Health Technology and Informatics*, *238*, 80–83. Retrieved from

IDF. (2017). *IDF Diabetes Atlas, Eight edition*.

International Diabetes Federation. (2017). *IDF Clinical Practice Recommendations for managing Type 2 Diabetes in Primary Care*. Retrieved from www.idf.org/managing-type2-diabetes

International Expert Committee, T. I. E. (2009). International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes Care*, *32*(7), 1327–34.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323.

Jothi, N., Aini, N. ', Rashid, A., & Husain, W. (2015). Data Mining in Healthcare – A Review. *Procedia Computer Science*, *72*, 306–313.

Kesumawati, A., & Setianingsih, D. (2016). A Segmentation Group by Kohonen Self-Organizing Maps (SOM) and K -Means Algorithms (Case Study : Malnutrition Cases in Central Java of Indonesia). *Int. J. Advance Soft Compu. Appl*, *8*(3), 100–115.

Kim, E., Oh, W., Pieczkiewicz, D. S., Castro, M. R., Caraballo, P. J., & Simon, G. J. (2014). Divisive Hierarchical Clustering towards Identifying Clinically Significant Pre-Diabetes Subpopulations. *AMIA 2014 Annual Symposium*, 1815–1824.

Koh, H., & Tan, G. (2005). Data mining applications in healthcare. *J Healthc Inf Manag*, *19*(2), 64–72.

Kohonen, T. (1990). The Self-Organizing Map. *Proceeding of the IEEE*, *78*(9), 1464–1480.

Kohonen, T. (2001). *Self-Organizing Maps*.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65.

Kumar, R., Nandhini, L., Sadishkumar, K., Sahoo, J., & Vivekanadan, M. (2016). Evidence for current diagnostic criteria of diabetes mellitus. *World J Diabetes*, *7*(717), 396–405.

Marinov, M., Mosa, A. S. M., Yoo, I., & Boren, S. A. (2011). Data-mining technologies for diabetes: a systematic review. *Journal of Diabetes Science and Technology*, *5*(6), 1549–56.

National Diabetes Data Group. (1979). Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. National Diabetes Data Group. *Diabetes*, *28*(12), 1039–57.

Nithya, N. S., Duraiswamy, K., & Gomathy, P. (2013). A Survey on Clustering Techniques in Medical Diagnosis. *International Journal of Computer Science Trends and Technology*, *1*(2), 17–22.

Observatório Nacional da Diabetes. (2016). *Diabetes Factos e Números, o ano de 2015 - Relatório Anual do Observatório Nacional da Diabetes*.

Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., … Makaroff, L. E. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, *128*, 40–50.

R Core Team R Foundation for, & Computing. (2018). R: A language and environment for statistical computing. Vienna.

Schatz, M., Hsu, J.-W. Y., Zeiger, R. S., Chen, W., Dorenbaum, A., Chipps, B. E., & Haselkorn, T. (2014). Phenotypes determined by cluster analysis in severe or difficult-to-treat asthma. *Journal of Allergy and Clinical Immunology*, *133*(6), 1549–1556.

Tabák, A., & et al. (2012). Prediabetes: a high-risk state for diabetes development. *The Lancet*, *379*(9833), 2279–2290.

Tabák, A., & et al. (2017). Prediabetes : A high-risk state for developing diabetes Progression from prediabetes to diabetes Reversion to normoglycaemia Risk prediction. *Pmc*, *379*(9833), 1–14.

Țăranu, I. (2015). Data mining in healthcare: decision making and precision. *Database Systems Journal*, *VI*(4).

The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. (1997). Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*, *20*(7),

1183–97.

Tirunagari, S., Poh, N., Hu, G., & Windridge, D. (2015). Identifying Similar Patients Using Self-Organising Maps: A Case Study on Type-1 Diabetes Self-care Survey Responses. Retrieved from

Toppila, I. (2016). Identifying novel phenotype profiles of diabetic complications and their genetic components using machine learning approaches.

Ultsch, A. (2003). U*-Matrix: a Tool to visualize Clusters in high dimensional Data. *Computer*, *52*(36), 1–10. Retrieved from http://www.informatik.uni-marburg.de/~databionics/papers/ultsch03ustar.pdf

Ultsch, A., & Mörchen, F. (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. *Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany*, 1–7.

Uri, D., & Breard, G. (2017). *Evaluating Self-Organizing Map Quality Measures as Convergence Criteria*.

Veloso, R., Portela, F., Santos, M. F., Silva, Á., Rua, F., Abelha, A., & Machado, J. (2014). ScienceDirect A clustering approach for predicting readmissions in Intensive Medicine. *Procedia Technology*, *16*(16), 1307–1316.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600.

Vesanto, J., Alhoniemi, E., & Member, S. (2000). Clustering of the Self-Organizing Map, *11*(3), 586–600.

Wehrens, M. R., & Kruisselbrink, J. (2018). Package ‘ kohonen .’

Wehrens, R. (2007). Self- and Super-organizing Maps in R : The kohonen, *21*(5).

Wehrens, R. (2015). Package ‘ kohonen .’ *R Topics Documented*, 25.

WHO, & IDF. (2006). *Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia*.

Wickham, H. (2016). Elegant Graphics for Data Analysis. Springer-Verlag New York.

World Health Organization. (1965). Diabetes mellitus. Report of a WHO expert committee. *World Health Organization Technical Report Series*.

World Health Organization. (1980). *WHO Expert Committee on Diabetes Mellitus*.

World Health Organization. (1985). Diabetes mellitus. Report of a WHO Study Group. *World Health Organization Technical Report Series*, *727*, 1–113.

World Health Organization. (2016). Global Report on Diabetes. *Isbn*, *978*, 88.

Wu, Y., Ding, Y., Tanaka, Y., & Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *International Journal of Medical Sciences*, *11*(11), 1185–1200.

Yin, H. (2008). The self-organizing maps: Background, theories, extensions and applications. *Studies in Computational Intelligence*, *115*, 715–762. https://doi.org/10.1007/978-3-540-78293-3_17

Zhao, Q. (2012). *Cluster Validity in Clustering Methods*.

8. ANEXOS

Anexo A

Estatísticas das variáveis selecionadas do Prevadiab 2.

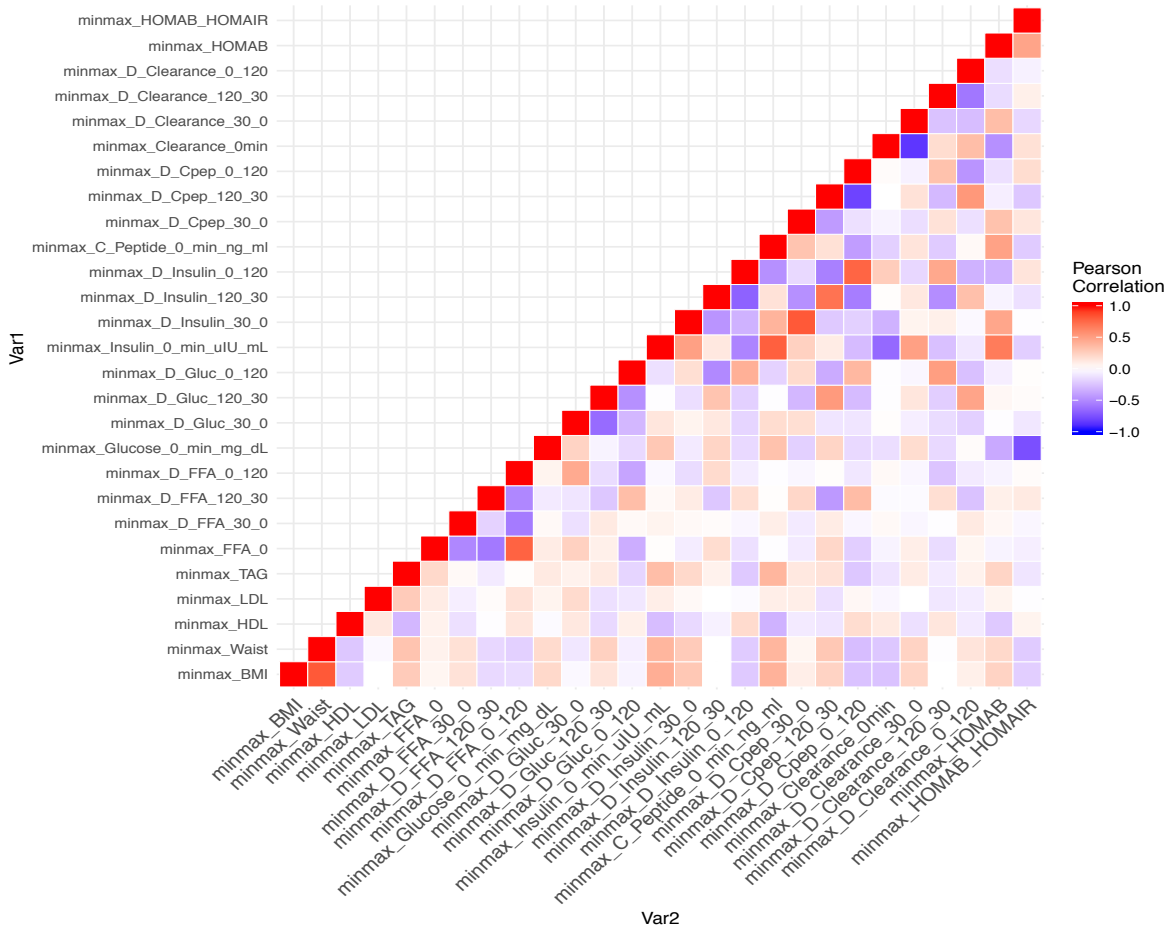
	Média	Desvio padrão	Valores omissos	Mínimo	Máximo	Mediana
IMC	27,4	4,3	10%	17	49	27
PA	72	12,6	10%	39	116	71
Idade	60	13	10%	22	86	62
Bioimpedância	33	7,6	23%	6,7	50	33,5
Plaquetas	231	66,9	98%	80	450	219
Albumina	4,0	0,25	98%	3,29	4,43	3,98
Creatinina						
AST	25,2	10,2	7%	9	150	23
ALT	23,85	12,9	7%	4,1	127	21
GGT	31,5	30,2	7%	7	323	22
Colesterol total	201	37,2	7%	65	198	81,4
Colesterol HDL	53,2	12,4	7%	25,2	106	52,4
Colesterol LDL	138	30,7	7%	49	268	135
Triglicéridos	118	60	7%	33	467	104
Ácidos gordos livres 0’	0,52	0,22	7%	0,08	1,49	0,49
Ácidos gordos livres 30’	0,37	0,18	7%	0,02	1,61	0,34
Ácidos gordos 120’	0,14	0,1	7%	0,01	0,97	0,12
Glicose 0’	93	13,5	7%	63	215	91
Glicose 30’	156	34	7%	56	310	153
Glicose 120’	118	43	7%	42	444	111
Insulina 0’	7,8	5,1	7%	0,64	46,9	6,2
Insulina 30’	54,7	36,4	7%	2,8	279	46
Insulina 120’	46,7	45,9	7%	0,4	325,5	34,3
Peptídeo C 0’	2,23	0,96	7%	0,5	7,78	2,06
Peptídeo C 30’	6,7	2,7	7%	1,21	24	6,3
Peptídeo C 120’	8,6	4	7%	1,14	30	7,88

	Moda	Valores omissos
Género	Feminino (60,3%)	-

Anexo B

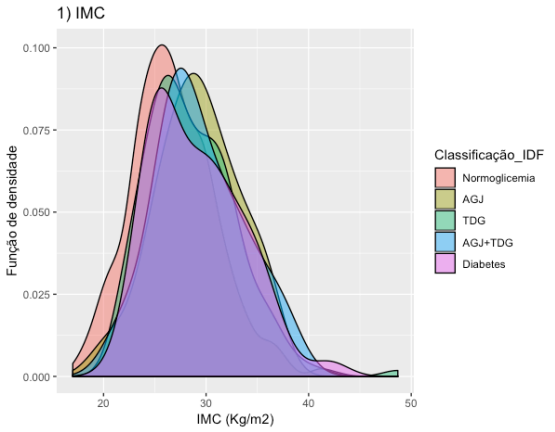
Correlação entre as variáveis selecionadas para as grelhas do *superSOM*

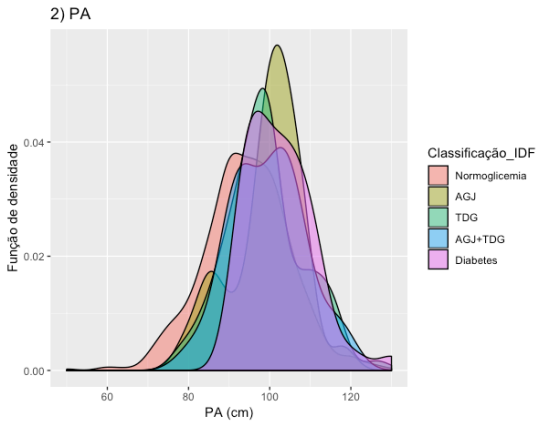
A corelação mais elevada é de 0,81 entre o PA e o IMC.

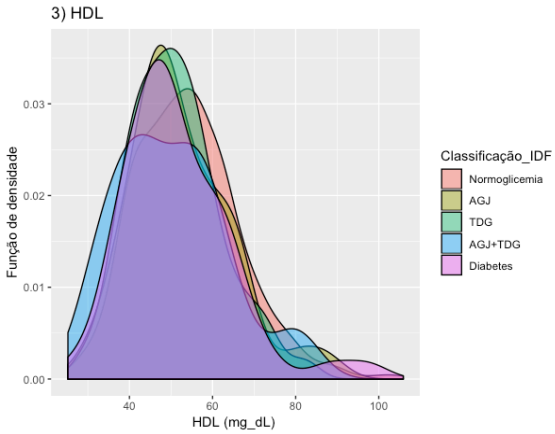


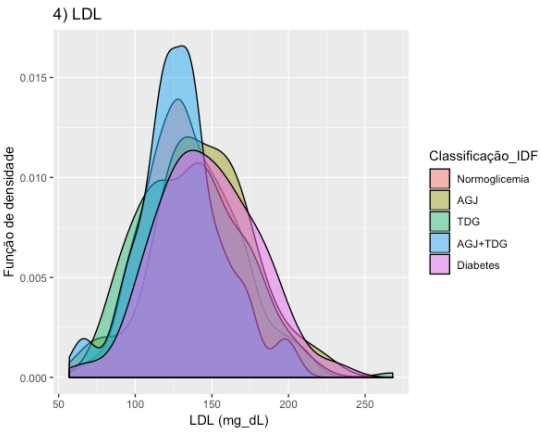
ANEXO C

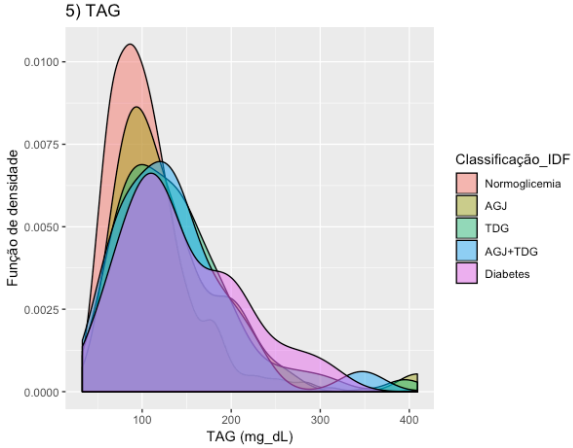
Distribuição das variáveis selecionadas para a modelação do SOM, pelas classes glicemia da IDF. Em algumas variáveis estas curvas estão sobrepostas, como é o caso do BMI, sugerindo que estes parâmetros não distinguem as classes de hiperglicemia definidas pela IDF. Noutras, como é o caso do HOMA B, as curvas são multimodais, sugerindo que podem existir diferentes grupos dentro das mesmas classes.

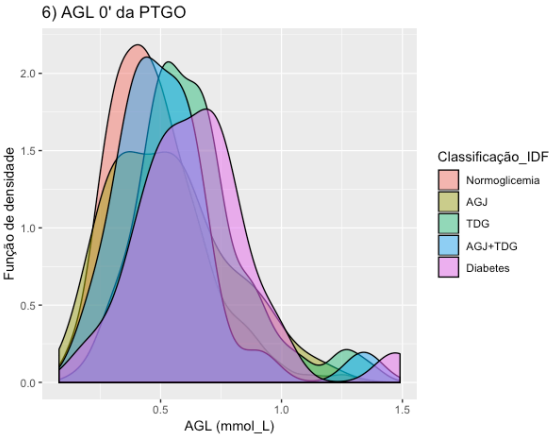




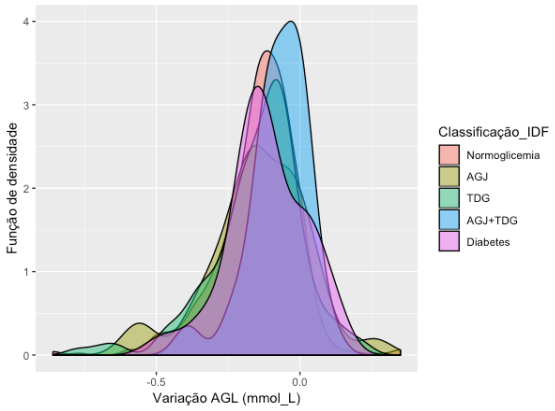




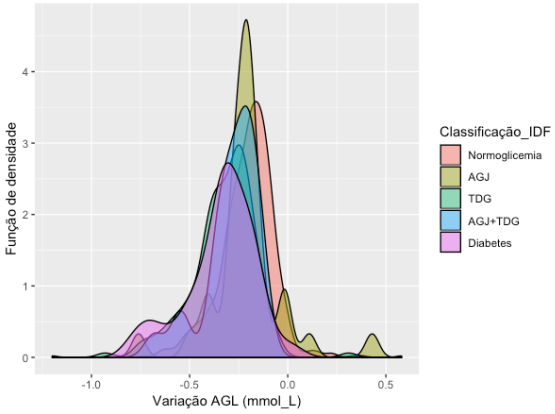




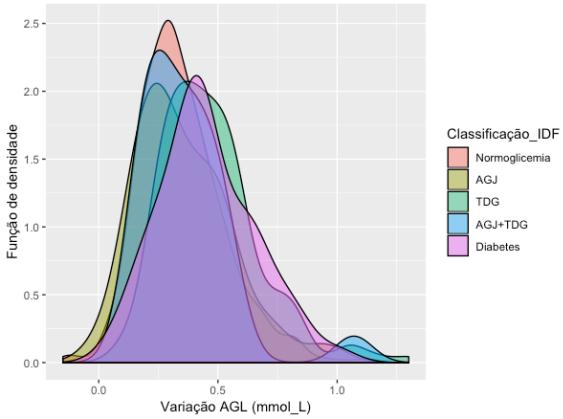
7) Variação de ácidos gordos livres dos 0' aos 30´ da PTGO



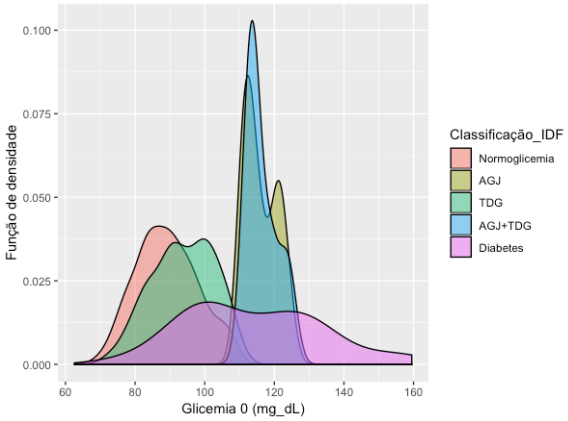
8) Variação de ácidos gordos livres dos 30' aos 120´ da PTGO



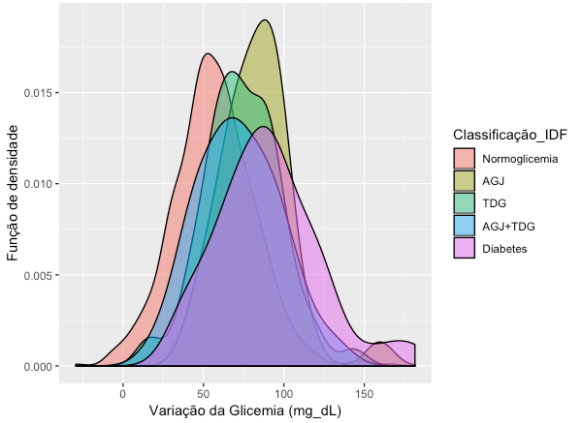
9) Variação de ácidos gordos livres dos 120' aos 0´ da PTGO



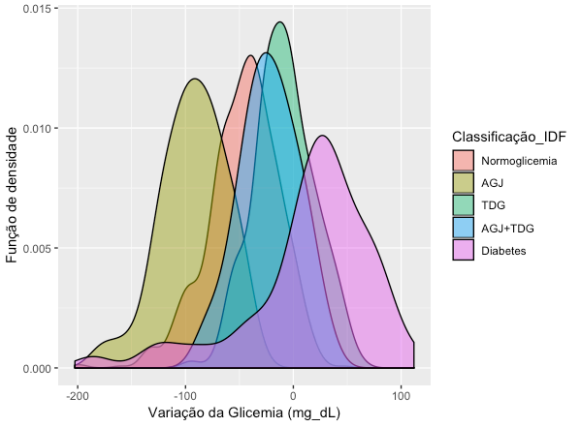
10) Glicemia aos 0´ da PTGO



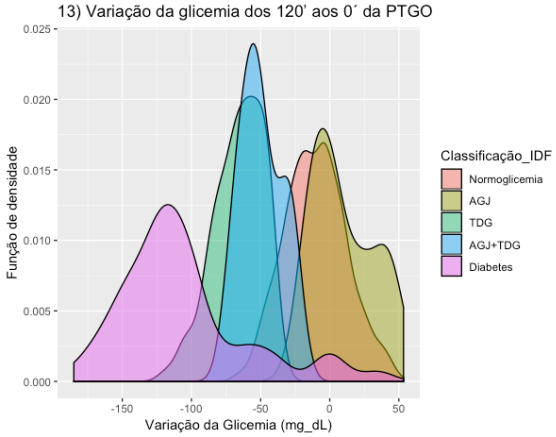
11) Variação da glicemia dos 0' aos 30´ da PTGO



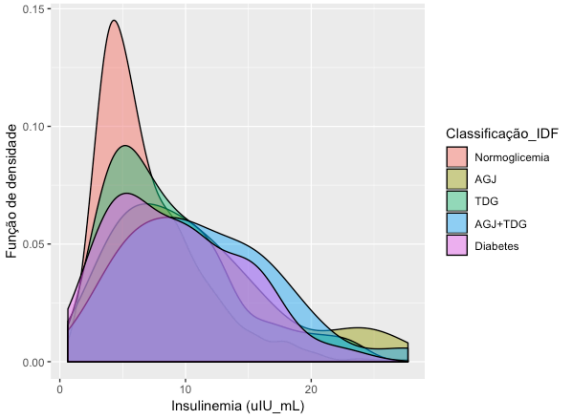
12) Variação da glicemia dos 30' aos 120´ da PTGO



13) Variação da glicemia dos 120' aos 0´ da PTGO

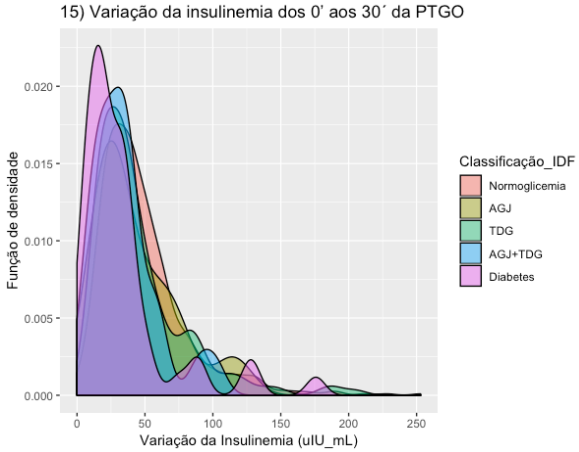


14) Insulinemia aos 0' da PTGO

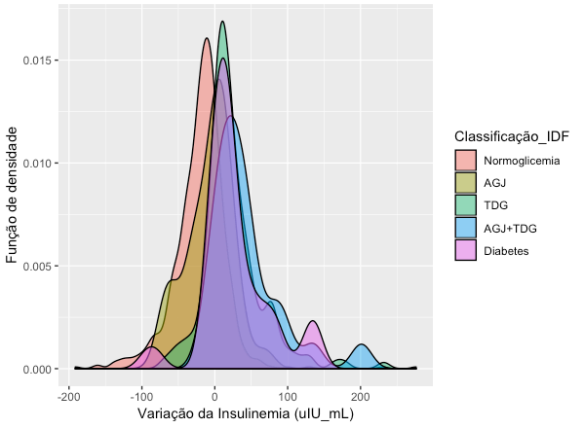


14) Insulinemia aos 0' da PTGO

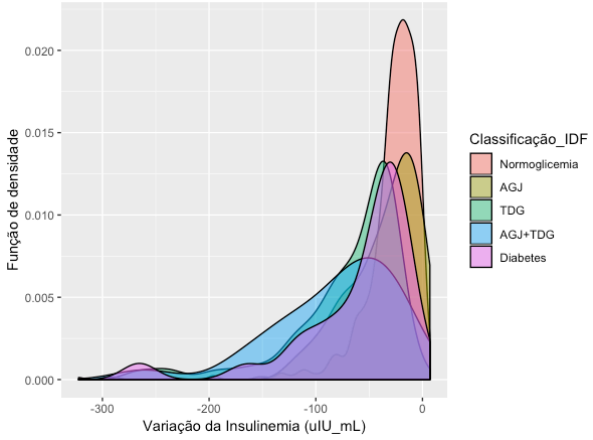
15) Variação da insulinemia dos 0' aos 30´ da PTGO



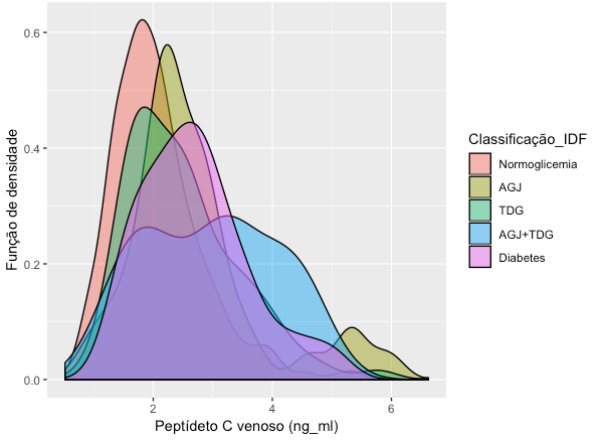
16) Variação da insulinemia dos 30' aos 120´ da PTGO



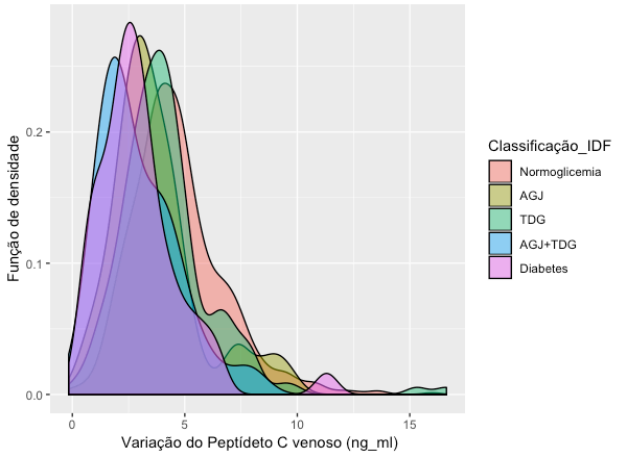
17) Variação da insulinemia dos 120' aos 0´ da PTGO



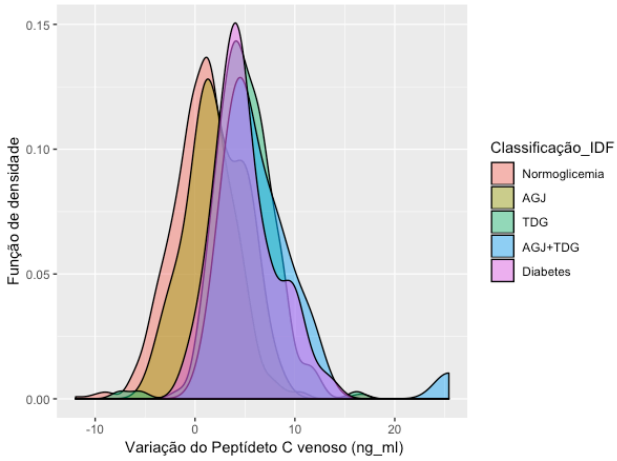
18) C-peptídeo venoso aos 0' da PTGO



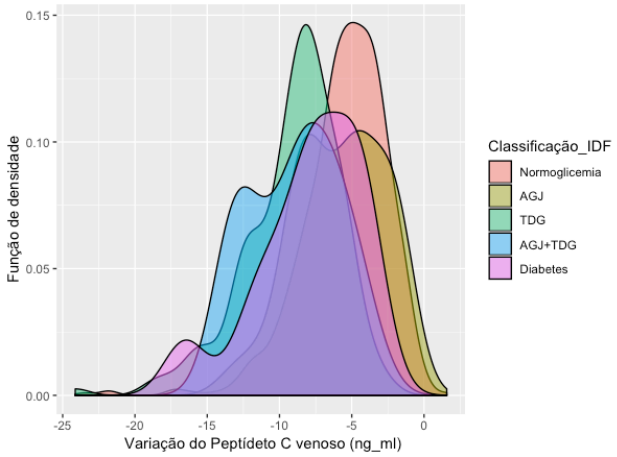
19) Variação da C-peptídeo venoso dos 0' aos 30´ da PTGO



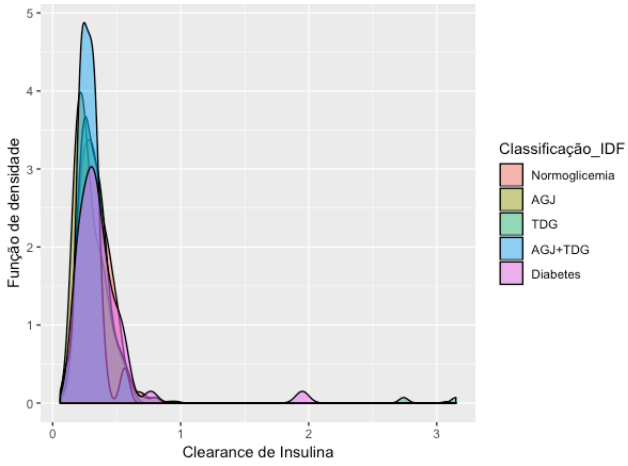
20) Variação da C-peptídeo venoso dos 30' aos 120´ da PTGO



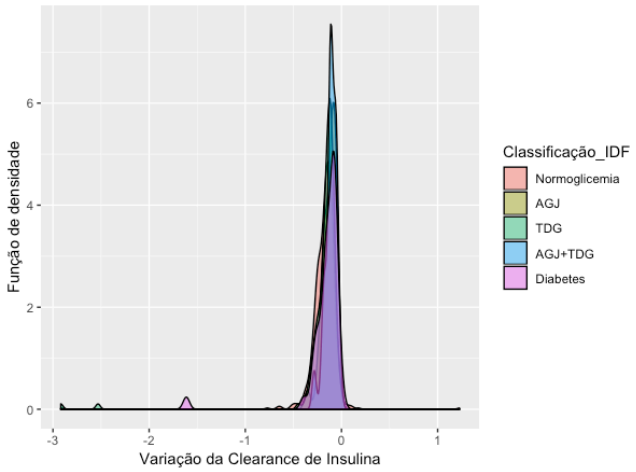
21) Variação da C-peptídeo venoso dos 12 0' aos 0´ da PTGO



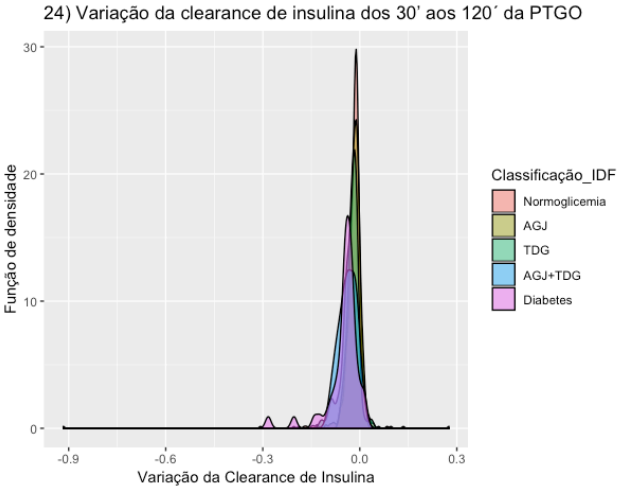
22) Clearance de insulina aos 0' da PTGO



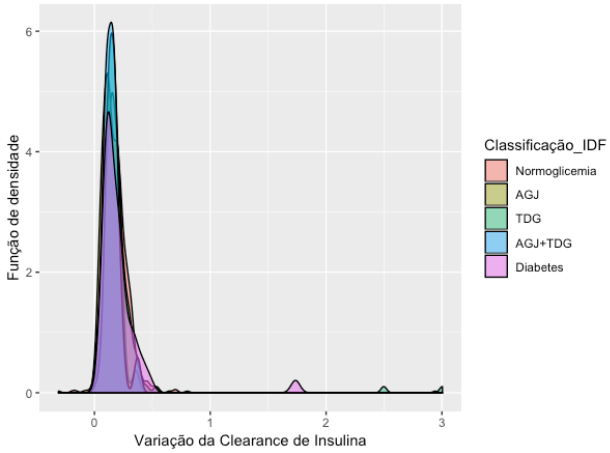
23) Variação da clearance de insulina dos 0' aos 30´ da PTGO



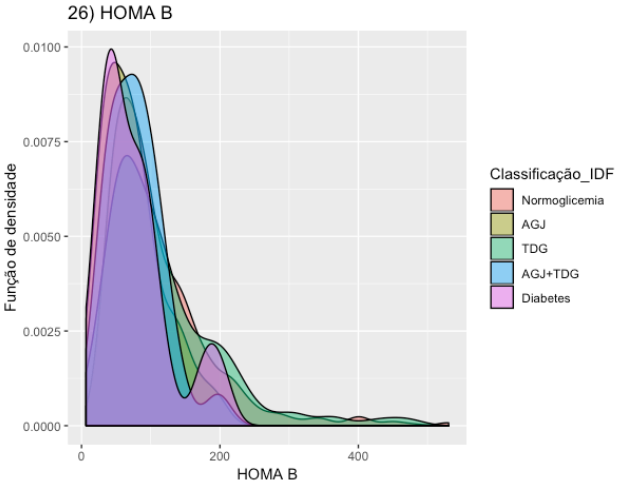
24) Variação da clearance de insulina dos 30' aos 120´ da PTGO



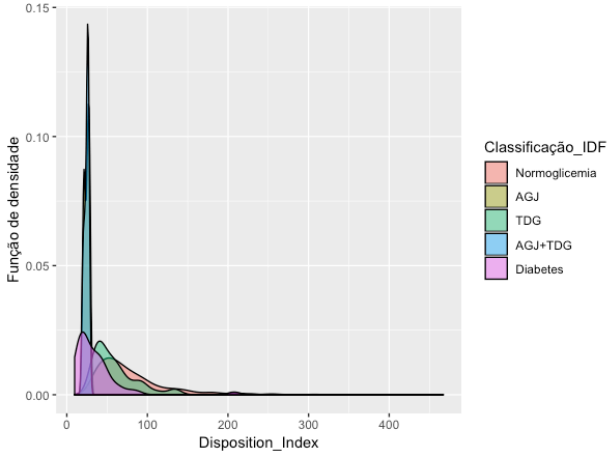
25) Variação da clearance de insulina dos 120' aos 0´ da PTGO



26) HOMA B



27) Disposition Index



ANEXO D

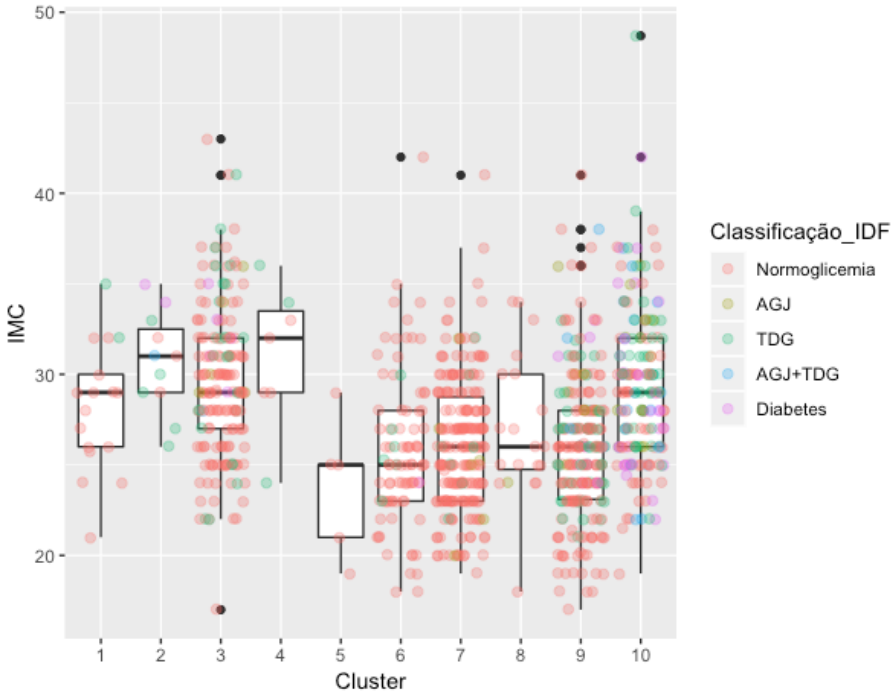
Distribuição da população pelas 27 unidades do *superSOM* selecionado

Unidade	Número de indivíduos	Unidade	Número de indivíduos
1	17	**15**	60
2	22	**16**	24
3	7	**17**	31
4	5	**18**	54
5	42	**19**	26
6	23	**20**	63
7	33	**21**	44
8	59	**22**	40
9	24	**23**	73
10	42	**24**	51
11	30	**25**	11
12	20	**26**	33
13	30	**27**	26
14	73	**NC**	47

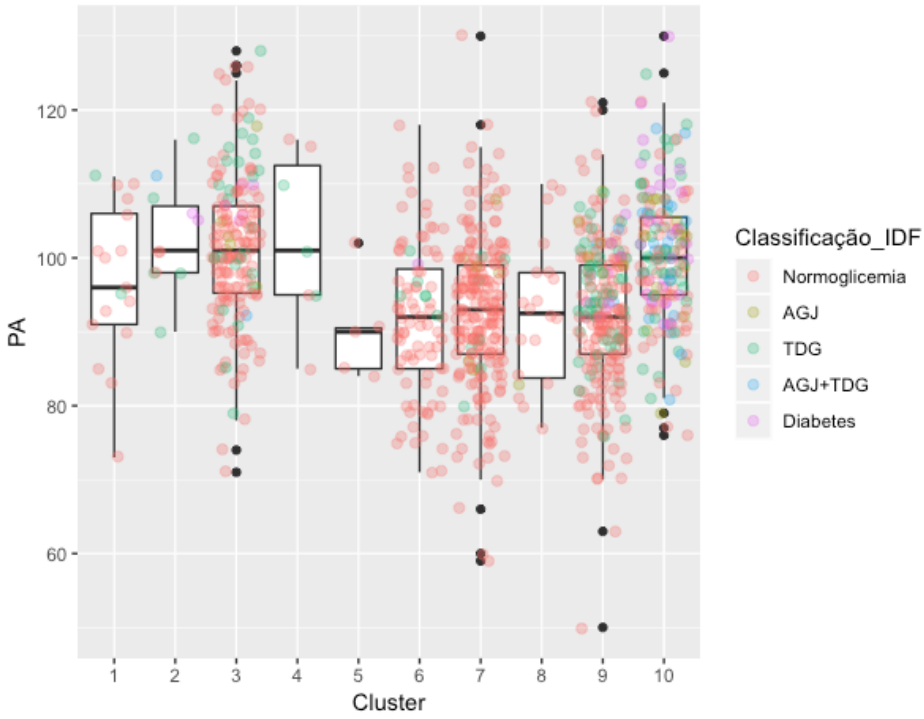
ANEXO E

Distribuição do IMC(a), PA(b), HOMA_IR(c), HOMA_B(d), TAG(e), HDL(f), LDL(g), Colesterol total(h), Idade(i) nos 10 *clusters* definidos pelo algoritmo *superSOM*.

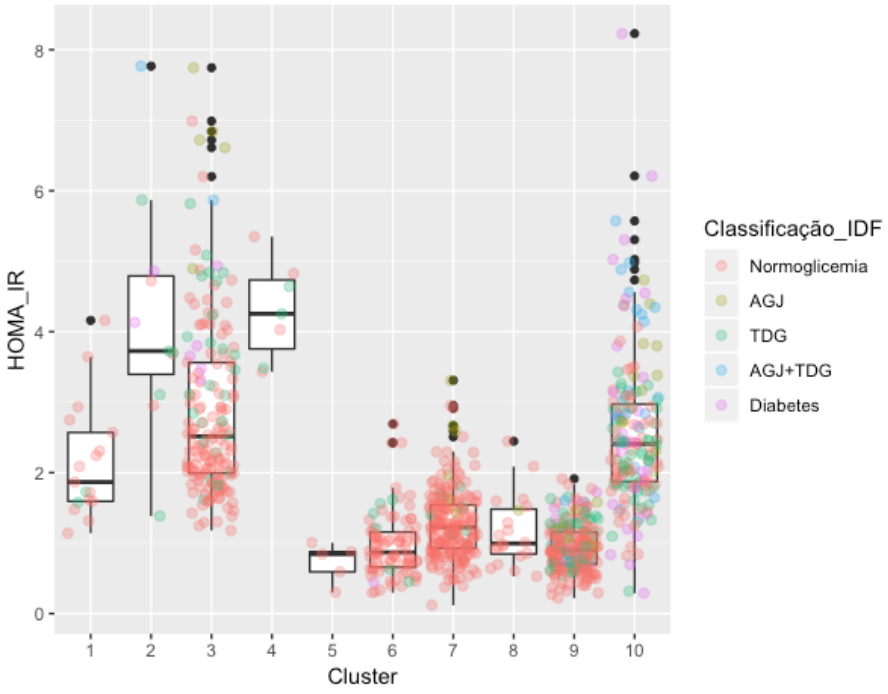
a) IMC



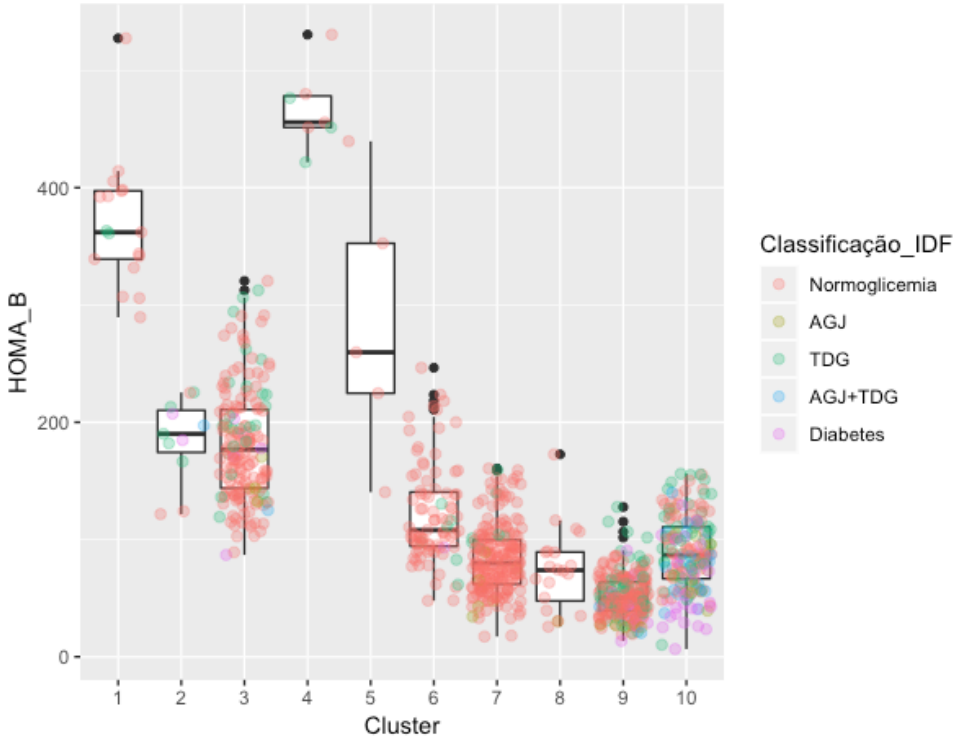
b) PA



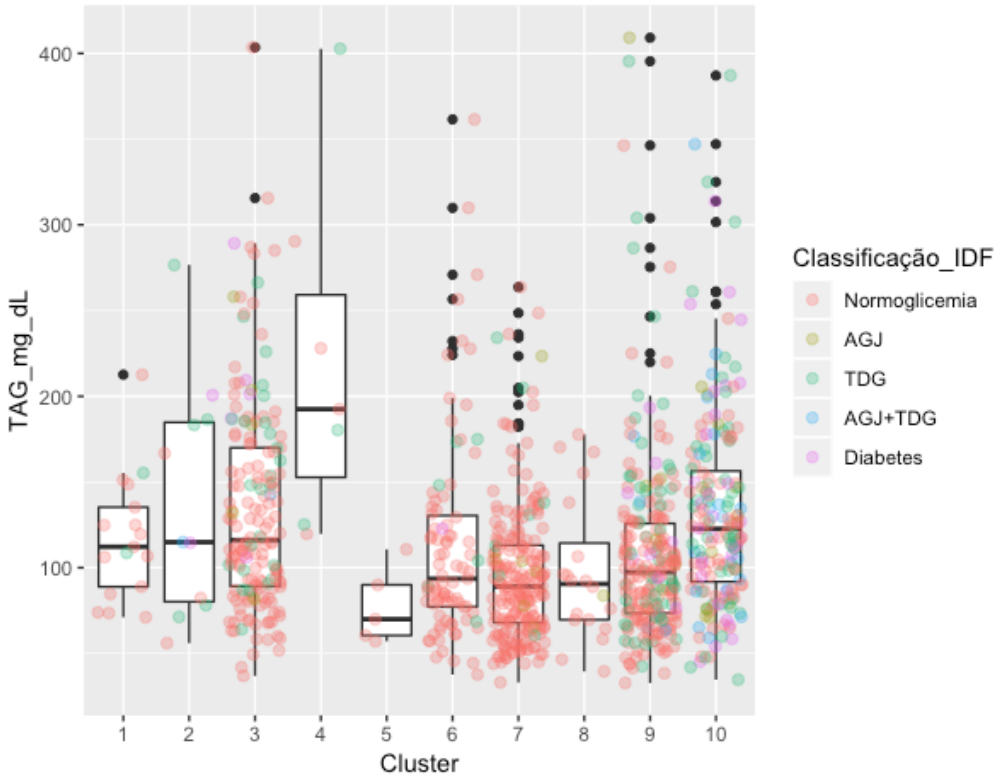
c) HOMA_IR



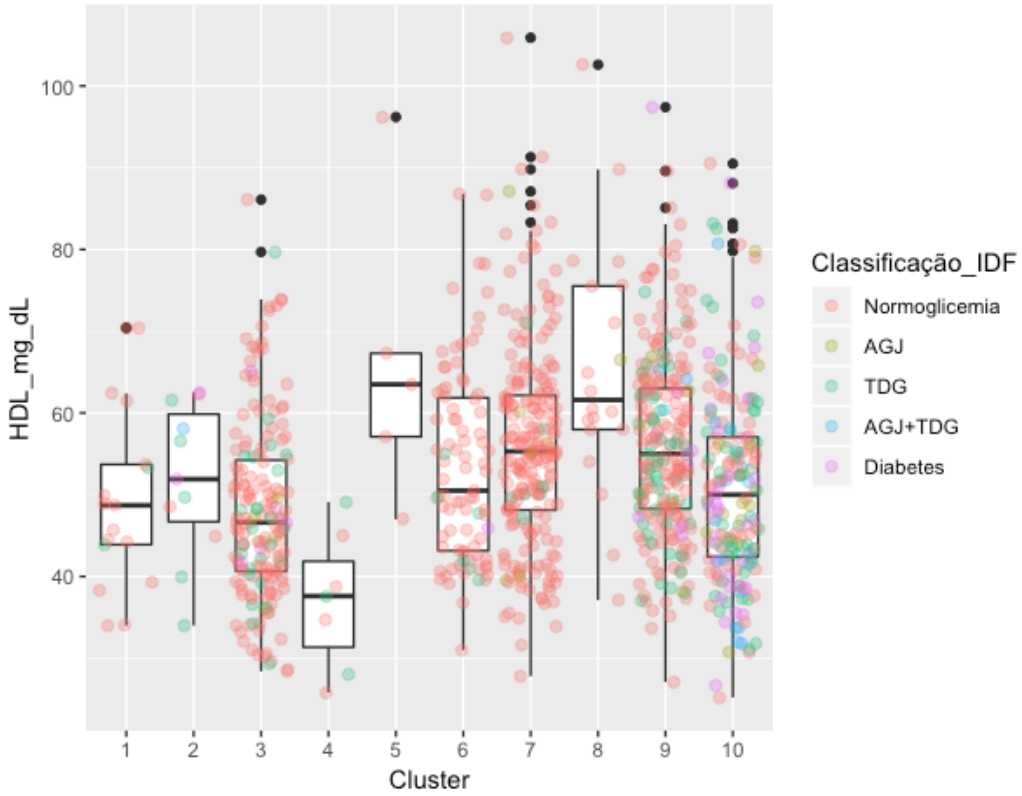
d) HOMA_B



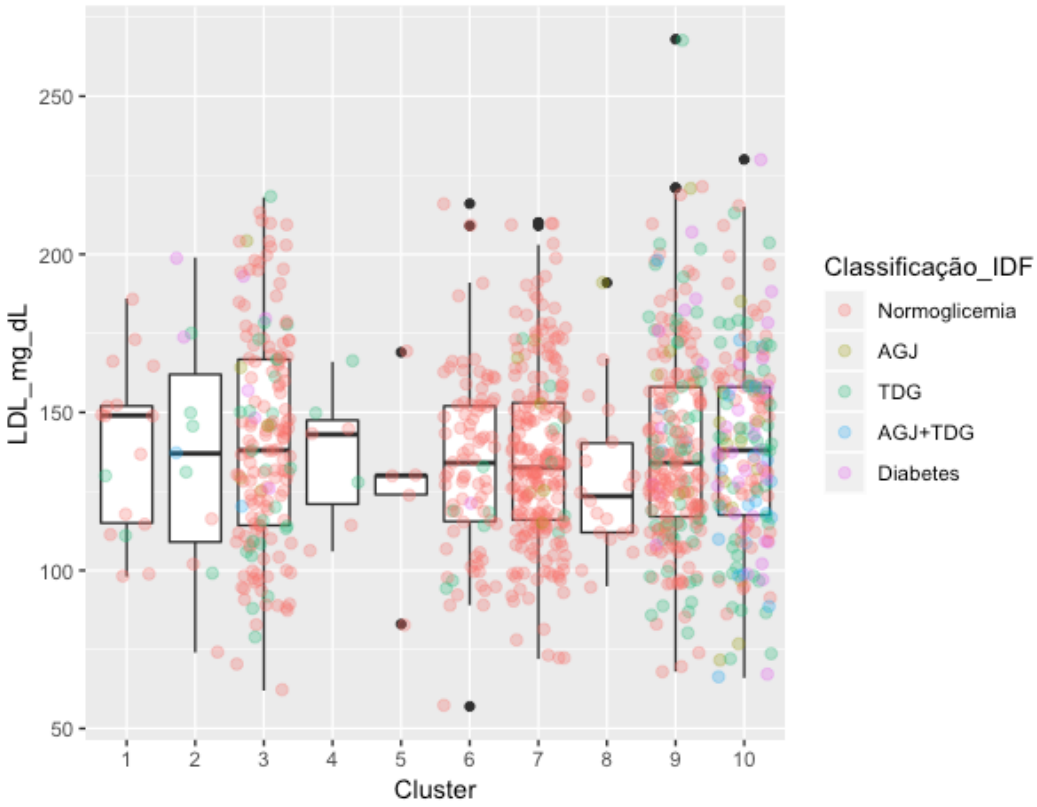
e) TAG


TAG_mg_dL
400
300
200
100
1
2
3
4
5
6
7
8
9
10
Cluster
Classificação_IDF
Normoglicemia
AGJ
TDG
AGJ+TDG
Diabetes

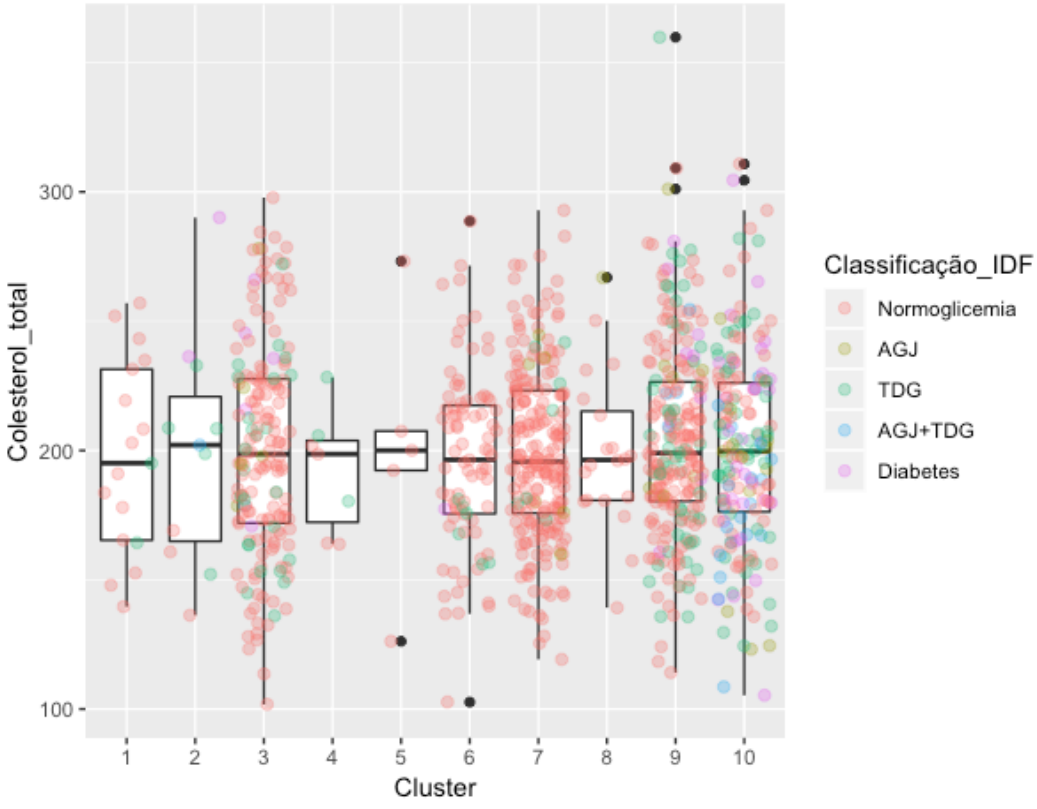
f) HDL


HDL_mg_dL
100
80
60
40
1
2
3
4
5
6
7
8
9
10
Cluster
Classificação_IDF
Normoglicemia
AGJ
TDG
AGJ+TDG
Diabetes

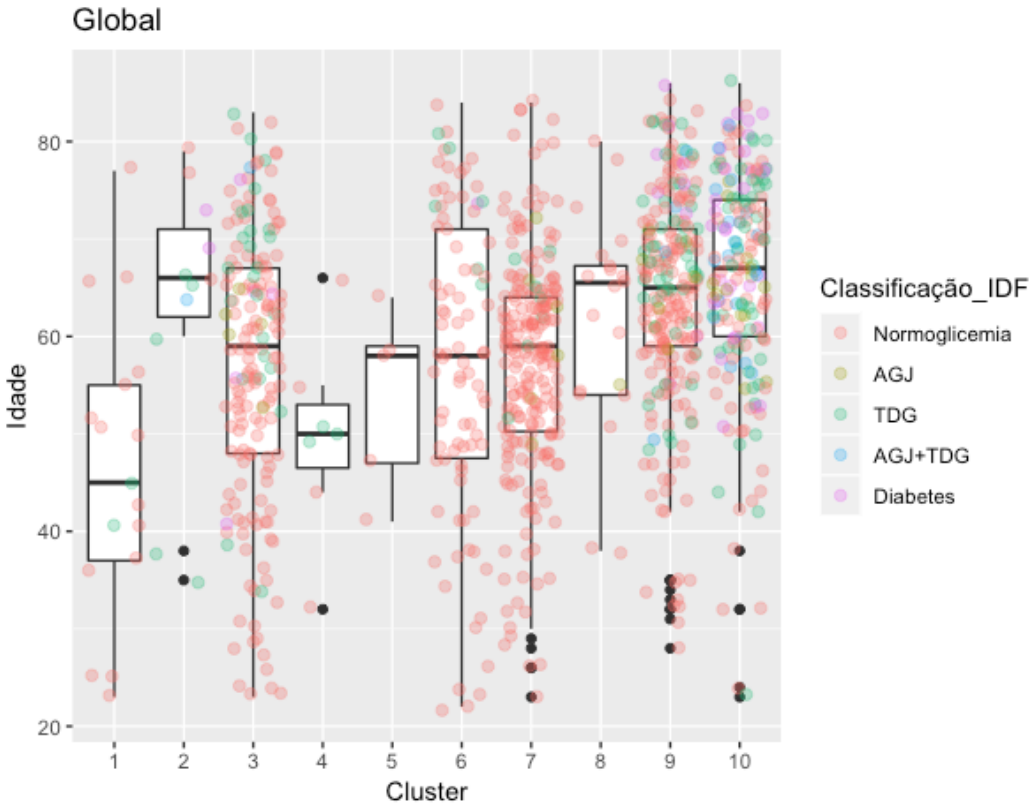
g) LDL



h) Colesterol total



i) Idade


Global
Idade
80
60
40
20
1
2
3
4
5
6
7
8
9
10
Cluster
Classificação_IDF
Normoglicemia
AGJ
TDG
AGJ+TDG
Diabetes