

UNIVERSIDADE FEDERAL DO PARANÁ

ANA PAULA FILUS BANDEIRA

**APLICAÇÃO DE REDE NEURAL ARTIFICIAL PARA O
RECONHECIMENTO DO DIABETES MELLITUS GESTACIONAL
COM MARCADORES NÃO-GLICÊMICOS**

CURITIBA

2015

ANA PAULA FILUS BANDEIRA

**APLICAÇÃO DE REDE NEURAL ARTIFICIAL PARA O
RECONHECIMENTO DO *DIABETES MELLITUS* GESTACIONAL
COM MARCADORES NÃO-GLICÊMICOS**

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração em Bioinformática.

Orientador: Professor Dr. Geraldo Picheth
Co-orientadora: Professora Dra. Jeroniza Nunes Marchaukoski

CURITIBA

2015

UNIVERSIDADE FEDERAL DO PARANA SISTEMA DE BIBLIOTECAS
CATALOGAÇÃO NA FONTE

B214 Bandeira, Ana Paula Filus
 Aplicação de rede neural artificial para o reconhecimento do diabetes
 mellitus gestacional com marcadores não-glicêmicos / Ana Paula Filus
Bandeira. - Curitiba, 2015.
 75 f.: il., tabs, grafs.

 Orientador: Prof^o. Dr. Geraldo Picheth
 Co-orientadora: Profa. Dra. Jeroniza Nunes Marchaukoski
 Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de
Educação Profissional e Tecnológica, Curso de Pós-Graduação em
Bioinformática.

 1. Diabetes mellitus - Gravidez. 2. Redes neurais (Computação).
 3. Sistemas de suporte de decisão. 4. Software - Desenvolvimento.
 5. Bioinformática. I. Picheth, Geraldo. II. Marchaukoski, Jeroniza Nunes.
 III. Título. IV. Universidade Federal do Paraná.

CDD 006.3

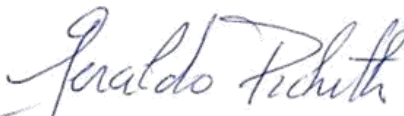
TERMO DE APROVAÇÃO

ANA PAULA FILUS BANDEIRA

“Aplicação de rede neural artificial para o reconhecimento do diabetes mellitus gestacional com marcadores não-glicêmicos”

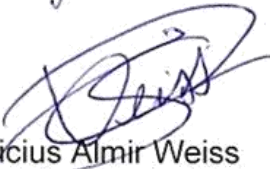
Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:


Orientador:


Prof. Dr. Geraldo Picheth

Coorientadora:


Profª Drª Jeroniza Nunes Marchaukoski


Dr. Vinicius Almir Weiss
Pós-Doc PNP/DCAPES/Projeto Biologia Computacional
Universidade Federal do Paraná - UFPR


Prof. Dr. Mauro Antonio Alves Castro
Universidade Federal do Paraná


Profª Drª Fabiane Gomes de Moraes Rego
Universidade Federal do Paraná

Curitiba, 13 de abril de 2015

DEDICATÓRIA

Aos meus pais e irmãos

Ao meu noivo

E principalmente à Deus

AGRADECIMENTOS

Às vezes a gente imagina que passar por todo esse processo pode ser algo inatingível, devido aos problemas que encontramos durante o caminho e até mesmo em nossas vidas. Mas como num bom livro, sempre há um final. Sempre aprendemos muitas coisas com nossos professores, colegas, amigos, pessoas com que convivemos dia a dia. E é por esse motivo que aqui escrevo à todas essas importantes pessoas que me ajudaram a concluir mais uma etapa em minha vida.

Agradeço aos meus queridos professores e orientadores, professor Geraldo Picheth, por primeiramente ter permitido em fazer parte desse projeto. Agradeço pela sua paciência e seus ensinamentos. O professor fora ótimo como pessoa e orientador. Obrigada professora Jeroniza Nunes Marchaukoski por toda a sua orientação e dedicação. Agradeço pela sua gentileza e por estar sempre presente em todos os momentos.

Obrigada à doutora Izabella Castilhos Ribeiro dos Santos-Weiss, por todo o seu conhecimento, amizade, contribuição e ensinamentos. Pelo seu apoio e presença em todo o projeto.

Agradeço ao mestre Waldemar Volanski, por toda a ajuda necessária. Pelo seu direcionamento e sua visão em meu trabalho.

Ao professor Roberto Tadeu Raitzz e professor Paulo Bracarense por suas instruções declaradas e concisas.

À secretaria geral do Programa de Pós-graduação em Bioinformática, e à CAPES. A cada um dos meus colegas mestrandos, Flávia, Eduardo, Venício, Caleb, Nilson, Elisa e Rodrigo, pela amizade e pela ajuda.

Aos meus familiares e amigos, e principalmente à Deus.

Em nenhum momento este processo foi um esforço solitário, muitas pessoas contribuíram para a realização e conclusão desse projeto. Então somente posso dizer à todos, com muito respeito, um grande muito obrigada.

“Owning our story can be hard but not nearly as difficult as spending our lives running from it. Embracing our vulnerabilities is risky but not nearly as dangerous as giving up on love and belonging and joy—the experiences that make us the most vulnerable. Only when we are brave enough to explore the darkness will we discover the infinite power of our light.”

Brené Brown

RESUMO

O *Diabetes mellitus* gestacional (DMG) afeta cerca de 7% das gestações e tem impacto importante para a gestante, o feto e o neonato. O diagnóstico do DMG é realizado com base na determinação da glicemia em jejum e após sobrecarga. A identificação de padrões com base em características antropométricas e laboratoriais pode ser de interesse no processo de triagem do DMG. Nesse trabalho, o toolbox, *open source*, “WEKA” foi empregado, utilizando os algoritmos *simple k-means* e *simple logistic*, para a classificação de variáveis não-glicêmicas. A rede neural artificial, multilayer perceptron (MLP), foi aplicada na busca da identificação automática do DMG. A base do estudo foi uma amostra de 997 gestantes (699 = gestantes saudáveis (controle); 298 = gestantes com diabetes gestacional (DMG)) utilizando critérios glicêmicos estabelecidos pela *American Diabetes Association*. Os marcadores discriminantes selecionados com auxílio dos algoritmos *k-means* e *simple logistic* foram idade, peso, índice de massa corporal, pressão arterial sistólica, pressão arterial diastólica, ácido úrico, triglicérides, colesterol não HDL e Log(TG/HDL-C). Estas variáveis foram aplicadas na rede neural MLP para a classificar gestantes com DMG. A MLP devidamente treinada e testada permitiu 88% de predições corretas, estabelecendo uma sensibilidade de 92,1%, especificidade de 83,8% e acurácia de 87,5%. Em síntese, a metodologia desenvolvida neste trabalho é de baixo custo e uma alternativa de “segunda opinião” para a triagem do DMG.

Palavras-chave: *Diabetes mellitus* gestacional; sistemas de apoio à decisão clínica; WEKA, *k-means*, *simple-logistic*, redes neurais artificiais.

ABSTRACT

The gestational diabetes mellitus (GDM) affects about 7% of pregnancies and has an important impact on the mother, the fetus and the newborn. The diagnosis of GDM is based on the determination of fasting and post-load glycemia. The identification of patterns based on anthropometric and laboratory parameters can be of interest in the screening process of DMG. In this work, the toolbox, open source, "WEKA" was employed, using the simple k-means and simple logistic algorithms, for non-glycemic variables rating. The artificial neural network, multilayer perceptron (MLP), was applied in the search for automatic identification of the DMG. The basis of the study was a sample of 997 patients (699 = healthy pregnant women (control); 298 = pregnant women with gestational diabetes (GDM)) using glycemic criteria established by the American Diabetes Association. The discriminant markers selected with the help of k-means and simple logistic algorithms were age, weight, body mass index, systolic blood pressure, diastolic blood pressure, uric acid, triglycerides, non-HDL cholesterol and Log(TG/HDL-C). These variables were applied in MLP to classify pregnant women with GDM. The MLP network properly trained and tested allowed 88% of correct predictions, establishing a sensitivity of 92.1%, specificity of 83.8% and accuracy of 87.5%. In summary, the methodology developed in this work is low cost and an appropriate alternative to "second opinion", allowing GDM screening.

Keywords: gestational diabetes; decision support systems; WEKA; *k-means*; *simple logistic*; artificial neural networks.

LISTA DE FIGURAS

FIGURA 1. MECANISMO DE AÇÃO DA INSULINA.....	15
FIGURA 2. ELEMENTOS ASSOCIADOS À PATOFISIOLOGIA DO DIABETES GESTACIONAL.	17
FIGURA 3. DIAGNÓSTICO DO DMG SEGUNDO OS CRITÉRIOS DO MINISTÉRIO DA SAÚDE DO BRASIL (2001).	19
FIGURA 4. PÁGINA INICIAL DO WEKA E A INTERFACE DE TRABALHO.	23
FIGURA 5. FORMATO ARFF PARA A INSERÇÃO DOS DADOS AO WEKA.....	24
FIGURA 6. ABA CLASSIFY E CLUSTER NA INTERFACE EXPLORER DO PROGRAMA WEKA.....	25
FIGURA 7. PORCENTAGEM DA CLASSIFICAÇÃO E MATRIZ DE CONFUSÃO.	56
FIGURA 8. ÉPOCAS VERSO ACERTOS COM A REDE MLP.....	59
FIGURA 9. ESTRUTURA E VARIÁVEIS DA REDE NEURAL MLP EM ESTUDO.	60
FIGURA 10. CLASSIFICAÇÃO DOS GRUPOS COM O USO DA REDE NEURAL MLP.	60

LISTA DE TABELAS

TABELA 1. PRINCIPAIS INFORMAÇÕES E ENSAIOS LABORATORIAIS OBTIDOS DAS GESTANTES EM ESTUDO (VARIÁVEIS NÃO GLICÊMICAS)...	21
TABELA 2. PRINCIPAIS INFORMAÇÕES E ENSAIOS LABORATORIAIS OBTIDOS DAS GESTANTES EM ESTUDO (VARIÁVEIS GLICÊMICAS)	22
TABELA 3. CLUSTERIZAÇÃO PELO ALGORITMO K-MEANS COM AS VARIÁVEIS SELECIONADAS, ENFATIZANDO A MÉDIA.	51
TABELA 4. COMPARAÇÃO E ANÁLISE COM O K-MEANS	52
TABELA 5. GRUPOS ESTABELECIDOS POR SUAS VARIÁVEIS.....	52
TABELA 6. VARIÁVEIS SIGNIFICATIVAS E EXCLUÍDAS APÓS O PROCESSO DE CLUSTERIZAÇÃO.	53
TABELA 7. RESULTANTE DA ANÁLISE DAS VARIÁVEIS SELECIONADAS COMBINADAS DA CLUSTERIZAÇÃO COM O ALGORITMO K-MEANS.	54
TABELA 8. COMPARAÇÃO E ANÁLISE COM AS VARIÁVEIS SELECIONADAS COM O K-MEANS.....	54
TABELA 9. VARIÁVEIS SELECIONADAS COM O ALGORITMO SIMPLE LOGISTIC.	55
TABELA 10. COMPARAÇÃO E ANÁLISE COM AS VARIÁVEIS SELECIONADAS COM ALGORITMO SIMPLE LOGISTIC.....	56
TABELA 11. INTERSECÇÃO DAS VARIÁVEIS SELECIONADAS COM OS ALGORITMOS K-MEANS E SIMPLE LOGISTIC E FORMAÇÃO DE MARCADORES SELECIONADOS COMBINADOS.	57
TABELA 12. COMPARAÇÃO DOS MARCADORES NÃO-GLICÊMICOS SELECIONADOS ENTRE OS GRUPOS CONTROLE E DMG.	58
TABELA 13. RESULTADOS PARA O GRUPO DE MARCADORES SELECIONADOS COM A REDE NEURAL MLP.....	61
TABELA 14. ANÁLISE DOS PARÂMETROS DE QUALIDADE DO TESTE E O INTERVALO DE CONFIANÇA.	62
TABELA 15. TOTAL DE VARIÁVEIS OBTIDAS INICIALMENTE.	71
TABELA 16. VARIÁVEIS ANTROPOMÉTRICAS.....	71
TABELA 17. VARIÁVEIS LABORATORIAIS	72

LISTA DE ABREVIATURAS

1.5AG	1.5 Anidroglucitol
A	Altura
ADA	Associação Americana de Diabetes
ALB	Albumina
ARFF	<i>Atributte-Relation File Format</i>
CDSS	<i>Clinical decision support systems</i>
COL	Colesterol
CTRL	Gestantes Controle
DM2	<i>Diabetes mellitus</i> tipo 2
DMG	<i>Diabetes mellitus</i> Gestacional
FRUTO	Frutosamina
GGT	Gama Glutamil Transferase
GLP	<i>General Public License</i>
GLU	Glucose
GRUPO	Grupamento de Semanas de gestação
HbA1C	Hemoglobina Glicada Fração A1c
HDL-C	<i>High Density Lipoprotein Cholesterol</i>
ID	Idade
IMC	Índice de Massa Corporal
LDL-C	<i>Low Density Lipoprotein Cholesterol</i>
LOG(TG/HDL-C)	Índice Aterogênico do Plasma
MLP	<i>Multilayer Perceptron</i>
N_HDL	Colesterol Não HDL
P	Peso
PAD	Pressão Arterial Diastólica
PAS	Pressão Arterial Sistólica
PT	Proteínas Totais
RNA	Rede Neural Artificial
SEM	Semanas de Gestação
SOP	Síndrome do Ovário Policístico
TG	Triglicerídeos
TOTG	Teste Oral de Tolerância à Glicose
URIC	Ácido Úrico

Sumário

1. INTRODUÇÃO	11
2. OBJETIVOS	13
2.1 OBJETIVO GERAL.....	13
2.2 OBJETIVOS ESPECÍFICOS.....	13
3. FUNDAMENTAÇÃO TEÓRICA	14
3.1 <i>DIABETES MELLITUS</i>	14
3.2 <i>DIABETES MELLITUS</i> GESTACIONAL.....	16
3.3 FERRAMENTAS DE METODOLOGIA	23
3.3.1 WEKA	23
3.3.2 ANÁLISE ESTATÍSTICA	27
4. ARTIGO CIENTÍFICO	29
5. MATERIAL SUPLEMENTAR	50
5.1 TRATAMENTO DA AMOSTRA.....	50
6. RESULTADOS	51
6.1 CLUSTERIZAÇÃO DA AMOSTRA COM <i>SIMPLE K-MEANS</i>	51
6.2 SELEÇÃO DAS VARIÁVEIS COM <i>SIMPLE LOGISTIC</i>	54
6.3 RESULTADOS DA ANÁLISE ESTATÍSTICA.....	57
6.4 UTILIZAÇÃO DA REDE NEURAL ARTIFICIAL - <i>MULTILAYER PERCEPTRON (MLP)</i>	58
6.5 AMOSTRAS PARA TREINAMENTO E TESTE E APLICAÇÃO DA REDE NEURAL MLP.....	61
7. CONCLUSÕES	63
REFERÊNCIAS BIBLIOGRÁFICAS	65
ANEXO	69

1. INTRODUÇÃO

O número de afetados pelo *diabetes mellitus* (DM), vem aumentando significativamente com o passar dos anos, simulando um processo epidêmico. O crescimento populacional, a urbanização e a maior longevidade são fatores que favorecem o aumento da frequência do diabetes (SBD, 2013-2014). O estilo de vida contemporâneo passou por várias mudanças, permitindo que houvesse a transição dos padrões nutricionais, o que aumentou a prevalência do sedentarismo e da obesidade (PEREIRA *et al.*; 2003; RÔAS e REIS, 2012). Com base no crescimento do diabetes em diferentes populações é possível prever que no ano de 2035, o mundo terá 592 milhões de afetados pela patologia (IDF, 2014).

O *diabetes mellitus* gestacional (DMG) é um tipo de diabetes associado à gestação. A prevalência é de 3-25% (SBD, 2013-2014) das gestações, dependendo da população estudada e dos critérios empregados para o seu diagnóstico. Sua incidência vem subindo paralelamente com o *diabetes mellitus* tipo 2 (DM2) (SCHMIDT *et al.*, 2001; SACKS *et al.*, 2012). O diagnóstico do DMG deve ser precoce para evitar a exposição da gestante e do feto à hiperglicemia crônica, responsável por várias complicações associadas à patologia. A principal terapêutica para o DMG é a orientação alimentar, estabelecendo à gestante um ganho de peso adequado e um controle em seu metabolismo (SCHIRMER, 2000).

Estudos recentes buscam melhorar o diagnóstico ou a predição do DMG através do uso de ferramentas da área da bioinformática. Nagarajan *et al.* 2014, utilizaram o *toolbox* WEKA, um minerador de dados, para aprimorar e melhorar o diagnóstico do DMG. Os algoritmos utilizados foram: ID3, Naïve Bayes, C4.5 e árvores de decisão aleatórias. A melhor acurácia e a menor taxa de erros foi obtida a partir, das árvores de decisão aleatórias. Entre as variáveis utilizadas por estes autores, algumas apresentaram uma capacidade discriminatória relevante na classificação do DMG, como: a concentração de glucose no plasma (2ª hora), o histórico familiar de diabetes e o número de gestações. O trabalho reafirma a necessidade de ferramentas que façam o uso da mineração de dados

no processo de tomada de decisão no campo da medicina, pois podem, eventualmente, melhorar o diagnóstico de doenças como o DMG.

Em outro estudo, Lakshmi e Padmavathamma (2013), modelaram um sistema usando *Feed Forward Neural Network*, uma rede neural artificial, para diagnosticar o DMG. Nesse modelo haviam duas fases, a de análise, onde reuniam as informações do paciente como: idade, histórico familiar, hábitos pessoais, complicações, exames físicos, história clínica e medidas; e a fase de avaliação, que permite o diagnóstico da gestante com DMG, através dos critérios estabelecidos pela Associação Americana de Diabetes (ADA). O programa desenvolvido com este sistema permite que a gestante faça um autoexame para a detecção do DMG.

Gandhi e Prajapati (2014), utilizando o classificador SVM (*Support vector machine*, algoritmo de aprendizado de máquina) conseguiram classificar gestantes diabéticas (do banco de dados *Pima Indian Diabetes Dataset*¹) com 98% de acurácia, 97,77% de sensibilidade e 97,79% de especificidade. Esses autores utilizaram como variáveis para classificação o número de gestações, a concentração de glucose no plasma (2ª hora- TOTG), a pressão arterial diastólica (PAD), a dobra cutânea tricipital, insulina sérica (2 horas), IMC, história familiar de diabetes e idade.

O objetivo do nosso estudo, é utilizar ferramentas de bioinformática para a busca de padrões e diagnóstico associados ao DMG. Neste estudo, enfatizamos o uso de variáveis não glicêmicas, ou seja, aquelas não relacionadas diretamente com a concentração de glicose no sangue (glicemia em jejum; curva glicêmica; HbA1c, frutossamina e 1,5 anidroglicitol). Buscamos, portanto, uma proposta alternativa ao diagnóstico do diabetes gestacional utilizando uma *toolbox* como o programa WEKA, um minerador de dados não comercial e software livre (*open source*) com interface amigável, que congrega múltiplos algoritmos para classificação, clusterização, associação e visualização de dados. Seleccionadas as variáveis não glicêmicas discriminantes, será empregada a rede neural artificial (MLP) para buscar padrões e uma proposta para o diagnóstico automático, caracterizado como de “segunda opinião”.

¹ Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>).

2. OBJETIVOS

2.1 OBJETIVO GERAL

Avaliar ferramentas e estratégias de bioinformática para a triagem do *Diabetes mellitus* gestacional (DMG) utilizando variáveis de fácil obtenção.

2.2 OBJETIVOS ESPECÍFICOS

- Identificar variáveis (NÃO GLICÊMICAS), disponíveis na rotina e de baixo custo, associadas ao *DMG*.
- Empregar a *toolbox* WEKA, com os algoritmos *k-means* e *simple logistic*, na seleção das variáveis não-glicêmicas para a classificação do *DMG*.
- Empregar a rede neural artificial MLP (*multilayer perceptron*) para a análise das variáveis não-glicêmicas selecionadas para desenvolver sistema de triagem automática do *DMG*.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 *DIABETES MELLITUS*

O *diabetes mellitus* (DM) é uma síndrome, caracterizada por hiperglicemia crônica, que afeta cerca de 8% da população mundial, e está associada com elevada morbimortalidade (IDF, 2014). A patologia não tem cura e os afetados evoluem com efeitos em múltiplos órgãos, com destaque para os olhos (retinopatia), rins (nefropatia) e doença vascular (infarto do miocárdio, aterosclerose) (IDF, 2014).

Os afetados pelo diabetes necessitam de tratamento contínuo e as características das complicações associadas ao processo patológico, apresentam elevado custo para o sistema de saúde (SBD 2013-2014).

Outro elemento que torna o DM uma patologia de grande interesse é o crescente número de afetados nas sociedades modernas, o que permite a designação atual de “epidemia de diabetes” (IDF, 2014).

Patofisiologia e Classificação do *diabetes mellitus*

Múltiplos hormônios participam da homeostase da glicose, um carboidrato importante para o metabolismo celular (DRUCKER, 2007).

O DM é uma síndrome resultante: da ausência do hormônio insulina, da deficiência na resposta à insulina ou de ambos os efeitos combinados; com resultante comum em aumento da glicose no sangue ou hiperglicemia (ADA 2015).

Apesar de conhecida há vários séculos e de ser muito estudada, a patofisiologia do DM não é completamente conhecida (MANDAL, 2014). A insulina, um hormônio polipeptídeo produzido por um conjunto de células especializadas do pâncreas (ilhotas de Langherans), é o único hormônio que reduz a glicose do sangue.

A ingestão de alimentos, estimula por diversos mecanismos hormonais, o pâncreas a liberar insulina e este hormônio quando interage com o seu receptor presente em diferentes tecidos, desencadeia uma série de reações no interior da

célula promovendo a entrada de glicose e afetando também outros metabolismos como o de lípidos e proteínas (DRUCKER, 2007). A Figura 1, apresenta o mecanismo de ação da insulina.

Importante ressaltar, como apresentado na Figura 1, que a ação da insulina afeta outras vias metabólicas como a dos lípidos e proteínas. Este efeito pleiotrópico da insulina, foi utilizado no presente estudo, com os biomarcadores não-glicêmicos, que tem potencial para sensoriar a ação da insulina e consequentemente a concentração da glicemia (CHHABRA, 2012).

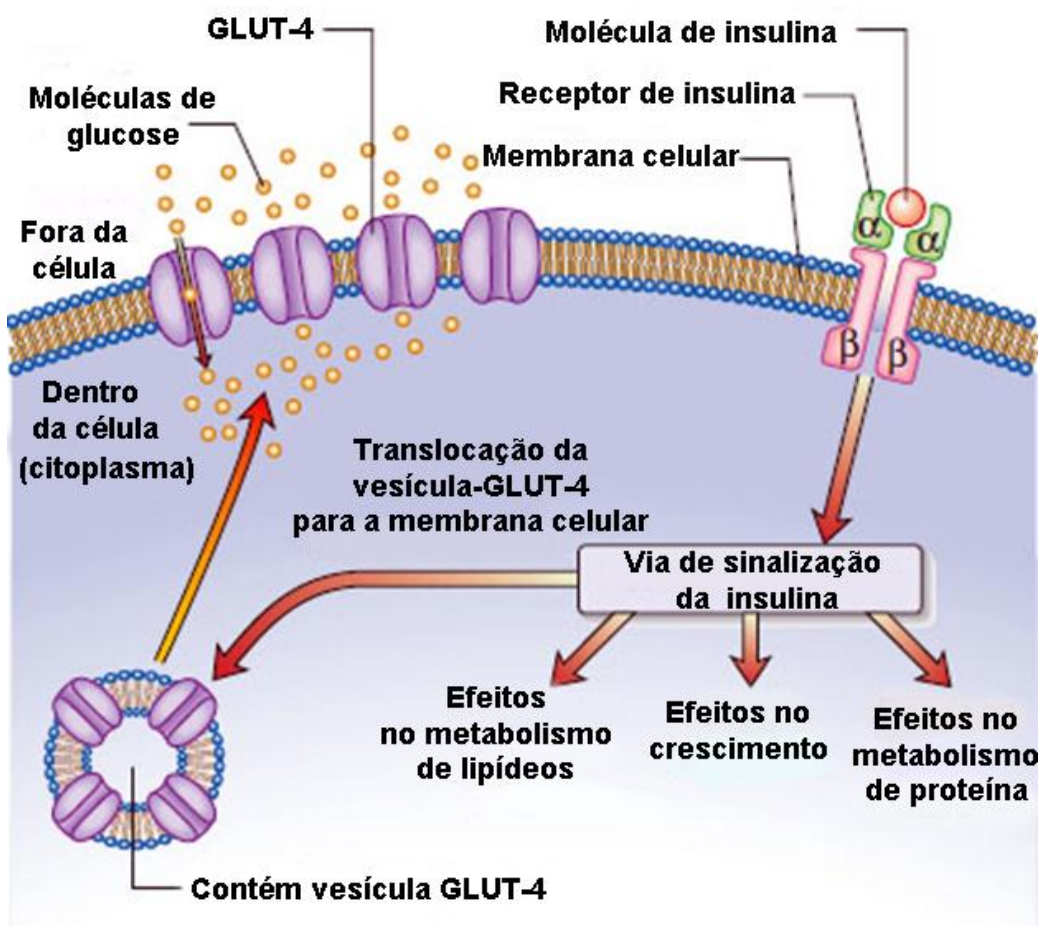


FIGURA 1. MECANISMO DE AÇÃO DA INSULINA

A insulina liga-se ao seu receptor estimulando a atividade intrínseca da tirosina quinase, levando o receptor a autofosforilação e ao recrutamento de sinais moleculares intracelulares (substratos do receptor de insulina - IRS). Os IRS e outros adaptadores de insulina começam a complexa cascata de reação de fosforilação e desfosforilação, difundindo os efeitos metabólicos e mitogênicos da insulina. A ativação do caminho fosfatidilinositol-3'-quinase (PI-3-kinase) estimula a translocação do transportador da glicose (GLUT4) para a superfície a célula, um evento que é crucial para a captação da glicose pelo músculo esquelético e a gordura. A ativação de outras vias de sinalização do receptor da insulina induz a síntese do glicogênio, síntese proteica, lipogênese e a

regulação de vários genes em células que são sensíveis à insulina (CHHABRA, 2012).

O *diabetes mellitus* atualmente é classificado em quatro grupos ou categorias (SBD 2013-2014, ADA 2015):

- (1) Tipo 1 – caracterizado primariamente por reação autoimune, em que o próprio sistema de defesa do organismo destrói as células β do pâncreas, e conseqüentemente afeta a produção de insulina. Essa patologia pode atingir qualquer faixa etária, mas é mais prevalente em crianças e adolescentes. Os afetados necessitam receber terapia de insulina exógena (IDF, 2014). Cerca de 10% dos diabéticos estão neste grupo.
- (2) Tipo 2 – É o tipo mais frequente, sendo responsável por 90% dos casos de diabetes. O elemento central neste tipo de diabetes é a resistência à ação da insulina. O tratamento pode ser realizado com hipoglicemiantes orais e atinge principalmente adultos após os 40 anos de idade.
- (3) Gestacional Tema de estudo desta dissertação que será explorado na sequência.
- (4) Outros tipos de diabetes - Neste grupo estão elencadas mais de cinquenta patologias associadas ao diabetes. O destaque são as patologias que envolvem doenças genéticas (mutações) em genes responsáveis pela produção ou mecanismo de ação da insulina, como o diabetes tipo MODY (diabetes da maturidade de início precoce, *Maturity Onset Diabetes of the Young*).

3.2 DIABETES MELLITUS GESTACIONAL

O *diabetes mellitus* gestacional (DMG) é definido de maneira clássica como a intolerância à glucose com início ou primeiro reconhecimento durante a gestação (BUCHANAN e XIANG, 2005; BUCHANAN *et al.*, 2007; AMERICAN DIABETES, 2011; WHO, 2014). No conceito de DMG ficam excluídas as gestantes que apresentem diabetes (DM1 e DM2) prévio à gestação.

Os principais elementos patofisiológicos do DMG estão apresentados na Figura 2.

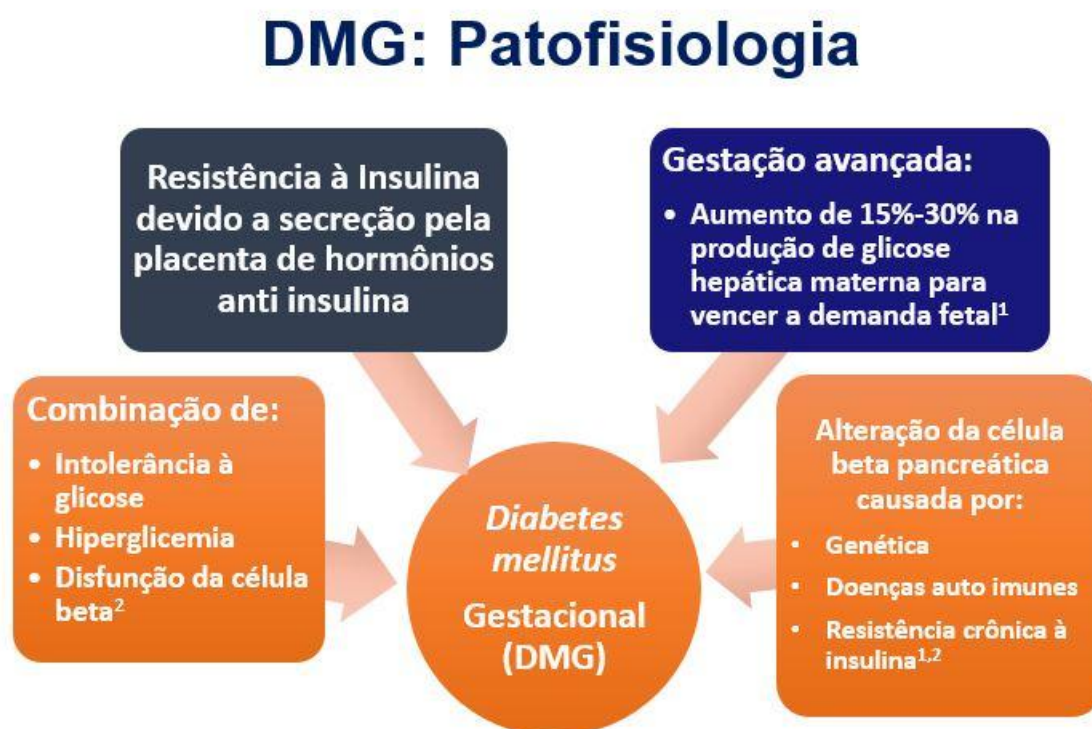


FIGURA 2. ELEMENTOS ASSOCIADOS À PATOFISIOLOGIA DO DIABETES GESTACIONAL.

A resistência à insulina, fisiológica é capitaneada pelos hormônios amplificados na gestação (como o hormônio lactogênio placentário). O aumento da produção de glicose amplifica a disponibilidade de glicose no sangue. Gestantes que apresentem disfunções na célula beta (subclínica) e/ou associação com genes que afetem a tolerância à glicose, propicia o diagnóstico do DMG.

1. Inturrisi M, et al. *Endocrinol Metab Clin N Am*. 2011; 40:703-26.

2. Metzger BE, et al. *Diabetes Care*. 2007; 30(2):S251- 60.

O DMG pode iniciar em qualquer fase da gravidez. Mas o período entre a 24ª e 28ª semanas de gestação é o período onde o diagnóstico é facilitado. Neste período gestacional há o incremento na produção de hormônios produzidos pela placenta, que tem como característica promover resistência à ação da insulina (SBD, 2013-2014). Esta ação hormonal, fisiológica da gravidez, capitaneado pelo hormônio lactogênio placentário favorece, em mulheres com predisposição genética, o desenvolvimento do DMG (ALEPPO e WALKER, 2013).

A frequência do DMG varia de 1-14% das gestações, dependendo da população avaliada e dos critérios empregados para o seu diagnóstico

(AMERICAN DIABETES, 2012). Para a população brasileira a frequência é de aproximadamente 7% das gestações (SHAAT e GROOP, 2007). A incidência do DMG está diretamente relacionada ao aumento da prevalência da obesidade e ao aumento da incidência do DM2 (DESISTO *et al.*, 2010; KWAK; JANG e PARK, 2012; SBD, 2013-2014).

Gestantes com DMG podem, após o parto, apresentar normoglicemia (glicemia normal, inferior a 100 mg/dL) o que ocorre com a maioria, ou podem permanecer como diabéticas, usualmente tipo 2 (SBD, 2013-1024). Pacientes com DMG apresentam risco aumentado para o desenvolvimento de diabetes no futuro (SBD, 2013-1024).

Os fatores de risco associados ao desenvolvimento do DMG são múltiplos (BUCHANAN e XIANG, 2005) com destaque para:

- Idade superior a 35 anos;
- Obesidade marcante, sobrepeso ou excesso de ganho de peso durante a gestação;
- História familiar;
- Baixa estatura (< 1.5m);
- Síndrome do Ovário Policístico (SOP);
- Crescimento fetal excessivo;
- Hipertensão arterial sistêmica (HAS);
- Polidrâmnio;
- Abortos repetitivos;
- Malformações;
- Morte fetal ou neonatal.

Diagnóstico do *diabetes mellitus* gestacional

O diagnóstico do diabetes gestacional realizado segundo os critérios do Ministério da Saúde do Brasil (2001) é apresentado na Figura 3. As amostras analisadas neste estudo foram caracterizadas com base nestes critérios.

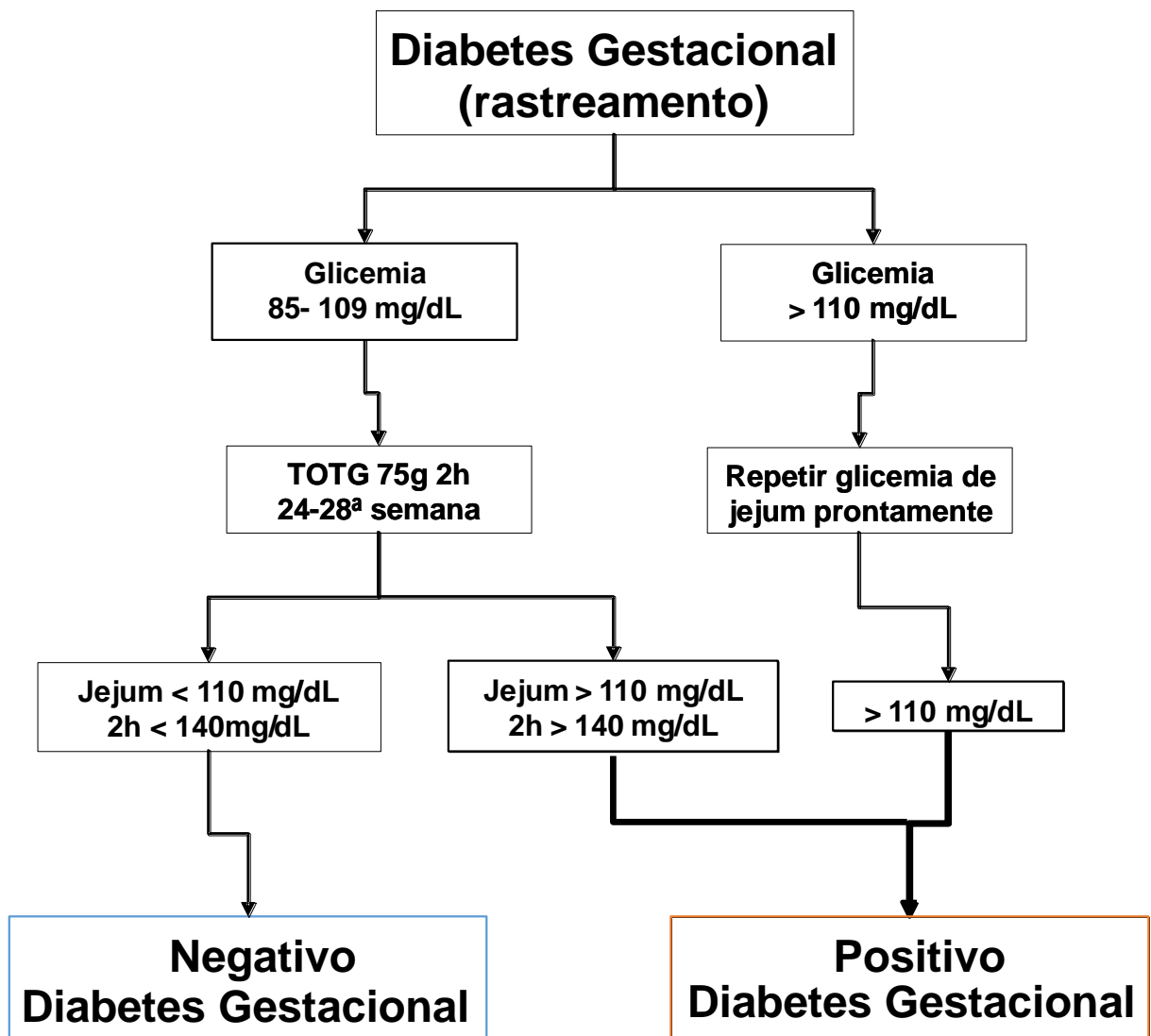


FIGURA 3. DIAGNÓSTICO DO DMG SEGUNDO OS CRITÉRIOS DO MINISTÉRIO DA SAÚDE DO BRASIL (2001).

As amostras estudadas seguem os critérios aqui apresentados.

FONTE: Retirado das informações do Ministério da Saúde do Brasil, 2001 e SBD, 2009.

Mulheres com DMG apresentam alto risco para sérias complicações como eclampsia, infecções urinárias e hemorragia puerperal (KHAN *et al.*, 2006), além de maior propensão para desenvolver DM2 entre 5-10 anos após o parto. As complicações para o feto (FUKS, 2008; DESISTO *et al.*, 2010) incluem:

- Macrossomia;
- Hipoglicemia fetal;
- Distócia do ombro;
- Hipocalcemia.

Na adolescência, os filhos de mães com DMG têm maior chance de apresentarem obesidade, intolerância à glucose ou diabetes, hipertensão e doença cardiovascular (KIM *et al.*, 2002; SIMMONS, 2005).

O diagnóstico precoce do DMG, minimiza a longa exposição dos tecidos à hiperglicemia crônica, que reduz as complicações e as severidades desta doença para a gestante e para o feto (ADA, 2015).

Não há um consenso mundial para o diagnóstico do DMG e diversos protocolos estão em uso. Todas as principais diretrizes centralizam o diagnóstico do DMG nas medidas da glicemia em jejum e após a sobrecarga padronizada com glucose oral. Os valores de corte apresentados vêm sendo modificados e reposicionados nos últimos anos, sem no entanto um consenso global.

A identificação de marcadores e de ferramentas que possam caracterizar a doença precocemente é de grande interesse pois várias complicações estão associadas (METZGER *et al.*, 2007). Estudos demonstram que a detecção e o tratamento precoce trazem benefícios para a mãe e o feto, diminuindo as chances de manifestação das complicações associadas à doença (CROWTHER *et al.*, 2005; LANDON *et al.*, 2009; SACKS, 2009).

Como a ação da insulina afeta o metabolismo de várias vias metabólicas, com ênfase nos lípidos e proteínas, estas biomoléculas podem contribuir para o diagnóstico do diabetes e do DMG.

Na literatura estão descritos alguns marcadores que podem identificar precocemente o risco para o DMG, como o $\log(\text{TG}/\text{HDL-C})$, uma razão denominada índice aterogênico do plasma (SANTOS-WEISS *et al.*, 2013), e outros que podem prever a doença em mulheres que pretendem engravidar como a gama glutamil transferase (SRIDHAR *et al.*, 2014) e a proteína C reativa de alta sensibilidade combinada com a dosagem do hormônio sexual de ligação de globulina (MAGED *et al.*, 2014).

O atendimento de gestantes pelo Sistema Único de Saúde (SUS), como o oferecido pela Prefeitura Municipal de Curitiba, respaldado por protocolos internacionais, realiza exames clínicos, coleta de dados antropométricos e diferentes ensaios laboratoriais destas pacientes. Um dos objetivos destes exames é a identificação de doenças sexualmente transmissíveis como HIV

(vírus da Imunodeficiência Humana), hepatites e tipo sanguíneo da mãe. Outro conjunto de variáveis, apresentado na Tabela 1 e 2, buscam informações para acompanhamento das pacientes e para o diagnóstico do DMG, foco deste estudo. As variáveis apresentadas foram àquelas abordadas neste projeto.

TABELA 1. PRINCIPAIS INFORMAÇÕES E ENSAIOS LABORATORIAIS OBTIDOS DAS GESTANTES EM ESTUDO (VARIÁVEIS NÃO GLICÊMICAS)

Variáveis	Abreviatura	Análise
Idade	ID	Fator de risco, idade de 35 anos ou mais (SBD, 2013-2014).
Peso	P	Apresentar sobrepeso, obesidade ou excesso de ganho de peso durante a gravidez é considerado fator de risco para DMG (SBD, 2013-2014).
Altura	A	Gestantes que apresentam baixa estatura (<1,5m) possuem maior risco de DMG (SBD, 2013-2014).
Índice de Massa Corporal	IMC	Se o IMC for >30kg/m ² , a incidência do DMG chega a ser 1,4 a 20 vezes mais alta (GALTIER-DEREURE; BOEGNER; BRINGER, 2000).
Semanas de Gestação	SEM	Pesquisa de tolerância à glicose na 1ª consulta pré-natal e investigação de DMG entre a 24ª e 28ª semanas de gestação com TOTG (SBD, 2013-2014).
Agrupamento de Semanas de Gestação	GRUPO	A partir das semanas de gestação, a amostra foi dividida em 4 grupos: (1) 12-23 semanas de gestação (2) 24-28 semanas de gestação (3) 29-32 semanas de gestação > 32 semanas de gestação
Pressão Arterial Sistólica	PAS	Valores de pressão arterial ≥140mmHg caracterizam HAS (SBC, 2009-2014).
Pressão Arterial Diastólica	PAD	Valores de pressão arterial ≥90mmHg caracterizam HAS (SBC, 2009-2014).
Colesterol total	COL	Vr: <200 mg/dL.
HDL-C	HDL	Lipoproteínas de alta densidade. Vr: Alto > 60 mg/dL (desejável) Baixo <40 mg/dL(risco para DAC)
LDL-C	LDL	Lipoproteínas de baixa densidade. Vr: <130 mg/dL.
Triglicérides	TG	Concentrações séricas do triglicerídeo associam-se ao DMG. Vrf: <150 mg/dL. (MCGROWDER <i>et al.</i> , 2009)
Log(TG/HDL-C)	Log(TG/HDLC)	Índice aterogênico do plasma (AIP), é um marcador de aterogenicidade no plasma. Vr.: <0,11 - Baixo risco para doença cardiovascular 0,11-0,21 - Risco médio para doença cardiovascular >0,21 - Alto risco para doença cardiovascular

		Calculador do índice aterogênico, disponível em http://www.biomed.cas.cz/fgu/aip/ (TAN, M.H.; JOHNS; GLAZER, 2004).
Colesterol não-HDL-C	nHDL	Reflete o colesterol total menos o colesterol HDL.
Proteína Total	PT	Indica o estado nutricional da gestante e a retenção de líquidos. A gestante com DMG não apresenta perda proteica. Vr.: 6,0-8,0 g/dL
Albumina	ALB	Indica o estado nutricional da gestante e a retenção de líquidos. Vr: 3,5-5,0 g/dL.
Creatinina	CREA	Marcador de função renal. Vr.: 0,6-1,1 mg/dL.
Uréia	UREIA	Marcador de função renal. Vr: 15-45 mg/dL.
Ácido Úrico	URIC	Associado com a resistência à insulina em gestantes com hipertensão gestacional. Vr: 2,6-6,0 mg/dL.
COL/HDL	COL/HDL	Fórmula da razão entre colesterol total e o colesterol HDL.
LDL/HDL	LDL/HDL	Fórmula da razão entre o colesterol LDL e o colesterol HDL.

Os dados estão de acordo com os critérios estabelecidos pelo Ministério da Saúde (2001).
HAS: hipertensão arterial sistêmica; Vr: Valor de referência.

TABELA 2. PRINCIPAIS INFORMAÇÕES E ENSAIOS LABORATORIAIS OBTIDOS DAS GESTANTES EM ESTUDO (VARIÁVEIS GLICÊMICAS)

Variáveis	Abreviatura	Análise
Glicemia de jejum	GLI	Apresentar concentração >110 mg/dL ou após 2h o valor > 140 mg/dL caracteriza o DMG (ADA, 2009).
Hemoglobina Glicada	Hb1Ac	Indica o controle glicêmico, através da média de glicose nos últimos 90 dias. Gestantes com valor de HbA1c \geq 6,5% são caracterizadas como diabéticas (DMG) (SBD, 2013-2014).
1,5 Anidroglicitol	1,5AG	Marcador de controle glicêmico, que mede a concentração média de glicose nas últimas 24-72 horas. Concentrações < 10 μ g/mL caracterizam o DMG (YAMANOUCHI <i>et al.</i> , 1992; SBD, 2013-2014).
Frutosamina	FRUTO	Medição da glicação das proteínas séricas (Albumina) em geral (SBD, 2013-2014)
FRUTO/PT	FRUTO/PT	Fórmula da razão entre a frutosamina e as proteínas totais.
FRUTO/ALB	FRUTO/ALB	Fórmula da razão entre a frutosamina e a albumina.

Os dados estão de acordo com os critérios estabelecidos pelo Ministério da Saúde (2001).

3.3 FERRAMENTAS DE METODOLOGIA

3.3.1 WEKA

O WEKA (atualização 3.7.10) é um software de mineração de dados, que utiliza um conjunto de algoritmos de aprendizado de máquina e ferramentas para pré-processamento de dados (<http://www.cs.waikato.ac.nz/ml/WEKA/>). O programa foi desenvolvido na Universidade de Waikato, na Nova Zelândia. O sistema está escrito na linguagem JAVA, sendo distribuído sob os termos da *General Public License* (GPL), podendo ser implementado em qualquer sistema operacional (Linux, Windows, Macintosh) e plataformas (WITTEN; FRANK; HALL; 2011). A Figura 4 mostra a página inicial e a interface de trabalho do programa.

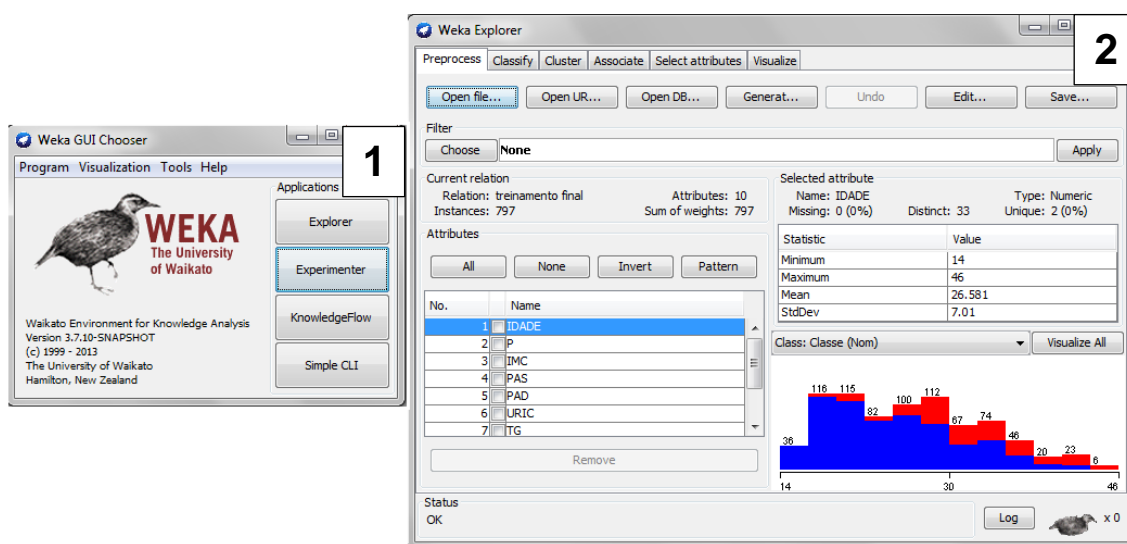


FIGURA 4. PÁGINA INICIAL DO WEKA E A INTERFACE DE TRABALHO.

Página de abertura e interface de seleção de algoritmos do programa WEKA.

1. WEKA GUI Chooser – Página inicial; 2. WEKA Explorer – interface de trabalho.

O WEKA apresenta interface simples, flexível, em que o usuário pode comparar diferentes tipos de métodos e escolher o mais adequado para a solução do seu problema. Os principais métodos para mineração de dados nesse software são a regressão, a classificação, a clusterização, as regras de associação e a seleção de atributos. Os dados são inseridos no WEKA com

arquivos no formato ARFF (*Atributte-Relation File Format*) (Figura 5) (WITTEN; FRANK; HALL; 2011).

```

@relation 'DMG final'

@attribute IDADE numeric
@attribute P numeric
@attribute A numeric
@attribute IMC numeric
@attribute PAS numeric
@attribute PAD numeric
@attribute URIC numeric
@attribute COL numeric
@attribute TG numeric
@attribute PT numeric
@attribute ALB numeric
@attribute HDLC numeric
@attribute LDLC numeric
@attribute 'N_ HDL' numeric
@attribute COL/HDL numeric
@attribute LDL/HDL numeric
@attribute 'LogTG/HDL-C (mmol/L)' numeric

@attribute Classe {1,2}

@data
26,74.3,1.5,32.2,110,70,3.4,215,138,6.9,4.2,42,145,173,5.1,3.5,0.2,1
27,90.5,1.6,34.1,120,60,3.7,137,95,8,4,26,92,111,5.3,3.5,0.2,1
36,67.6,1.6,26.4,120,80,2.5,215,137,7.3,4.3,39,149,176,5.5,3.8,0.2,1
26,63,1.5,26.6,110,70,3.4,199,103,7.6,4.9,39,139,160,5.1,3.6,0.1,1
38,62.8,1.7,21.7,120,70,3.1,188,103,7.5,4.6,38,129,150,5,3.4,0.1,1
29,68,1.6,26.2,100,60,3.1,182,87,7.5,4.6,42,123,140,4.3,2.9,0,1
28,73.5,1.6,28.4,110,80,3,156,123,7,4.2,41,90,115,3.8,2.2,0.1,1
39,54,1.6,21.1,100,70,3.6,254,143,8.4,4.9,43,182,211,5.9,4.2,0.2,1
33,79.6,1.5,33.6,100,70,3.4,216,207,8,4.5,25,150,191,8.6,6,0.6,1

```

(1)

(2)

(3)

FIGURA 5. FORMATO ARFF PARA A INSERÇÃO DOS DADOS AO WEKA.

Legenda: (1), declaração dos atributos (podendo ser do tipo nominal e numérico); (2) atributo sobre a separação das classes (1 ou 2); e (3), é a base de dados.

A melhor forma de utilizar o minerador de dados WEKA, é através da interface da comunicação “EXPLORER”, que permite acessar todos os métodos, e seus respectivos algoritmos. Nesse trabalho foi utilizado na aba *Cluster*, o algoritmo *k-means* e na aba *Classify*, o algoritmo *simple logistic*. Ao final de todo o procedimento, utilizou-se a rede neural *MLP*, da aba *Classify* (Figura 6).

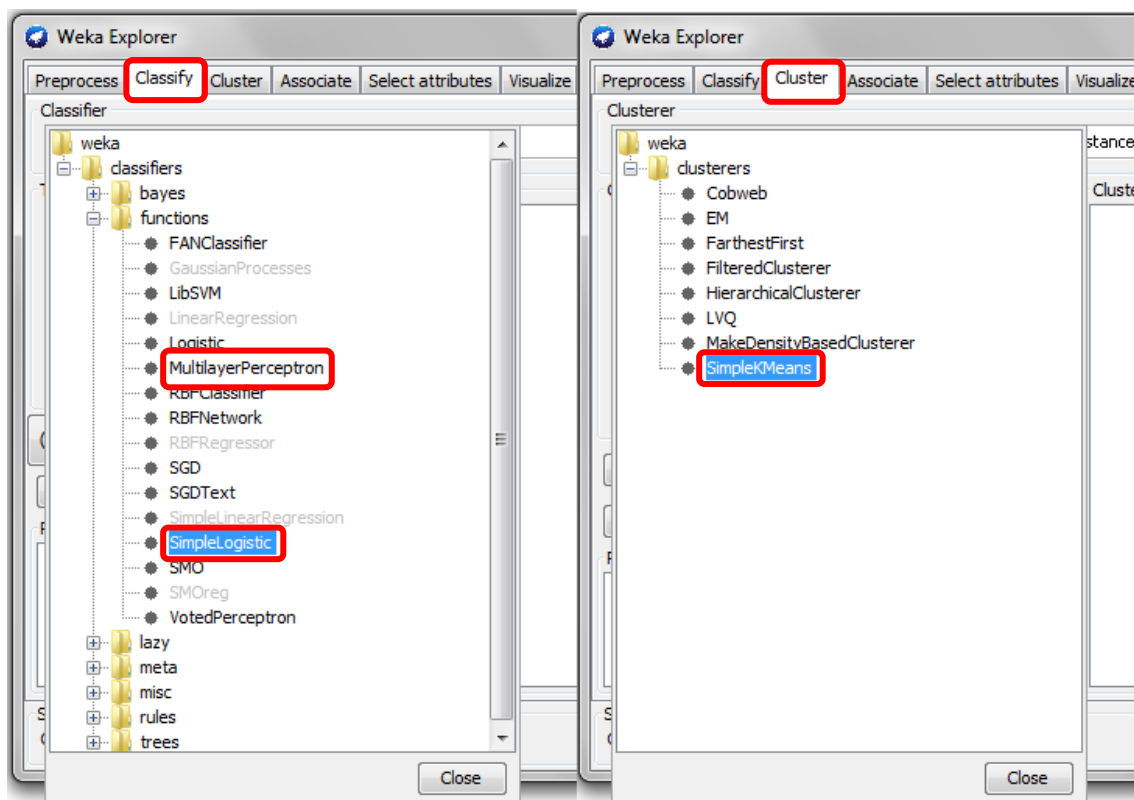


FIGURA 6. ABA CLASSIFY E CLUSTER NA INTERFACE EXPLORER DO PROGRAMA WEKA.

Na aba Classify os algoritmos *Simple Logistic* e *Multilayer Perceptron*.
Na aba Cluster o algoritmo *Simple K-means*.

3.3.1.1 *Simple k-means*

A presença de características irrelevantes em um conjunto de dados, pode piorar a qualidade da aprendizagem, consumindo a memória e o tempo computacional. Na clusterização, a remoção dessas características, não prejudicará a precisão dos agrupamentos quando houver uma redução no armazenamento ou no tempo computacional. No entanto, diferentes características que apresentam relevância, podem produzir agrupamentos diferentes, que podem ajudar a descobrir diferentes padrões ocultos presente nos dados. O *simple k-means* foi aplicado como um método de seleção de variáveis, que elimina características irrelevantes ou redundantes, e mantém características discriminantes, a fim de melhorar a eficiência e qualidade do agrupamento (ALELYANI; TANG; LIU, 2013).

O algoritmo *simple k-means* realiza o agrupamento de objetos com base em atributos. Esse agrupamento ocorre pela minimização da soma dos quadrados das distâncias entre os dados, e corresponde ao centro geométrico de uma característica (*centroid*) (TEKNOMO, 2007).

Essa clusterização, quando utilizada, requer várias interações, e em cada uma delas procura-se encontrar a distância dos centros dos k grupos de cada instância para poder determinar um agrupamento. O k é especificado pela quantidade de *clusters* que se deseja obter no estudo, então pontos k são escolhidos aleatoriamente como centros de *cluster*. Todas as variáveis do estudo são designadas para o centro de *cluster* mais próximo de acordo com a distância Euclidiana. Em seguida, o centroide de cada *cluster* é calculado, e são tomados como novos valores de centro para cada grupo. Todo o processo é realizado novamente com esses novos centros de conjuntos (WITTEN; FRANK; HALL; 2011).

Essa iteração continuará até que os mesmos pontos sejam designados para cada *cluster* em rodadas consecutivas, e que os centroides mantenham uma estabilização (WITTEN; FRANK; HALL; 2011).

3.3.1.2 *Simple logistic*

É um classificador que cria modelos de regressão logística linear (WITTEN; FRANK; HALL; 2011). A regressão logística é uma abordagem para a predição de um desfecho dicotômico, em que analisa a relação entre uma ou mais variáveis que podem predizer a probabilidade de ocorrência de uma determinada doença (Equação I) (DEVORE, 2006).

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad \text{Equação I}$$

Onde $p(x)$ indica a dependência da probabilidade em relação ao valor x , e β_0 e β_1 são variáveis estimadas (DEVORE, 2006).

3.3.1.3 *Multilayer Perceptron*

O *multilayer perceptron* ou MLP, é um classificador que usa *backpropagation* para classificar as instâncias (WITTEN; FRANK; HALL; 2011). É o tipo mais frequente de rede neural utilizada para realizar tarefas como: reconhecimento de padrões, processamento e controle de sinais (CASTRO e CASTRO, 2001; POPESCU *et al.*, 2009). *Backpropagation* é um algoritmo de aprendizagem supervisionada, em que seu processo de treinamento se baseia por tentativa e erro, e isso influencia em seu tempo. Durante o processo, ocorre atualização das taxas de pesos, e se essas taxas definidas tiverem um caráter muito baixo, o tempo de treinamento será prolongado (BARCA; SILVEIRA; MAGINI, 2005).

O MLP possui a capacidade de se adaptar (para pesos e topologias com base em mudanças de ambiente); conseguir ignorar entradas não relevantes e ruídos; ser capaz de modelar funções complexas (não-linear); é de fácil uso; é eficiente em várias áreas de aplicação (ALMEIDA, 2010). Apesar de todas estas vantagens, um dos problemas relevantes ao seu entendimento, é o de apresentar-se como “*out of box*” ou seja um modelo de estilo caixa-preta fechada, o que não permite o conhecimento sobre o relacionamento das funções a serem modeladas. Outro ponto a ser mencionado é o fato de ser susceptível ao *overfitting*, ou seja, à incapacidade de reconhecer padrões não observados durante o treinamento (PINHEIRO e NUNES, 2007).

3.3.2 ANÁLISE ESTATÍSTICA

As variáveis contínuas foram testadas para normalidade com o teste de Kolmogorov-Smirnov. Variáveis contínuas com distribuição normal foram apresentadas como média e 1-desvio padrão. Variáveis com distribuição normal foram comparadas com o teste t-Student (bicaudal).

Uma probabilidade menor que 5% ($P < 0,05$) foi considerada significativa em todas as análises.

O programa *Statistica for windows* versão 8.0 (StatSoft Inc., Tulsa, OK) foi utilizado nas análises.

4. ARTIGO CIENTÍFICO

O presente artigo integra-se a esta dissertação e foi submetido na revista *International Journal of Medical Informatics* (IJMI).

Use of artificial neural network with non-glycemic markers for recognition of gestational *diabetes mellitus*

Ana Paula Filus Bandeira^a, Waldemar Volanski^{a,c}, Izabella Castilhos Ribeiro do Santos-Weiss^{a,c}, Emanuel Maltempi de Souza^{a,b}, Roberto Tadeu Raittz^a, Jeroniza Nunes Marchaukoski^a, Geraldo Picheth^{a,c*}

^a Department of Bioinformatics, Federal University of Parana, Curitiba, Parana, Brazil

^b Department of Biochemistry, Federal University of Parana, Curitiba, Parana, Brazil

^c Department of Clinical Analysis, Federal University of Parana, Curitiba, Parana, Brazil

*Corresponding author: Geraldo Picheth,
Department of Clinical Analysis, Federal University of Parana, Curitiba, Parana, Brazil

Rua Prefeito Lothário Meissner, 632

80210-170 Curitiba, PR, Brazil

Phone/Fax: +55-41-3360-4067

E-mail: geraldopicheth@gmail.com, gpicheth@ufpr.br

Abstract

Background: Gestational *diabetes mellitus* (GDM) is present in about 7% of pregnancies and affects the mother, fetus and newborn. The diagnosis of GDM is

made based on the determination of fasting and postprandial glycemia. Another interesting screening method is the patterns identification based on anthropometric and laboratory characteristics.

Objective: In this work, we used artificial neural networks for automatic identification of GDM using non-glycemic variables.

Methods: We used the software "WEKA" in this project. A sample of 997 pregnant women was used and the subjects were classified as healthy pregnant women (control group, n=699) and gestational diabetic patients (GDM, n=298) according to the glycemic criteria established by American Diabetes Association 2009. *Results:* The discriminant markers selected with the *k-means* and *simple logistic* algorithms were age, weight, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), uric acid (UA), triglycerides (TG), non-HDL cholesterol (nHDL) and Log (TG/HDL-C). These variables were used to feed a neural network MLP (*multilayer perceptron*) able to screen GDM patients. After the training and testing, we found 88% of correct predictions with sensitivity of 92.1%, specificity of 83.8% and accuracy of 87.5%.

Conclusion: The methodology developed in this work is low cost and an effective alternative for GDM screening.

Keywords: gestational diabetes; decision support systems; WEKA; *k-means*; *simple logistic*; artificial neural networks.

1 Introduction

Gestational *diabetes mellitus* (GDM) is defined as glucose intolerance with onset or first recognition during pregnancy, or as diabetes diagnosed in the second or third trimester of pregnancy that is not clearly overt diabetes [1-5].

The frequency of the disease varies from 1-14% depending on the population tested and the criteria used for its diagnosis [6]. For the Brazilian population the frequency is approximately 7% of all pregnancies [7]. The GDM incidence is directly related to the increased prevalence of obesity and type 2 *diabetes mellitus* (DM2) [8-10].

GDM diagnosis is constant changing and different protocols are in use [7, 11]. Several complications are associated with GDM, so the identification of biomarkers and tools that can identify the disease early in pregnancy is necessary [12]. Studies have shown that the early detection and treatment reduces complications associated to the disease [13-15].

Alternative markers to detect GDM are present in the literature. Some of them can identify early risk for GDM, such as the Log (TG/HDL-C), a ratio called atherogenic index of the plasma (AIP) [16]. Others can predict the disease in women who intend to become pregnant, as gamma glutamyl transferase (GGT) [17] and high sensitivity C-reactive protein combined with the measurement of sex hormone binding globulin [18].

These data support the search for early biomarkers in the GDM identification and emphasize the significance of this work. The data analysis of a large number of pregnant women can identify new biomarkers that are not being investigated, as well as set of biomarkers that can predict the disease with high sensitivity and specificity.

Clinical decision support systems (CDSS) are programs designed to help healthcare professionals to make clinical decisions wisely. These systems use artificial intelligence, a computational process that aims to explain and emulate actions intelligently [19]. The early diagnosis and screening of diseases can be obtained by joining biomarkers to the application of bioinformatics. A higher quality of care, efficiency and lower costs compared to what is currently used is a possibility for CDSS. Research and new techniques on this topic can bring many benefits in the future, opening up new possibilities of study and new care protocols and better assistance.

Artificial neural networks (ANN) have been used to develop classification models for medical diagnostic purposes. The advantages of using neural network in diagnosis include its generalization ability considering both the nonlinear and the diffuse nature of the data set [20].

Using neural networks strategies, many authors have demonstrated higher performance and accuracy in predicting clinical outcomes of diabetes diagnosis and other diseases (Table 1).

In this work, we used the WEKA toolbox with *k-means* and simple *logistic* algorithms and the artificial neural network MLP, for the analysis of laboratory variables in the automatic identification of GDM. Through bioinformatics tools, we searched for a low cost and easy way to obtain routine biomarkers (non-glycemic) that can characterize pregnant women with GDM.

Table 1. Artificial neural network strategies for diabetes and other diseases diagnosis.

Techniques	Disease	Accuracy (%)	Sensitivity (%)	Specificity (%)	Reference
<i>k-means</i> , SVM	DM2	94	93	94	Barakat <i>et al.</i> 2010 [21]
<i>k-means</i> , CFS-GA, KNN	DM	96.68	100	88	Karegowda <i>et al.</i> 2012 [22]
RBF neural network, logistic regression	DM2	-	95.2	-	Mansour <i>et al.</i> 2010 [23]
Design ANN with Matlab	Cirrhosis	-	86.6	92.7	Pournik <i>et al.</i> 2014 [24]
Neural network system	DM	92.8	-	-	Kumari and Singh 2012 [25]
<i>k-means</i> , SVM	DM	93.65	-	-	Inan <i>et al.</i> 2014 [26]
Logistic regression, fisher linear, MLP, SVM, fuzzy c-mean, random forests	DM	-	82	100	Tapak <i>et al.</i> 2013 [27]
Bayesian networks, MLP networks, radial basis function and logistic regression	CAD	64	-	-	Purwanto <i>et al.</i> 2012 [28]
MLP, generalized feed forward neural network model and modular neural network models	Cardiac arrhythmia	86.67	93.75	-	Jadhav <i>et al.</i> 2011 [29]
<i>k-means</i> and decision tree	Heart disease	83.9	-	-	Shouman <i>et al.</i> 2013 [30]

SVM: support vector machine; CFS-GA: correlation based feature selection and genetic algorithm; KNN: K-nearest neighbor; RBF: radial basis function; ANN: artificial neural network; MLP: multilayer perceptron; DM2: diabetes mellitus type 2; DM: diabetes mellitus; CAD: coronary artery disease.

2 Material and Methods

The data for this work was obtained at the Clinical Hospital of Federal University of Parana and from the Curitiba Municipal Laboratory. The University's Human Research Ethics Committee approved this study.

2.1 Sample and analyzed parameters

The subjects were classified as healthy pregnant women (control group, n=699) and gestational diabetic patients (GDM, n=298) classified according to the American Diabetes Association criteria [5] and Brazilian Society of Diabetes [10]. For GDM the fasting glucose and glucose 1-hour and 2-hour postprandial with 75 g glucose is used as diagnosed criteria. Pregnant women with other complications despite the GDM like kidney failure, thyroid and heart disease, were not included in this work.

For all pregnant women we analyzed 17 variables including anthropometric data and laboratory parameters described in Table 2, plus the class (0 – GMD; 1 – CTRL). Parameters associated with blood sugar levels as fasting plasma glucose and post-overload, glycated hemoglobin and 1.5 anhydroglucitol were not used.

Table 2. Anthropometric and laboratory parameters.

1. Age (A)	10. Total Protein (TP)
2. Weight (W)	11. Albumin (ALB)
3. Height (H)	12. HDL-C
4. Body Mass Index (BMI)	13. LDL-C
5. Systolic Blood Pressure (SBP)	14. Non-HDL (nHDL)
6. Diastolic Blood Pressure (DBP)	15. COL/HDL
7. Cholesterol (COL)	16. LDL/HDL
8. Uric Acid (URIC)	17. Log TG/HDL-C (mmol/L)
9. Triglycerides (TG)	18. Class 0 or 1

Class 0 - control (CTRL) and Class 1 - gestational *diabetes mellitus* (GDM). The variables abbreviation are in brackets.

2.2 Analysis tools

The software used in the analysis was the open source WEKA 3.7 (<http://www.cs.waikato.ac.nz/ml/weka/>), which has a collection of algorithms to perform the data mining [31]. For classification and variables selection, we used the *k-means* and the *simple logistic* algorithms. The artificial neural network used was the multilayer perceptron (MLP) [32]. All algorithms were used as presented in Weka without modification. The samples were trained and tested to find the best discrimination variables between healthy pregnant women and those with GDM.

Normality was assessed using the Kolmogorov-Smirnov test. Variables with normal distribution were compared with the "t" test for independent samples, and those without normal distribution with Mann-Whitney test. The program Statistica for Windows version 8.0 (StatSoft, Inc. Tulsa, CA, USA) was used for statistical analysis.

A P value lower than 5% ($P < 0.05$) was considered significant in all analysis.

The Figure 1 shows major steps for this project. In the first step, data was inserted in WEKA program for algorithms analysis. By the second step, the selection for variables with better discrimination percentual was performed. In the third step, by the intersection of the obtained markers those most significant were selected; the training time was established; creating files for training and testing in the MLP; and classification of data.

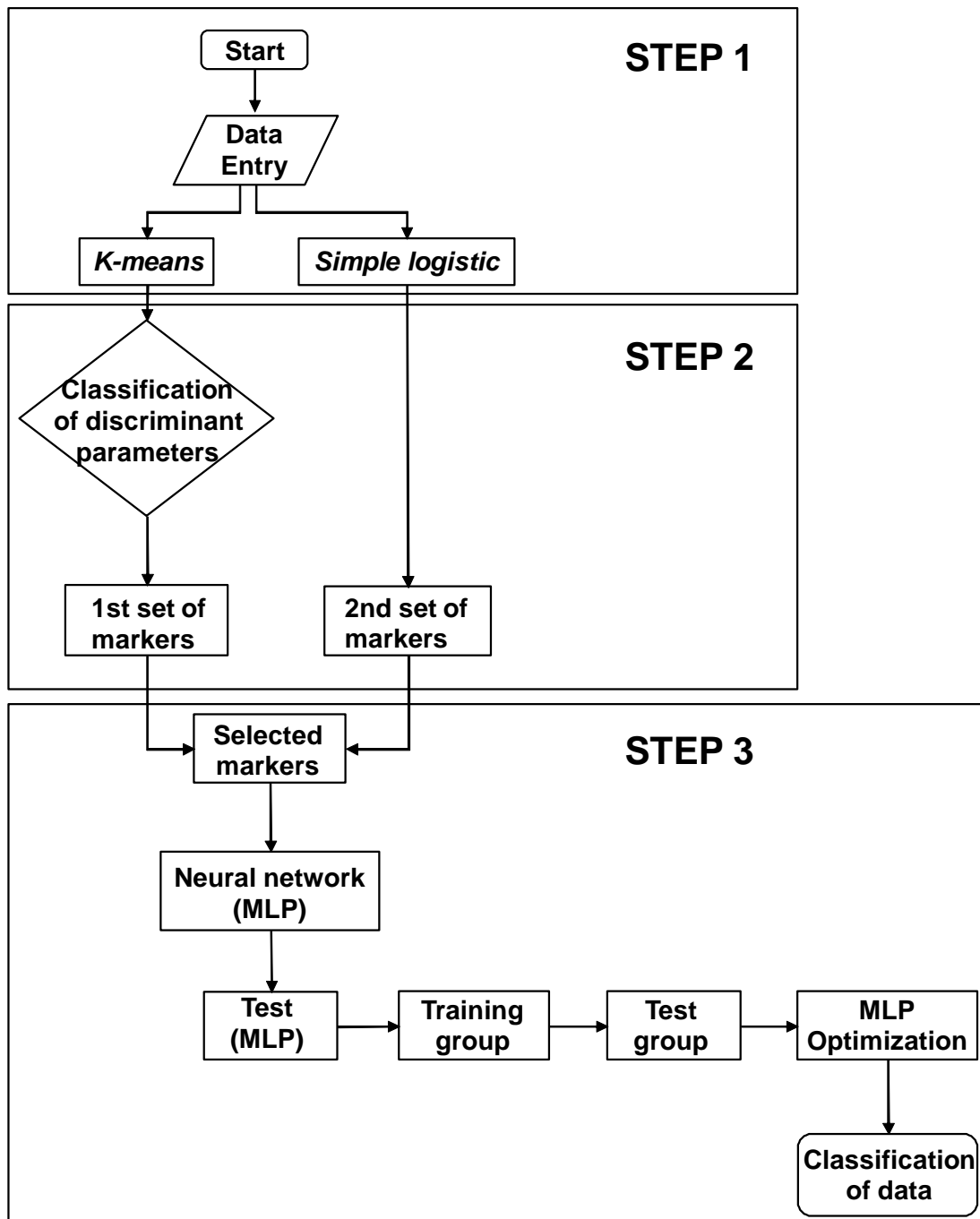


Figure 1. Proposed flowchart for the project.

Step 1. Data entry in WEKA program for algorithms analysis;
 Step 2. Variables selection through *k-means* and *simple logistic* algorithms;
 Step 3. Intersection of the obtained markers; establishment of training time;
 training and testing in the MLP; and classification of data.

3 Results

Simple k-means was applied for selection of variables via elimination of irrelevant or redundant features. This way we could improve the efficiency and quality of group determination. In the first step, we used all variables as a single group and we applied the *k-means* clustering algorithm. With this approach, the GDM group (cluster 1) showed a false positive increment of about 13%.

To improve the corrected classification we analyzed the variables separated in four groups as presented in Table 3. For each group we added and removed the variables step by step and proceed the new clustering. This was done in order to reach the known initial classification (CTRL = 699; GDM = 298) and to have an efficient final set that discriminated the groups.

Table 3. Biomarkers groups and its variables.

Groups	Variables	Categories
Group 1	A, W, H, BMI, SBP, DBP	Anthropometric
Group 2	TP, ALB, URIC	Biochemical markers
Group 3	COL, TG, HDL-C, LDL-C, nHDL	Lipid profile
Group 4	COL/HDL, LDL/HDL, Log (TG/HDL-C)	Lipid ratios

Group 1: anthropometric variables; group 2: biochemical variables associated with kidney function and nutritional status; group 3: biochemical variables of the lipid profile; group 4: ratios of lipid variables obtained by calculation.

With this approach, the variables associated with GDM were A, W, BMI, SBP, DBP, URIC, TG, nHDL, Log (TG/HDL-C). A new clustering with this new group of variables resulted in 50% of false positive reduction when compared with the first results for *k-means*.

The second method used for the classification of variables through logistic regression was the *simple logistic*. It showed the more related variables to both groups (control and GDM) through their weights. The selected variables with this algorithm were A, W, H, BMI, SBP, DBP, URIC, TG, TP, ALB, HDL-C, LDL/HDL, nHDL, Log (TG/HDL-C). We observed with this approach about 5% of false negative to GDM.

Therefore, we decided to combine the variables discriminated by *k-means* (A, W, BMI, SBP, DBP, URIC, TG, N_HDL, Log[TG/HDL-C]) with those selected by *simple logistic* (A, W, H, BMI, SBP, DBP, URIC, TG, PT, ALB, HDL-C, N_HDL, LDL/HDL, Log[TG/HDL-C]) to balance false positive and negative and to improve the GDM discrimination. Table 4 shows the selected markers.

Table 4. Selected variables to discriminated GDM by different classifiers.

Clustering (<i>k-means</i>)	Classification (<i>simple logistic</i>)	Selected markers
A	A	A
W	W	W
BMI	H	BMI
SBP	BMI	SBP
DBP	SBP	DBP
URIC	DBP	URIC
TG	URIC	TG
N_HDL	TG	N_HDL
Log (TG/HDL-C)	PT	Log(TG/HDL-C)
	ALB	
	HDL-C	
	N_HDL	
	LDL/HDL	
	Log(TG/HDL-C)	

In bold are the variables that are common to clustering and classification.

The markers A, W, BMI, SBP, DBP, URIC, TG, N_HDL and Log(TG/HDL-C), were present in the selections made by both simple k-means and simple logistic. And thus they were characterized as the “selected markers”.

The selected markers were compared between the studied groups with classic statistics (Table 5). All markers were statistically different between groups ($P < 0.001$), suggesting that the data obtained from the *k-means* and *simple logistic* algorithms have potential to group classification.

Table 5. Comparison of selected non-glycemic markers between the control and GDM groups.

Selected markers	CTRL	GDM	P
Age (years)	24.8 ± 6.3	31.9 ± 6.1	<0.001
Weight (kg)	64.4 ± 12.2	84.1 ± 17.5	<0.001
BMI (kg/m²)	24.8 ± 4.3	32.9 ± 6.4	<0.001
SBP (mmHg)	106.4 ± 11.9	118.1 ± 12.7	<0.001
DBP (mmHg)	66.0 ± 8.5	74.0 ± 10,2	<0.001
Uric Acid (µmol/L)	214.1 ± 47.6	267.6 ± 61.8	<0.001
Triglycerides (mmol/L)	1.3 ± 0.6	2.6 ± 0.9	<0.001
Non HDL (mmol/L)	3.6 ± 1.2	4.4 ± 1.1	<0.001
Log (TG/HDL-C) (mmol/L)	-0.04 ± 0.2	0.24 ± 0.2	<0.001

Data presented as mean ± 1-SD (standard deviation).

CTRL: control group; GDM: gestational diabetes group. BMI: body mass index; SBP: systolic blood pressure; DBP: diastolic blood pressure; Non-HDL: obtained by the equation (total cholesterol - HDL-C); Log(TG/HDL-C): calculated through the logarithm of (TG/HDL-C) in mmol/L. P, probability: t test for independent samples.

We applied the neural network MLP to classify the studied groups using the selected variables.

To detect overfitting, we verified the relationship between rate of correct prediction x epochs. An optimal of 800 epochs was identified without overfitting (Appendix A).

From study sample, containing 997 pregnant women (CTRL=699 and DMG=298), we selected two groups by random choice and without repetition; Training group (797 subjects) and test group (200 subjects). Both files of training and test were processed by the MLP grid, with the selected markers, through 800 epochs and without changing the network structure, using the default configuration of WEKA. The analysis and results are shown on table 6.

Table 6. MLP performance with selected variables.

MLP		Results	
Training	Confusion Matrix	a	b
		553	46
		47	151
	Classification	88.3% correct 11.7% wrong	
Test <i>(supplied test set)</i>	Confusion Matrix	a	b
		93	7
		18	82
	Classification	87.5% correct 12.5% wrong	
Test Performance Variables	Sensitivity	92.1%	
	Specificity	83.8%	
	Accuracy	87.5%	
	Efficiency	88%	

Training results with the 797 samples (599 - CTRL and 198 - GDM). In training, the confusion matrix explains the amount of trial and error that the program had. Of the 599 CTRL samples, 553 were identified correctly, missing 46 samples; of the 198 GDM samples, 151 were identified correctly, missing 47 samples. For the test set, of the 100 CTRL samples, 93 were identified correctly as CTRL, missing 7; and of the 100 GDM samples, 82 were correctly identified as GDM, missing 18 samples.

The MLP network classified 87.5% of the data correctly with sensitivity of 92.1%, specificity of 83.8%, accuracy of 87.5%, and efficiency of 88%. The results for evaluating and validating the test performance were obtained with EPR-Val Test Pack 2 (www.hutchon.net/EPRval.htm)

4 Discussion

Over the last years, the diagnosis of gestational diabetes (GDM) has been changing significantly [5]. The determination of fasting and after oral glucose load are the central elements for the laboratory diagnosis of GDM [31]. Besides fasting glucose, other glycemc markers, those that reflect or that are directly affected by blood sugar levels, as the concentrations of glycated hemoglobin (HbA1c) and 1.5 anhydroglucitol may also be useful in the diagnosis and monitoring of GDM [5, 34, 35].

Changes in carbohydrate metabolism, characteristic of diabetes, also affect the lipid and protein metabolism [36, 37]. Serum triglyceride concentration increases up to three times in the presence of GDM [38]. The ratio of two

elements of the lipid profile, defined as atherogenic index of plasma (AIP; \log [triglyceride/HDL cholesterol]) allows you to identify low-risk pregnant women for GDM, before 24 weeks of gestation [16]. These examples, among others, suggest that non-glycemic markers may be relevant in identifying GDM.

The use of bioinformatics tools to recognize patterns is well established and has been growing in healthcare [34]. The WEKA software was selected because it is a free platform (no charge) and intuitive, which provides all the desired tools for this study, and is widely used for academic and research purposes.

Initially, in this prospective study, we selected non-glycemic variables associated with GDM, which are easy to obtain and whose routine tests are low cost (Table 4), and are available in care centers to pregnant women in public health services.

Using two sets of pregnant data well characterized as healthy (control) and with GDM, two algorithms, the simple *k-means* and the *simple logistic*, were applied to identify the variables with better power discrimination. The selected variables were those identified as relevant by both algorithms used (Table 4). These were used for pattern recognition studies with MLP neural network.

After training and testing, the MLP identified 88.3% and 87.5%, respectively, for training and test, of corrected cases for GDM. We had a sensitivity of 92.1%, specificity of 83.8%, and accuracy of 87.5% (Table 6).

The sensitivity (92.1%) and accuracy (87.5%) obtained are similar to other works described in literature that use ANN (MLP) and some algorithms to detect diabetes and other diseases (as shown in Table 1).

Studies that use the bioinformatics tools for GDM classification are scarce for comparison. Nagarajan *et al.* [39], took the WEKA tool with the following algorithms: ID3, Naïve Bayes, C4.5 and random decision trees, in order to improve the diagnosis of GDM through the application of data mining techniques. Among the employed algorithms, random decision trees showed better accuracy and low error rate in discriminating the GDM. The variables used in the analysis included glucose concentration in plasma (2hour- OGTT), family history of diabetes, number of pregnancies, DBP, BMI and age. Of these, only three had a vital role in the classification of GDM: the concentration of glucose in plasma

(2hour- OGTT), family history of diabetes and the number of pregnancies. We observed that some variables used by the authors were used in our analysis (DBP, BMI and age); however, when we compare the most significant variables we find that these are different from ours. Population characteristics may help to explain the observed differences. It is important to mention that the use of tools that employ data mining in the decision-making process in medical field can eventually improve the diagnosis of diseases such as gestational diabetes.

Lakshmi and Padmavathamma [20] used Feed Forward Neural Network to model a system to diagnose GDM. This model had two phases; the first one: analysis, containing patient information such as age, family history, gender, personal habits, complications, physical examinations, past history and laboratory measurements; and the second one: evaluation, which diagnoses the patient with GDM through the parameters established by the American Diabetes Association (ADA). The system allows the patient to perform self-examination, for the characterization of GDM. In the same way, our study intends, in the future, after confirming the findings in a larger population, to design a software for automatic and electronic "second opinion" for the diagnosis of GDM for South-Brazilian population.

Using the variables number of pregnancies, glucose concentration in plasma (2hour- OGTT), DBP, triceps skinfold thickness, serum insulin (2 hours), BMI, family history of diabetes and age, in the classifier SVM (Support vector machine), Gandhi and Prajapati (2014) managed to classify diabetic pregnant women (Pima Indian Diabetes Dataset) with 98% of accuracy, 97.77% of sensitivity and 97.79% of specificity [40]. We found that some variables used by these authors were also important in our analysis in the discrimination of GDM and healthy pregnant women (DBP, BMI and age).

The identification system of non-glycemic variables and the application of MLP neural network, allowed the classification of groups (control and diabetic) efficiently. Even with the limited sample size, the prospective feature of the study, the system and the proposed algorithms were robust and simple to be used. Therefore, the presented approach has the potential to identify the variables of interest in other studies with similar characteristics.

Further studies with larger sample size should be conducted with the proposed structure. The presented process using non-glycemic routine and inexpensive variables has the potential to recognize patterns and provide alternative support for the diagnosis of GDM. The selected open source software, contributes to the widespread use, as desired by the public health service, where diagnosis costs are relevant factors in the adoption of new procedures.

The proposed methodology combined with blood glucose tests established in the guidelines for the diagnosis of diabetes has potential to facilitate the identification of the disease and possibly extend the spectrum of risk stratification for gestational diabetes.

In summary, the *k-means* and *simple logistic* algorithms, with emphasis in MLP neural network, using non-glycemic variables such as age, weight, body mass index, systolic and diastolic blood pressure, uric acid, triglycerides, non-HDL cholesterol and the ratio log (triglycerides / HDL-cholesterol), allowed the discrimination of pregnant women with gestational diabetes from healthy pregnant women with high sensitivity, specificity and accuracy (higher than 80%).

This approach has potential application in public health services as decision support tool in the diagnosis of gestational diabetes.

Authors' contributions

Following are the contributions of the authors to the work described in this paper.

Ana Paula Filus Bandeira: development of the methodology, data analysis and the writing of this manuscript. Waldemar Volanski: development of the methodology, data analysis and critical review. Izabella Castilhos Ribeiro do Santos-Weiss: statistical analysis, interpretation of data, data collection and the writing of this manuscript. Emanuel Maltempi de Souza: quantification and analysis of laboratory parameters. Roberto Tadeu Raittz: analysis and validation of the proposed methodology. Jeroniza Nunes Marchaukoski: conception and study design, data analysis and critical review of the manuscript. Geraldo Picheth: responsible for conception and study design, statistical analysis, critical review and final approval of the manuscript.

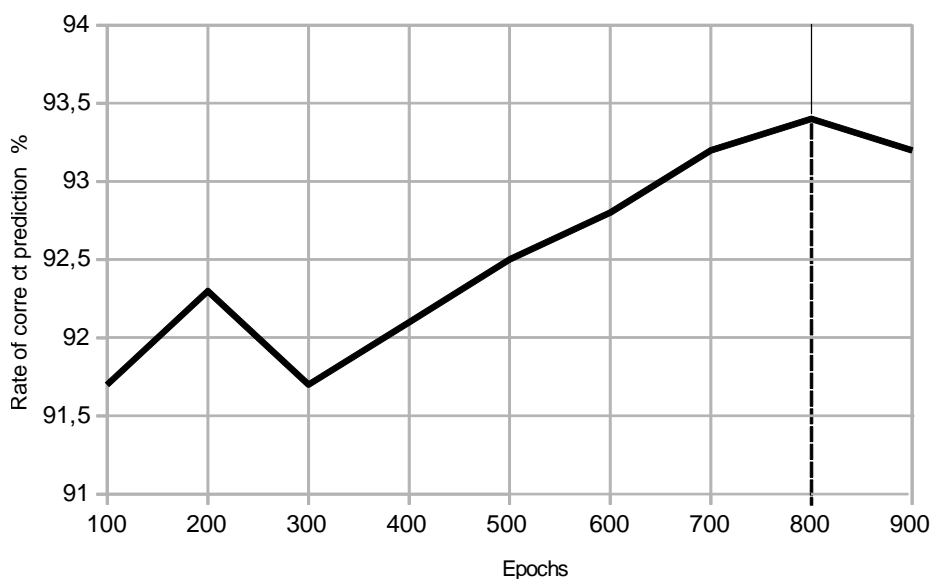
Conflicts of interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and Fundação Araucária supported this work.

Appendix A



Appendix A. Rate of correct prediction x Epochs
Training time optimized in 800 epochs (dotted line).

Summary table

What is already known on this topic?

- Gestational diabetes mellitus is a frequent disease and important for health care system
- Diagnostic of gestational diabetes is based on serum glycemia and the recognition of the disease is relevant for the pregnant and fetus.

- Artificial neural networks have been used to identify diabetes

What this study has added to our knowledge?

- We proposed identify gestational diabetes through non-glycemic biomarkers with multilayer perceptron (MLP) neural network.
- Variables were selected combine *k-means* and logistic regression in the software Weka toolbox
- The approach allowed screening gestational diabetes from healthy pregnant women with accuracy of 87%.

References

[1] T.A. Buchanan, A.H. Xiang. Gestational diabetes mellitus. J Clin Invest. 115 (2005) 485–91. doi: 10.1172/JCI200524531

[2] T.A. Buchanan, A. Xiang, S.L. Kjos, R. Watanabe. What is gestational diabetes? Diabetes Care. 30 Suppl 2 (2007) S105–11. doi: 10.2337/dc07-s201

[3] A. American Diabetes. Diagnosis and classification of diabetes mellitus. Diabetes Care. 34 Suppl 1 (2011) S62–69. doi: 10.2337/dc11-S062

[4] WHO – World Health Organization. Diabetes. <<http://www.who.int/mediacentre/factsheets/fs312/en/>>. Access: 2014/12/09.

[5] American Diabetes Association. Standards of medical care in diabetes. Diabetes Care. 38 Suppl 1 (2015) S8–16. doi: 10.2337/dc15-S005

[6] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 35 Suppl 1 (2012) S64–71. doi: 10.2337/dc12-s011

[7] N. Shaat, L. Groop. Genetics of gestational diabetes mellitus. Curr Med Chem. 14 (2007) 569–83. doi: 10.2174/092986707780059643

- [8] C.L. Desisto, S.Y. Kim, A.J. Sharma. Prevalence Estimates of Gestational Diabetes Mellitus in the United States, Pregnancy Risk Assessment Monitoring System (PRAMS), 2007–2010. *Prev Chronic Dis*.11 (2014) 130415. doi: 10.5888/pcd11.130415
- [9] S.H. Kwak, H.C. Jang, K.S. Park. Finding genetic risk factors of gestational diabetes. *Genomics Inform*. 10 (2012) 239–43. doi: 10.5808/GI.2012.10.4.239
- [10] SBD - Diretrizes da Sociedade Brasileira do Diabetes. *Diabetes mellitus gestacional: diagnóstico, tratamento e acompanhamento pós-gestação*. Rio de Janeiro: Gen, 2013-2014.
- [11] H. Long. Diagnosing gestational diabetes: can expert opinions replace scientific evidence? *Diabetologia*. 54 (2011) 221-2213. doi: 10.1007/s00125-011-2228-z
- [12] B.E. Metzger, T.A. Buchanan, D.R. Coustan, A. de Leiva, D.B. Dunger, D.R. Hadden, M. Hod, J.L. Kitzmiller, S.L. Kjos, J.N. Oats, D.J. Pettitt, D.A. Sacks, C. Zouzas. Summary and recommendations of the Fifth International Workshop-Conference on Gestational Diabetes Mellitus. *Diabetes Care*.30 Suppl 2 (2007) S251–60. doi: 10.2337/dc07-s225
- [13] C.A. Crowther, J.E. Hiller, J.R. Moss, A.J. McPhee, W.S. Jeffries, J.S. Robinson, Australian Carbohydrate Intolerance Study in Pregnant Women (ACHOIS) Trial Group. Effect of treatment of gestational diabetes mellitus on pregnancy outcomes. *N Engl J Med*. 352 (2005) 2477–86. doi: 10.1056/NEJMoa042973
- [14] M.B. Landon, C.Y. Spong, E. Thom, M.W. Carpenter, S.M. Ramin, B. Casey, R.J. Wapner, M.W. Varner, D.J. Rouse, J.M. Jr. Thorp, A. Sciscione, P. Catalano, M. Harper, G. Saade, K.Y. Lain, Y. Sorokin, A.M. Peaceman, J.E. Tolosa, G.B. Anderson; Eunice Kennedy Shriver National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network. A multicenter,

randomized trial of treatment for mild gestational diabetes. *N Engl J Med.* 361 (2009) 1339–48. doi: 10.1056/NEJMoa0902430

[15] D.A Sacks. Gestational diabetes–whom do we treat? *N Engl J Med.* 361 (2009) 1396–8. doi: 10.1056/NEJMe0907617

[16] I.C.R. Santos-Weiss, R.R. Réa, C.M.T. Fadel-Picheth, F.G.M. Rego, F.O Pedrosa, P. Gillery, E.M. Souza, G. Picheth. The plasma logarithm of the triglyceride/HDL-cholesterol ratio is a predictor of low risk gestational diabetes in early pregnancy. *Clin Chim Acta.* 418 (2013) 1-4. doi: <http://dx.doi.org/10.1016/j.cca.2012.12.004>

[17] S.B. Sridhar, F. Xu, J. Darbinian, C.P. Quesenberry, A. Ferrara, M.M. Hedderson. Pregravid liver enzyme levels and risk of gestational diabetes mellitus during a subsequent pregnancy. *Diabetes Care.* 37 (2014) 1878-1884. doi: 10.2337/dc13-2229

[18] A.M. Maged, G.A. Moety, W.A. Mostafa, D.A. Hamed. Comparative study between different biomarkers for early prediction of gestational diabetes mellitus. *J Matern Fetal Neonatal Med.* 27 (2014) 1108-1112. doi: 10.3109/14767058.2013.850489

[19] R.J. *Schalkoff. Artificial Intelligence: An Engineering Approach.* McGraw-Hill, 1990.

[20] K.V. Lakshmi, M. Padmavathamma. Modeling an Expert System for Diagnosis of Gestational Diabetes Mellitus Based On Risk Factors. *J Computer Eng (IOSRJCE).* 8 (2013) 29-32. ISSN: 2278-0661, ISBN: 2278-8727

[21] N.H. Barakat, A.P. Bradley, M.N.H. Barakat. Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. *Trans Info Tech Biomed.* 14 (2010) 4. Doi: 10.1109/TITB.2009.2039485.

- [22] A.G. Karegowda, M.A. Jayaram, A.S. Manjunath. Cascading *K-means* Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients. *Int J Eng Adv Tech.* 1 (2012). e-ISSN: 2249 – 8958.
- [23] R. Mansour, Z. Eghbal, H. Amirhossein. Comparison of Artificial Neural Network, Logistic Regression and Discriminant Analysis Efficiency in Determining Risk Factors of Type 2 Diabetes. *W App Sci J.* 23 (2013). e-ISSN: 1818-4952.
- [24] O. Pournik, S. Dorri, H. Zabolinezhad, S.M. Alavian, S. Eslami. A diagnostic mode for cirrhosis in patients with non-alcoholic fatty liver disease: an artificial neural network approach. *Med J Islam Repub Iran.* 28 (2014).
- [25] S. Kumari, A. Singh. A Data Mining Approach for the Diagnosis of Diabetes Mellitus. *Int Conf Intel Sys Contr.* (2013). Doi: 10.1109/ISCO.2013.6481182.
- [26] O. Inan, N. Yilmaz, M.S. Uzer. A New Data Elimination Method Based on Clustering Algorithms for Diagnosis of Diabetes Diseases. *Glob J tech.* 5 (2014).
- [27] L. Tapak, H. Mahjub, O. Hamidi, J. Poorolajal. Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran. *Healthc Inform Res.* 19 (2013). Doi: <http://dx.doi.org/10.4258/hir.2013.19.3.177>.
- [28] Purwanto, C. Eswaran, R. Logesswaran, A.R.A. Rahman. Prediction models for early risk detection of cardiovascular event. *J Med Sys.* 36 (2012)
- [29] S.M. Jadhav, S.L. Nalbalwar, A.A. Ghatol. Artificial Neural Network Based Cardiac Arrhythmia Disease Diagnosis. *Int Conf Process Autom Contr Comp.* (2011). Doi: 10.1109/PACC.2011.5979000.
- [30] M. Shouman, T. Turner, R. Stocker. Integrating Decision Tree and *K-means* Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients. *Int Conf Data Mining.* 20 (2013). e-ISSN: 2043-9091.

[31] I.H. Witten, E. Frank, Hall, A. Mark. Data Mining – Practical Machine Learning Tools and Techniques. 3 ed. United States: Elsevier, 2011.

[32] A. Krogh. What are artificial neural networks? *Nature Biotechnol.* 26 (2008) 195-7. doi: 10.1038/nbt1386

[33] HAPO Study Cooperative Research Group, B.E. Metzger, L.P. Lowe, A.R. Dyer, E.R. Trimble, U. Chaovarindr, D.R. Coustan, D.R. Hadden, D.R. McCance, M. Hod, H.D. McIntyre, J.J. Oats, B. Persson, M.S. Rogers, D.A. Sacks. Hyperglycemia and adverse pregnancy outcomes. *N Engl J Med.* 358 (2008) 1991–2002. doi: 10.1056/NEJMoa0707943

[34] K.C. Boritza, I.C.R. Santos-Weiss, A.S.C. Alves, R.R. Réa, F.O. Pedrosa, E.M. Souza, G. Picheth, F.G.M. Rego. 1,5 Anhydroglucitol serum concentration as a biomarker for screening gestational diabetes in early pregnancy. *Clin Chem Lab Med.* 52 (2014) e179-181. doi: 10.1515/cclm-2013-1042

[35] L.S. Weinert. International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy. *Diabetes Care.* 33 (2010) 676-682. doi: 10.2337/dc10-0544

[36] N.F. Butte. Carbohydrate and lipid metabolism in pregnancy: normal compared with gestational diabetes mellitus. *Am J Clin Nut.* 71 (2000) 1256S–61S.

[37] E. Koukkou, G.F. Watts, C. Lowy. Serum lipid, lipoprotein and apolipoprotein changes in gestational diabetes mellitus: a cross-sectional and prospective study. *J Clin Pathol.* 49 (1996) 634–7.

[38] B.E. Metzger, R.L. Phelps, N. Freinkel, I.A. Navickas. Effects of gestational diabetes on diurnal profiles of plasma glucose, lipids, and individual amino acids. *Diabetes Care.* 3 (1980) 402–9.

[39] S. Nagarajan, R.M. Chandrasekaran, P. Ramasubramanian. Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes. Int J Curr Res Acad Rev.2 (2014).91-98.

[40] K.K. Gandhi, N.B. Prajapati. Diabetes prediction using feature selection and classification. Int J Adv Eng Res Dev. 1 (2014) 1-7. e-ISSN: 2348 - 4470

5. MATERIAL SUPLEMENTAR

Esse material suplementar fornece detalhes do procedimento realizado durante a aplicação da rede neural artificial para o reconhecimento do *diabetes mellitus* gestacional (DMG). Informações detalhadas sobre a metodologia empregada e o processamento e tratamento das amostras antes da aplicação da ferramenta são apresentados.

5.1 TRATAMENTO DA AMOSTRA

A amostra inicialmente coletada possuía 1006 gestantes com 28 variáveis, sendo 699 classificadas como gestantes controle (CTRL) e 307 como gestantes com *diabetes mellitus* gestacional (DMG). Os dados foram predispostos no *Excel*, relacionando cada paciente com suas variáveis. Durante a análise, foi identificado que algumas amostras não apresentavam todos os dados, e a fim de evitar algum erro durante o processamento, essas amostras foram excluídas das análises, o que resultou em um conjunto de 997 amostras, sendo 699 do grupo CTRL e 298 do grupo DMG.

Em relação as variáveis, houve a retirada do número da amostra, por ser um rótulo; da hemoglobina glicada, que é um diferencial somente para as pacientes com DMG; de variáveis glicêmicas, como: Glicose, Frutosamina, fórmulas relacionadas e 1,5 Anidroglicitol; e da Creatinina e Ureia, por estarem relacionados com doença renal. As semanas de gestação e o grupamento de semanas de gestação foram retirados do estudo por serem variáveis discriminantes para DMG. Por fim o Log (TG/HDL-C) em mg/dl, foi também excluído do estudo, por apresentar valores de triglicérides não normalizados. O Log (TG/HDL-C) em mmol/L apresenta os valores de triglicérides normalizados e corrigidos para a unidade mmol/L, sendo então incluídos no estudo. O total final de variáveis foi de 17 mais a classe.

6. RESULTADOS

6.1 CLUSTERIZAÇÃO DA AMOSTRA COM SIMPLE K-MEANS

O primeiro algoritmo utilizado nesse estudo é o *k-means*, na aba *cluster* do WEKA. Realizou-se a análise de clusters com todos os atributos, sem a classe, no *cluster mode* em *Use training set*. Os clusters estabelecidos são: cluster 0 que é dito para pacientes CTRL e o cluster 1 para pacientes com DMG. Na Tabela 3, é apresentada a análise desse processo com todos os dados, estabelecendo 573 amostras como CTRL (*cluster 0*) e 424 como DMG (*cluster 1*). Com esse método o grupo DMG apresentou 13% de falsos positivos (Tabela 4).

TABELA 3. CLUSTERIZAÇÃO PELO ALGORITMO K-MEANS COM AS VARIÁVEIS SELECIONADAS, ENFATIZANDO A MÉDIA.

Variáveis	Todos (997)	Cluster 0 CTRL (573)	Cluster 1 DMG (424)
ID	26.94	24.08	30.80
P	70.26	62	81.42
A	1.61	1.62	1.61
IMC	27.24	23.88	31.79
PAS	109.90	105.97	115.23
PAD	68.34	65.66	71.97
URIC	3.87	3.50	4.37
COL	202.77	177.44	237.00
TG	150.97	101.15	218.30
PT	6.79	7.00	6.50
ALB	3.85	4.09	3.53
HDLC	54.07	54.18	53.93
LDLC	118.48	103.01	139.40
nHDL	148.70	123.26	183.08
COL/HDL	3.91	3.41	4.59
LDL/HDL	2.32	2.01	2.74
LogTG/HDL-C(mmol/L)	0.04	-0.10	0.24
Cluster 0 (CTRL)		573 (57%)	
Cluster 1 (DMG)			424 (43%)

Unidades e características das variáveis descritas na Tabela 1.

Resultado estabelecido pela clusterização: 573 amostras como controle (*cluster 0*) e 424 amostras como DMG (*cluster 1*). Classificação conhecida das amostras (padrão inicial) CTRL: 699 e DMG:298.

TABELA 4. COMPARAÇÃO E ANÁLISE COM O K-MEANS

Etapas	Referência Marcadores Glicêmicos	Teste – <i>k-means</i> Marcadores não-glicêmicos
1	Total = 997 = 100%	Total = 997 = 100%
	Controles 699 = 70,1%%	Controles = 573 = 57,5%%
	DMG = 298 = 29,9%	DMG = 424 = 42,5%
2	Comparação entre teste e referencia	
	DMG 42,5% - 29,9% = 12,6% (Falso positivo)	

Referência, classificação da amostra pela glicêmica de jejum, como descrito na Figura 3. Teste, resultados obtidos com o algoritmo *k-means*.

Etapa 1, Análise pelo *k-means*.

Etapa 2. Diferença observada entre a referência e o obtido pelo *k-means*.

Para melhorar a clusterização, as variáveis foram combinadas em 4 grupos, o primeiro grupo contendo o conjunto das variáveis antropométricas; o segundo, variáveis bioquímicas; o terceiro, variáveis bioquímicas do perfil lipídico; e o quarto, as razões entre as variáveis bioquímicas do perfil lipídico obtidas por cálculo (Tabela 5).

TABELA 5. GRUPOS ESTABELECIDOS POR SUAS VARIÁVEIS.

Grupos	Variáveis	Classificação
Grupo 1	ID, P, A, IMC, PAS, PAD	Variáveis antropométricas
Grupo 2	URIC, PT, ALB	Variáveis bioquímicas
Grupo 3	COL, TG, HDLC, LDLC, nHDL,	Variáveis bioquímicas de perfil lipídico
Grupo 4	COL/HDL, LDL/HDL, Log(TG/HDL)	Razões entre variáveis bioquímicas do perfil lipídico obtidas por cálculo

As abreviaturas das variáveis estão descritas na Tabela 1 e 2.

Para cada grupo formado, foi realizado o processo de adição e remoção das variáveis, seguido da clusterização, a fim de que o resultado se aproximasse da classificação inicial de referência (CTRL=699; DMG=298). As variáveis finais selecionadas pelo algoritmo *k-means* foram: ID, P, IMC, PAS, PAD, URIC, TG, N_HDL e Log(TG/HDL-C) (mmol/L) (Tabela 6). Para os grupos estudados, o

algoritmo *k-means* apresentou resultado falso positivo, número maior de doentes quando comparado à referência, para o grupo DMG.

TABELA 6. VARIÁVEIS SIGNIFICATIVAS E EXCLUÍDAS APÓS O PROCESSO DE CLUSTERIZAÇÃO.

Grupos	Variáveis significativas	Resultados 1	Variáveis Excluídas
Grupo 1	ID, P, IMC, PAS, PAD	Cluster 0 – 649 (65,1%) Cluster 1 – 348 (34,9%)	A
Grupo 2	URIC	Cluster 0 – 671 (67,3%) Cluster 1 – 326 (32,7%)	PT, ALB
Grupo 3	TG, nHDL	Cluster 0 – 673 (67,5%) Cluster 1 – 324 (32,5%)	COL, HDL-C, LDL-C
Grupo 4	Log (TG/HDL-C)	Cluster 0 – 671 (67,3%) Cluster 1 – 326 (32,7%)	COL/HDL, LDL/HDL

Cluster 0, controle (gestantes saudáveis); Cluster 1, DMG, gestantes com diabetes gestacional.

As variáveis significativas (PS) são aquelas que após a clusterização apresentaram uma classificação próxima à referência (CRTL – 699 (70,1%), DMG – 298 (29.9%)). As variáveis excluídas (PE) são aquelas que não modificaram a classificação dos grupos no processo de clusterização pelo *k-means*.

Nova clusterização com as variáveis significativas combinadas foi realizada, e o resultado está descrito na Tabela 7. O algoritmo *k-means* neste ensaio classificou 624 (62,6%) das amostras como saudáveis (CRTL) e 373 (37,4%) como DMG, resultando em 7,5% de falsos positivos, quando comparado à referência.

Com a clusterização das variáveis combinadas após a seleção por grupos, uma redução dos falsos positivos foi observada (tabela 8 - 12,6% vs. 7,5%). Embora os resultados falsos positivos tenham reduzido com a seleção das variáveis, este efeito não foi eliminado. Com as análises realizadas, temos como hipótese que o algoritmo *k-means* apresenta tendência para resultados falsos positivos, pelo menos nas condições de ensaio apresentadas neste estudo.

TABELA 7. RESULTANTE DA ANÁLISE DAS VARIÁVEIS SELECIONADAS COMBINADAS DA CLUSTERIZAÇÃO COM O ALGORITMO K-MEANS.

<i>k-means</i>	Variáveis selecionadas
	ID P IMC PAS PAD URIC TG N_HDL Log (TG/HDL-C)
Resultado	<i>Cluster 0 – 624 (62,6%)</i> <i>Cluster 1 – 373 (37,4%)</i>

Cluster 0: Grupo controle (CTRL) e cluster 1: grupo DMG. Classificação conhecida das amostras (padrão inicial) CTRL: 699 (70,1%) e DMG:298 (29,9%).

TABELA 8. COMPARAÇÃO E ANÁLISE COM AS VARIÁVEIS SELECIONADAS COM O K-MEANS

Etapas	Referência Marcadores Glicêmicos	Teste – <i>k-means</i> Marcadores não-glicêmicos
1	Total = 997 = 100%	Total = 997 = 100%
	Controles 699 = 70,1%%	Controles = 624 = 62,6%
	DMG = 298 = 29,9%	DMG = 373 = 37,4%
2	Comparação entre teste e referencia	
	DMG 37,4% - 29,9%= 7,5% (Falso positivo)	

Referência, classificação da amostra pela glicêmica de jejum, como descrito na Figura 3. Teste, resultados obtidos com o algoritmo *k-means*.

Etapa 1, análise pelo *k-means*.

Etapa 2. Diferença observada entre a referência e obtido pelo *k-means*.

6.2 SELEÇÃO DAS VARIÁVEIS COM SIMPLE LOGISTIC

Para nova classificação e seleção das variáveis, a regressão logística foi aplicada, através do algoritmo *simple logistic*, mostrada na Tabela 9. As variáveis selecionadas pelo algoritmo foram: ID, P, A, IMC, PAS, PAD, URIC, TG, PT, ALB, HDL-C, N_HDL, LDL/HDL, Log(TG/HDL-C).

TABELA 9. VARIÁVEIS SELECIONADAS COM O ALGORITMO SIMPLE LOGISTIC.

Classe 0 (CTRL)	2.2	Classe 1 (DMG)	- 2.2
ID	-0.05	ID	0.05
P	0	P	0
A	0.93	A	0.93
IMC	-0.09	IMC	0.09
PAS	-0.02	PAS	0.02
PAD	-0.01	PAD	0.01
URIC	-0.34	URIC	0.34
TG	-0.01	TG	0.01
PT	0.09	PT	-0.09
ALB	1.34	ALB	-1.34
HDLC	-0.01	HDLC	0.01
nHDL	0	nHDL	0
LDL/HDL	0.03	LDL/HDL	-0.03
Log(TG/HDL-C) (mmol/L)	-1.13	Log(TG/HDL-C) (mmol/L)	1.13

Ao lado das variáveis são apresentados os pesos para cada classe (0 e 1).

Essa classificação identificou 667 (66,9%) amostras como controle, e 256 (25,7%) amostras como DMG (Figura 7). Essa abordagem mostrou cerca de 4% de resultados falsos negativos para DMG (Tabela 10).

=== Summary ===

```

Correctly Classified Instances      923          92.5777 %
Incorrectly Classified Instances    74           7.4223 %
Kappa statistic                    0.8212
Mean absolute error                0.1097
Root mean squared error            0.2342
Relative absolute error            26.1681 %
Root relative squared error        51.1582 %
Coverage of cases (0.95 level)    99.5988 %
Mean rel. region size (0.95 level) 66.5998 %
Total Number of Instances          997

```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,954	0,141	0,941	0,954	0,947	0,821	0,978	0,991	1
	0,859	0,046	0,889	0,859	0,874	0,821	0,978	0,945	2
Weighted Avg.	0,926	0,112	0,925	0,926	0,925	0,821	0,978	0,977	

=== Confusion Matrix ===

```

a  b  <-- classified as
667 32 | a = 1
42 256 | b = 2

```

FIGURA 7. PORCENTAGEM DA CLASSIFICAÇÃO E MATRIZ DE CONFUSÃO.

Matriz de confusão: primeira linha, 667 amostras foram classificadas como CTRL (a=1), e os 32 dados restantes foram classificados como DMG (b=2); segunda linha, 42 amostras classificadas como CTRL (a=1), e 256 amostras como DMG (b=2). As letras a e b indicam as classes.

TABELA 10. COMPARAÇÃO E ANÁLISE COM AS VARIÁVEIS SELECIONADAS COM ALGORITMO SIMPLE LOGISTIC

Etapas	Referência Marcadores Glicêmicos	Teste – <i>simple logistic</i> Marcadores não-glicêmicos
1	Total = 997 = 100%	Total = 997 = 100%
	Controles 699 = 70,1%%	Controles = 667 = 66,9%
	DMG = 298 = 29,9%	DMG = 256 = 25,7%
2	Comparação entre teste e referencia	
	DMG 25,7% - 29,9% = -4,2% (Falso negativo)	

Referência, classificação da amostra pela glicêmica de jejum, como descrito na Figura 3. Teste, resultados obtidos com o algoritmo *simple logistic* Tabela 9 e Figura 7.

Etapa 1, análise pelo *simple logistic*.

Etapa 2. Diferença observada entre a referência e obtido pelo *simple logistic*.

Com os resultados das análises anteriores, em nova etapa do projeto, as variáveis de melhor discriminação, selecionadas pelo *k-means* (Tabela 7) foram comparadas com as variáveis selecionadas pelo algoritmo *simple logistic* (Tabela 9), como mostrado na Tabela 11.

As variáveis que foram selecionadas por ambos os algoritmos estudados (*k-means* e *simple logistic*) foram utilizadas (marcadores selecionados) para o estudo com rede neural. Como racional, buscamos nesta abordagem, uma estratégia de combinar algoritmo que produz falso-positivo (*k-means*) com outro que promove resultados falso-negativos (*simple logistic*), na tentativa de uma harmonização e minimização destes efeitos indesejados.

TABELA 11. INTERSECÇÃO DAS VARIÁVEIS SELECIONADAS COM OS ALGORITMOS K-MEANS E SIMPLE LOGISTIC E FORMAÇÃO DE MARCADORES SELECIONADOS COMBINADOS.

Clusterização (<i>k-means</i>)	Classificação (<i>simple logistic</i>)	Marcadores Selecionados
IDADE P IMC PAS PAD URIC TG N_HDL Log(TG/HDL-C)	IDADE P A IMC PAS PAD URIC TG PT ALB HDL-C N_HDL LDL/HDL Log(TG/HDL-C)	IDADE P IMC PAS PAD URIC TG N_HDL Log(TG/HDL-C)

Em negrito estão as variáveis que foram comuns aos dois processos de classificação (*k-means* e *simple logistic*) e que foram selecionados como marcadores finais para as análises seguintes com a rede neural MLP.

6.3 RESULTADOS DA ANÁLISE ESTATÍSTICA

Com o grupo de marcadores selecionados estabelecido, o teste t de Student para amostras independentes foi realizado (Tabela 12). Todos os marcadores foram estatisticamente diferentes entre os grupos ($P < 0.001$).

TABELA 12. COMPARAÇÃO DOS MARCADORES NÃO-GLICÊMICOS SELECIONADOS ENTRE OS GRUPOS CONTROLE E DMG.

Variáveis	Grupo CTRL	Grupo DMG	P
Idade (anos)	24.8 ± 6.3	31.9 ± 6.1	<0.001
Peso (kg)	64.4 ± 12.2	84.1 ± 17.5	<0.001
IMC (kg/m ²)	24.8 ± 4.3	32.9 ± 6.4	<0.001
PAS (mmHg)	106.4 ± 11.9	118.1 ± 12.7	<0.001
PAD (mmHg)	66.0 ± 8.5	74.0 ± 10,2	<0.001
Ácido úrico (mg/dL)	3.6 ± 0.8	4.5 ± 1.04	<0.001
Triglicérides (mg/dL)	116.6 ± 53.7	231.6 ± 80.1	<0.001
Não HDL (mg/dL)	139.0 ± 47.1	171.4 ± 45.9	<0.001
Log (TG/HDL-C) (mmol/L)	-0.04 ± 0.2	0.24 ± 0.2	<0.001

Dados apresentados como média ± 1-DP (desvio padrão).

CTRL: grupo controle; DMG: grupo *diabetes mellitus* gestacional. IMC: índice de massa corporal; PAS: pressão arterial sistólica; PAD: pressão arterial diastólica; Não HDL: obtido pela equação (Colesterol total - HDL-C); Log(TG/HDL-C): calculado através do logaritmo da razão dos TG/HDL-C, as concentrações em mmol/L.

Valor de probabilidade (P), teste t para amostras independentes.

P < 0,05 foi considerado significativo e em negrito.

6.4 UTILIZAÇÃO DA REDE NEURAL ARTIFICIAL - MULTILAYER PERCEPTRON (MLP)

Para identificar *overfitting* e otimizar os tempos de treinamento (épocas), foi realizado o experimento mostrado na figura 8. O tempo de treinamento de melhor estabilidade da rede, sem *overfitting* foi de 800 épocas.

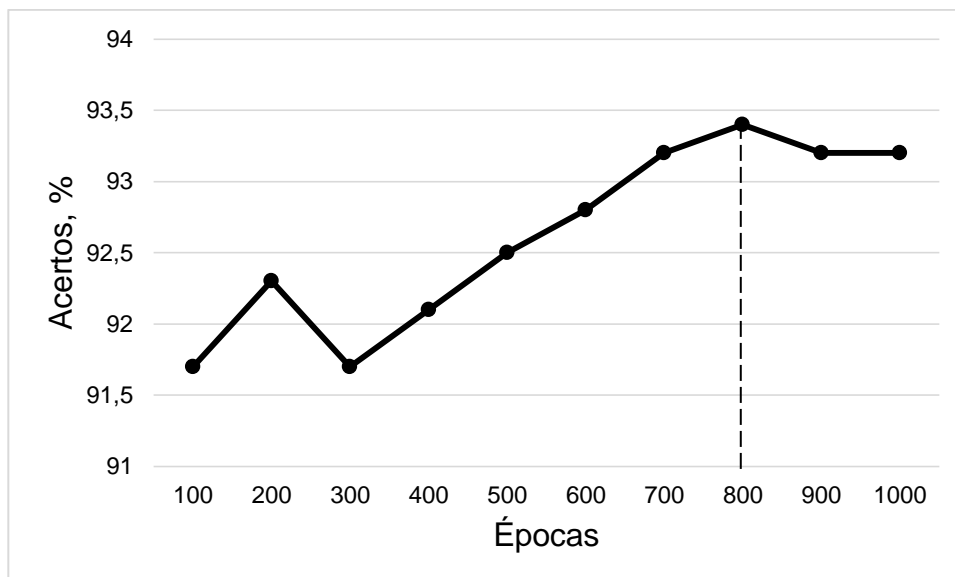


FIGURA 8. ÉPOCAS VERSO ACERTOS COM A REDE MLP.

Tempo de treinamento escolhido de 800 épocas representa a maior porcentagem de acertos (93,4%) sem *overfitting*.

A rede neural MLP (Figura 9) foi utilizada com o grupo de marcadores selecionados, com as 800 épocas e um total de 87,4% de instâncias classificadas corretamente como DMG foi identificada (Figura 10).

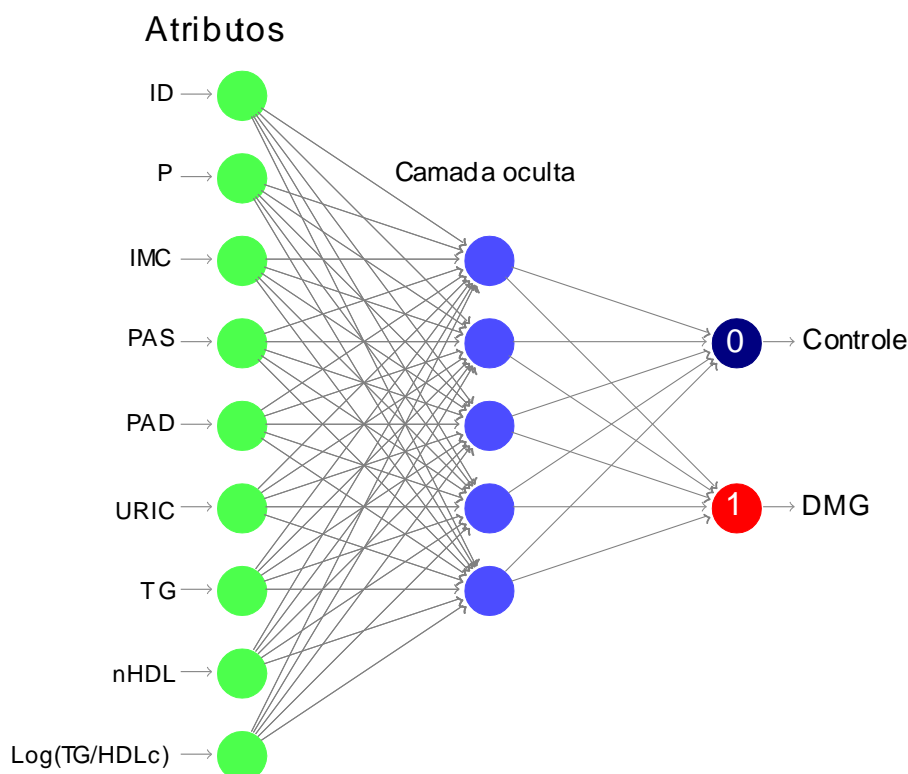


FIGURA 9. ESTRUTURA E VARIÁVEIS DA REDE NEURAL MLP EM ESTUDO.

A figura apresenta a entrada dos atributos; cinco camadas ocultas e a saída (0=controle e 1=DMG) da rede neural MLP de acordo com as configurações próprias do *toolbox* WEKA.

```

Correctly Classified Instances      876      87.8636 %
Incorrectly Classified Instances   121      12.1364 %
Kappa statistic                    0.7124
Mean absolute error                0.1323
Root mean squared error            0.3081
Relative absolute error            31.5421 %
Root relative squared error        67.3074 %
Coverage of cases (0.95 level)    95.5868 %
Mean rel. region size (0.95 level) 61.0331 %
Total Number of Instances         997

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,908    0,191    0,918     0,908   0,913     0,712    0,938    0,969     1
      0,809    0,092    0,790     0,809   0,799     0,712    0,938    0,873     2
Weighted Avg.   0,879    0,161    0,880     0,879   0,879     0,712    0,938    0,941

=== Confusion Matrix ===

  a  b  <-- classified as
635 64 |  a = 1
 57 241 | b = 2

```

FIGURA 10. CLASSIFICAÇÃO DOS GRUPOS COM O USO DA REDE NEURAL MLP.

Resultado da rede neural MLP, como apresentado pelo programa WEKA, empregando as variáveis finais selecionadas. 87,9% das amostras foram classificadas corretamente, e 12,1% classificadas incorretamente. Na matriz de confusão, 635 dados foram classificados

corretamente como grupo CTRL e 241 amostras foram classificadas corretamente como DMG. Nenhuma entrada de índice remissivo foi encontrada.

6.5 AMOSTRAS PARA TREINAMENTO E TESTE E APLICAÇÃO DA REDE NEURAL MLP

Da amostra em estudo, contendo 997 gestantes (CTRL=699 e DMG=298), foram construídos dois grupos (arquivos), sem repetições e de escolha aleatória, um para treinamento (CTRL=599 e DMG=198) e outro para teste (CTRL=100 e DMG=100).

Ambos os grupos (arquivos), treinamento e teste, foram processados com as variáveis selecionadas (Tabela 11, marcadores selecionados), aplicando 800 épocas, na rede MLP, sem modificações na estrutura da rede (*default*) apresentada no programa WEKA.

Os resultados e análises estão mostrados na Tabela 13 e Tabela 14.

TABELA 13. RESULTADOS PARA O GRUPO DE MARCADORES SELECIONADOS COM A REDE NEURAL MLP.

<i>MLP</i>		Resultados	
Treinamento	Matriz de Confusão	a 553 47	b 46 151
	Classificação	88,3% correta 11,7% incorreta	
Teste (<i>supplied test set</i>)	Matriz de Confusão	A 93 18	b 7 82
	Classificação	87,5% correta 12,5% incorreta	

Resultado do treinamento da rede neural com as 797 amostras (599 – CTRL e 198 – DMG). No treinamento, a matriz de confusão explana a quantidade de acertos e erros que o programa obteve. Dos 599 casos de CTRL, foi possível identificar 553 amostras corretamente, errando 46 amostras; dos 198 casos de DMG, foi possível identificar 151 amostras corretamente, errando 47 amostras. Para o conjunto de teste, dos 100 casos de CTRL, foram identificadas 93 amostras corretamente como CTRL, errando 7; e dos 100 casos de DMG, foram identificadas 82 amostras corretamente como DMG, errando 18 amostras.

TABELA 14. ANÁLISE DOS PARÂMETROS DE QUALIDADE DO TESTE E O INTERVALO DE CONFIANÇA.

Parâmetros de qualidade do teste		IC 95% Inferior-Superior
Sensibilidade	92,1%	84,64 – 96,14
Especificidade	83,8%	75,82 – 89,49
Acurácia	87,5%	
Eficiência	88%	

Parâmetros de qualidade obtidos com o software livre **EPR-Val Test Pack 2** (www.hutchon.net/EPRval.htm). IC – intervalo de confiança obtido calculadora de código aberto **OpenEpi** (<http://www.openepi.com/DiagnosticTest/DiagnosticTest.htm>).

Como mostrado na Tabela 13, no conjunto de Teste, o programa conseguiu classificar 87,5% dos dados corretamente, diferenciando gestantes controle e DMG (93 amostras como grupo 1 e 82 amostras como grupo 2).

Os indicadores de performance da rede (Tabela 14) mostraram sensibilidade de 92,1%, especificidade de 83,8%, acurácia de 87,5% e eficiência de 88%. O software livre **EPR-Val Test Pack 2** foi utilizado para obtenção destes parâmetros (www.hutchon.net/EPRval.htm).

O intervalo de confiança ficou entre 84,64% e 96,14% para o parâmetro de sensibilidade, e para especificidade foram observados limite inferior de 75,82% e superior de 89,49%. Os resultados foram obtidos da calculadora de código aberto **OpenEpi** (<http://www.openepi.com/DiagnosticTest/DiagnosticTest.htm>).

A rede neural MLP, na configuração de uso, permitiu discriminar gestantes saudáveis daquelas com DMG, com sensibilidade, especificidade, acurácia e eficiência acima de 80% para todas as variáveis. Estes resultados foram considerados adequados para um teste de triagem para o diabetes gestacional.

Portanto, a proposta em tela, apresentada como um estudo prospectivo, tem potencial para novos estudos com maior tamanho amostral e características de aplicabilidade na discriminação do DMG.

7. CONCLUSÕES

- O uso combinado dos algoritmos *k-means* e *simple logistic* disponíveis no software WEKA, permitiu identificar variáveis não-glicêmicas associadas ao DMG de forma eficaz.
- As variáveis não-glicêmicas que foram selecionadas para o estudo de rede neural pela capacidade de discriminação do DMG combinando as características dos algoritmos *k-means* e *simple logistic* foram: idade, peso, índice de massa corporal, pressão arterial sistólica, pressão arterial diastólica, ácido úrico, triglicerídeos, não-HDL e índice aterogênico [$\log(TG/HDL-C)$].
- A rede neural artificial MLP (*multilayer perceptron*) utilizada na configuração padrão do WEKA, adequadamente treinada e testada, com 800 épocas, permitiu discriminar gestantes com DMG de gestantes saudáveis, utilizando o conjunto de marcadores não-glicêmicos selecionados conjuntamente pelos algoritmos *k-means* e *simple logistic*, com sensibilidade de 92,1%, especificidade de 83,8%, acurácia de 87,5% e eficiência de 88%, para a predição da doença.
- As ferramentas disponíveis no software WEKA (*open source*) tem potencial para serem utilizadas em pesquisas que buscam identificar parâmetros associados, reconhecimento de padrões, e discriminação de processos patológicos.

PERSPECTIVAS FUTURAS

O *toolbox*, WEKA, tem a capacidade de discriminar gestantes saudáveis de gestantes com diabetes *mellitus* gestacional, se realizado através dos passos propostos. O programa é uma ferramenta de fácil uso, e considerado como um bom classificador automático de “segunda opinião”.

Os marcadores descritos neste estudo, e a proposta de classificação do DMG apresentada, tem potencial para aplicação no Sistema Único de Saúde, por serem de fácil obtenção, e por apresentarem um custo reduzido, associado a uma acurácia superior a 87% na discriminação da patologia. No entanto estudos com maior tamanho amostral são necessários para comprovar a eficiência da metodologia proposta bem como para avaliar a consistência do estudo em tela, estruturado como um estudo prospectivo.

Recomendamos testes com novas variáveis, com grupos balanceados (igual número de controles e doentes), bem como a possibilidade de testar outros tipos de redes neurais.

Está em planejamento, para futuro próximo, o desenvolvimento de um software para análise automática do diabetes *mellitus* gestacional, com base em redes neurais, amigável e de fácil alimentação com dados clínicos e laboratoriais, para fornecer um diagnóstico eletrônico ou “segunda opinião”, para esta relevante patologia.

REFERÊNCIAS BIBLIOGRÁFICAS

ADA - AMERICAN DIABETES ASSOCIATION. A. Diagnosis and classification of diabetes mellitus. *Diabetes Care*, v. 34 Suppl 1, p. S62–69, 2011.

ADA - AMERICAN DIABETES ASSOCIATION. A. Diagnosis and classification of diabetes mellitus. *Diabetes Care*, v. 35 Suppl 1, p. S64–71, 2012.

ADA - AMERICAN DIABETES ASSOCIATION. A. Standards of medical care in diabetes. **Diabetes Care**, v. 38 Suppl 1, p. S8–16, 2015.

ALELYANI, S.; TANG, J.; LIU, H. **Feature Selection for Clustering: A Review**. *Data Clustering: Algorithms and Applications*, 2013: 29-60.

ALEPPO, G.; WALKER, K. A. **Diabetes durante a gravidez**. Disponível em: <<http://www.diabete.com.br/diabetes-gestacional-sintomas-e-fatores-de-risco/>>. Acesso em: 26/03/2015.

ALMEIDA, L. M. **Redes Neurais Artificiais**. Disponível em:<<http://www.cin.ufpe.br/~lma3/sih2010/aula-02-sih-10-rna.pdf>>. Acesso em: 02/03/2015.

BARCA, M. C. S.; SILVEIRA, T. R. S. S.; MAGINI, M. **Treinamento de Redes Neurais Artificiais: o Algoritmo Backpropagation**. Disponível em: <http://www.inicepg.univap.br/cd/INIC_2005/inic/IC1%20anais/IC1-17.pdf>. Acesso em: 06/08/2014.

BUCHANAN, T. A.; XIANG, A. H. Gestational diabetes mellitus. *The Journal of clinical investigation*, v. 115, n. 3, p. 485–91, 2005.

BUCHANAN, T. A. et al. What is gestational diabetes? *Diabetes Care*, v. 30 Suppl 2, p. S105–11, 2007.

CASTRO, F. C. C.; CASTRO, M. C. F. **Multilayer Perceptrons**. Disponível em: <http://www.feng.pucrs.br/~decastro/pdf/RNA_C4.pdf>. Acesso em: 06/08/2014.

CHHABRA, N. **Insulin Biosynthesis, Secretion and Action**. Disponível em: <<http://www.namrata.co/insulin-biosynthesis-secretion-and-action/>>. Acesso em: 16/03/2014.

CROWTHER, C. A. et al. Effect of treatment of gestational diabetes mellitus on pregnancy outcomes. *N Engl J Med*, v. 352, n. 24, p. 2477–86, 2005.

DESISTO, C.L.; KIM, S.Y.; SHARMA, A.J. Prevalence Estimates of Gestational Diabetes Mellitus in the United States, Pregnancy Risk Assessment Monitoring System (PRAMS), 2007–2010. *Prev Chronic Dis*, v. 11, n.19, p. 130415, 2014.

DEVORE, J. L. **Probabilidade e Estatística**: para Engenharia e Ciências. São Paulo: Pioneira Thomson Learning, 2006.

DRUCKER, D. J. The role of gut hormones in glucose homeostasis. *J Clin Invest*, v.117, p.24-32, 2007.

FUKS, A. G. Insulinoterapia no Diabetes mellitus tipo 2: quando e como iniciar. *Int Clin Med.*, v. 1, n. A2, 2008.

GALTIER-DEREURE, F.; BOEGNER, C.; BRINGER, J. Obesity and pregnancy: complications and cost. *Am J Clin Nutr*, v. 71, n.5 Suppl, p. 1242S-1248S. 2000.

GANDHI, K.K.; PRAJAPATI, N.B. Diabetes prediction using feature selection and classification. *International Journal of Advance Engineering and Research Development*, v. 3, n. 5, p. 1-7, 2014.

IDF – International Diabetes Federation. Disponível em:<<http://www.idf.org/>>. Acesso em: 24/06/2014.

KHAN, K. S. et al. WHO analysis of causes of maternal death: a systematic review. *Lancet*, v. 367, n. 9516, p. 1066–74, 2006.

KIM, C.; NEWTON, K. M.; KNOPP, R. H. Gestational diabetes and the incidence of type 2 diabetes: a systematic review. *Diabetes Care*, v. 25, n. 10, p. 1862–8, 2002.

KWAK, S. H.; JANG, H. C.; PARK, K. S. Finding genetic risk factors of gestational diabetes. *Genomics Inform*, v. 10, n. 4, p. 239–43, 2012.

LAKSHMI, K.V.; PADMAVATHAMMA, M. Modeling an Expert System for Diagnosis of Gestational Diabetes Mellitus Based On Risk Factors. *IOSR Journal of Computer Engineering (IOSRJCE)*, v. 8, n.3, p. 29-32, 2013.

LANDON, M. B. et al. A multicenter, randomized trial of treatment for mild gestational diabetes. *N Engl J Med*, v. 361, n. 14, p. 1339–48, 2009.

MAGED, A.M. et al. Comparative study between different biomarkers for early prediction of gestational diabetes mellitus. *J Matern Fetal Neonatal Med*, v. 27, n.11, p. 1108-1112, 2014.

MANDAL, A. **Patofisiologia Gestacional do Diabetes**. Disponível em:<[http://www.news-medical.net/health/Gestational-Diabetes-Pathophysiology-\(Portuguese\).aspx](http://www.news-medical.net/health/Gestational-Diabetes-Pathophysiology-(Portuguese).aspx)>. Acesso em: 26/03/2015.

MCGROWDER, D. et al. Lipid profile and clinical characteristics of women with gestational diabetes mellitus and preeclampsia. *JMB*, v. 28, p. 72-81. 2009.

METZGER, B. E. et al. Summary and recommendations of the Fifth International Workshop-Conference on Gestational Diabetes Mellitus. **Diabetes Care**, v. 30 Suppl 2, p. S251–60, 2007.

Ministério da Saúde. Plano de Reorganização da Atenção à Hipertensão arterial e ao Diabetes *mellitus*. MS, 2001.

NAGARAJAN, S. et al. Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes. *International Journal of Current research and Academic Review*, v.2, n. 10, p.91-98, 2014.

NELSON, David L., COX, Michael M., "Princípios de Bioquímica de Lehninger", 6ª edição, W. H. Artmed, 2014

PEREIRA, LO. *Et al.* Obesidade: hábitos nutricionais, sedentarismo e resistência à insulina. *Arq. Bras. Endocrinol. Metab.*, v.47, 2003.

PINHEIRO, R. B.; NUNES, G. B. **O que são Redes Neurais Artificiais?** Disponível em: <<http://webserver2.tecgraf.puc-rio.br/~mgattass/RedeNeural/redeneural.html#>>. Acesso em: 02/03/2015.

POPESCU, M. C.; *et al.* **Multilayer Perceptron and Neural Networks.** Disponível em; <<http://www.wseas.us/e-library/transactions/circuits/2009/29-485.pdf>>. Acesso em: 06/08/2014.

RÔAS, YAS; REIS, EJB. Causas y consecuencias de um estilo de vida sedentário y posibilidades de transformar el conocimiento de hábitos saludables em acciones prácticas y concretas. *EFDeportes*, n.168, 2012.

SACKS, D. A. Gestational diabetes—whom do we treat? *N Engl J Med*, v. 361, n. 14, p. 1396–8, 2009.

SACKS DA; *et al.* HAPO Study Cooperative Research Group. Frequency of gestational diabetes mellitus at collaborating centers based on IADPSG consensus panel-recommended criteria: the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study. *Diabetes Care*. 2012 Mar; 35(3): 526-8.

SANTOS-WEISS, I.C.R.; REA, R.R.; FADEL-PICHETH, C.M.T.; REGO, F.G.M.; PEDROSA, F.O.; GILLERY, P.; SOUZA, E.M.; PICHETH, G. The plasma logarithm of the triglyceride/HDL-cholesterol ratio is a predictor of low risk gestational diabetes in early pregnancy. *Clin. Chim. Acta*, v. 418, p. 1-4, 2013.

SBC – Diretrizes da Sociedade Brasileira de Cardiologia. **I Diretriz sobre aspectos específicos de diabetes mellitus (tipo 2) relacionados á cardiologia.** Rio de Janeiro: SBC, 2009-2014.

SBD – Diretrizes da Sociedade Brasileira do Diabetes. **Diabetes mellitus gestacional: diagnóstico, tratamento e acompanhamento pós-gestação.** Rio de Janeiro: Gen, 2013-2014.

SHAAT, N.; GROOP, L. Genetics of gestational diabetes mellitus. *Curr Med Chem*, v. 14, n. 5, p. 569–83, 2007.

SCHIRMER J et al. *Assistência pré-natal: manual técnico*. Brasília: Secretaria de Políticas de Saúde-SP/Ministério da Saúde, 2000; 66p.

SCHMIDT MI, *et al.* Brazilian Gestational Diabetes Study Group. Gestational diabetes mellitus diagnosed with a 2-h 75-g oral glucose tolerance test and adverse pregnancy outcomes. *Diabetes Care*. 2001 Jul;24(7):1151-5.

SIMMONS, R. Developmental origins of adult metabolic disease: concepts and controversies. *Trends Endocrinol Metab*, v. 16, n. 8, p. 390–4, 2005.

SRIDHAR, S.B. et al. Pregravid liver enzyme levels and risk of gestational diabetes mellitus during a subsequent pregnancy. *Diabetes Care*, v. 37, p. 1878-1884, 2014.

TAN, M. H.; JOHNS, D.; GLAZER, N. B. Pioglitazone reduces atherogenic index of plasma in patients with type 2 diabetes. *Clin Chem*, v. 50, n.7, p. 1184-1188. 2004.

TEKNOMO, K. ***K-Means Clustering Tutorial***. Disponível em: <[http://sigitwidiyanto.staff.gunadarma.ac.id/Downloads/files/38034/M8-Note-kMeans .pdf](http://sigitwidiyanto.staff.gunadarma.ac.id/Downloads/files/38034/M8-Note-kMeans.pdf)>. Acesso em: 06/08/2014.

WHO – World Health Organization. **Diabetes**. Disponível em: <<http://www.who.int/mediacentre/factsheets/fs312/en/>>. Acesso em: 09/12/2014.

WITTEN, I. H.; FRANK, E.; HALL, MARK. A. **Data Mining – Practical Machine Learning Tools and Techniques**. 3 ed. United States: Elsevier, 2011.

YAMANOUCHI, T. *et al.* Plasma 1,5-anhydro-D-glucitol as new clinical marker of glycemic control in NIDDM patients. *Diabetes*, v. 38, n.6, p. 723-729. 1989.

ANEXO

ANEXO 1 – DADOS DA AMOSTRA.....67

ANEXO 1 – DADOS DA AMOSTRA

Os dados utilizados neste estudo foram coletados pela Dra. Izabella Castilhos Ribeiro do Santos-Weiss do Programa de Pós-Graduação em Ciências-Farmacêuticas da Universidade Federal do Paraná. Os dados são de gestantes com diabetes gestacional atendidas no Hospital de Clínicas da Universidade Federal do Paraná e de gestantes saudáveis atendidas no Laboratório Municipal de Curitiba. Este estudo tem a aprovação do Comitê de Ética em Pesquisa da Universidade Federal do Paraná sob os registros CEP/SD: 927.052.10.05 e CAAE: 1924.0.000.091-10.

As gestantes incluídas nas análises apresentam idade entre 13 e 49 anos e foram classificadas como controle (CTRL) (1) e como diabéticas (DMG) (2). O tamanho amostral foi de 1006 gestantes, sendo 699 classificadas como gestantes controle (CTRL) e 307 classificadas como gestantes diabéticas (DMG). Abaixo estão descritos os critérios empregados para a caracterização da amostra analisada de acordo com as recomendações da Associação Americana de Diabetes (2010) e do Ministério de Saúde do Brasil (2001).

- Gestantes CTRL: Glicemia de jejum inferior a 85mg/dl.
- Gestantes com DMG: Glicemia de jejum superior a 92mg/dl com posterior confirmação através do teste oral de tolerância à glucose com 75g (TOTG) e glicemia 2 horas pós sobrecarga superior a 140mg/dl.

Foram obtidas 28 variáveis no total (Tabela 15). O número da amostra e a classe não se enquadram como variáveis.

TABELA 15. TOTAL DE VARIÁVEIS OBTIDAS INICIALMENTE.

1.	Número da Amostra	16.	Triglicerídeos (TG)
2.	Idade (ID)	17.	Proteínas Totais (PT)
3.	Peso (P)	18.	Albumina (ALB)
4.	Altura (A)	19.	HDL-C
5.	Índice de Massa Corporal (IMC)	20.	LDL-C
6.	Semanas de Gestação (SEM)	21.	Log (TG/HDL-C) mg/dL
7.	Grupamento de Semanas (GRUPO)	22.	Não HDL (N_HDL)
8.	Pressão Arterial Sistólica (PAS)	23.	COL/HDL
9.	Pressão Arterial Diastólica (PAD)	24.	LDL/HDL
10.	Glucose (GLU)	25.	Frutosamina (FRUTO)
11.	Hemoglobina Glicada (HbA1C)	26.	FRUTO/PT
12.	Creatinina (CREA)	27.	FRUTO/ALB
13.	Ureia (UREIA)	28.	Log (TG/HDL-C) mmol/L
14.	Ácido Úrico (URIC)	29.	1.5 Anidroglicitol
15.	Colesterol (COL)	30.	Classe (1 ou 2)

A caracterização da amostra e as concentrações séricas dos variáveis laboratoriais, são descritas nas tabelas 16 e 17 a seguir.

TABELA 16. VARIÁVEIS ANTROPOMÉTRICAS

Variáveis	Controle (n=699)	DMG (n=307)	P
Idade (anos)	26 (21-30)	32 (28-36)	<0,001
Peso (kg)	63,3 (56,5-70,0)	82,0 (70,5-94,3)	<0,001
Altura (m)	1,62 (1,57-1,65)	1,60 (1,55-1,65)	0,002
IMC (kg/m ²)	24,0 (21,80-27,33)	32,0 (28,04-36,93)	<0,001
PAS (mmHg)	110 (100-110)	120 (110-120)	<0,001
PAD (mmHg)	60 (60-70)	70 (70-80)	<0,001
Hipertensão arterial* (%)	1,2	32	<0,001*

Valores são mediana (intervalo interquartil) ou frequência (%). P, Mann-Whitney U test ou * Teste do Chi-Quadrado.

Fonte: SANTOS-WEISS, 2013.

TABELA 17. VARIÁVEIS LABORATORIAIS

Variáveis	Controle (n=699)	DMG (n=307)	P
Glicemia de jejum (mg/dL)	82 (78-87)	93 (92-106)	<0,001
2-h 75g de glicose (mg/dL)	ND	162 (148-180)	ND
HbA1C (%)	ND	5,6 (5,3-6,1)	ND
1,5 anidroglicitol	15,3±8,3	9,6±4,9	<0,001*
Colesterol (mg/dL)	188 (160-224)	224 (196-260)	<0,001
HDL-C (mg/dL)	55 (47-65)	55,5 (47-64)	0,596
LDL-C (mg/dL)	109,4 (90,2-134,6)	125,8 (99-152,2)	<0,001
Triglicérides (mg/dL)	107 (83-139)	219 (178-271)	<0,001*
Log (TG/HDL-C)	0,30±0,20	0,60±0,19	<0,001*
Colesterol não-HDL-C (mg/dL)	130,5 (109-162)	169,0 (142-201)	<0,001
Proteína Total (g/dL)	6,9 (6,5-7,4)	6,4 (6,0-6,7)	<0,001
Albumina (g/dL)	4,0 (3,7-4,5)	3,4 (3,1-3,6)	<0,001
Creatinina (mg/dL)	0,8 (0,7-0,8)	0,7 (0,6-0,7)	<0,001
Ureia (mg/dL)	21 (17-25)	16 (13-19)	<0,001
Ácido úrico (mg/dL)	3,6 (3,1-4,1)	4,4 (3,8-5,2)	<0,001

Valores são mediana e intervalo interquartil ou média \pm 1-Desvio Padrão, P, Mann-Whitney U test ou *Student t test.

Fonte: SANTOS-WEISS, 2013