UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA BIOMÉDICA

Trabalho de Conclusão de Curso

ANÁLISE DE DADOS EPIDEMIOLÓGICOS E CLÍNICOS EM PACIENTES COM DIABETES MELLITUS TIPO 2

NATÁLIA EMÍLIA PEREIRA DE FREITAS

Natal/RN

2019

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA BIOMÉDICA

ANÁLISE DE DADOS EPIDEMIOLÓGICOS E CLÍNICOS EM PACIENTES COM DIABETES MELLITUS TIPO 2

Trabalho de conclusão de curso apresentado ao Departamento de Engenharia Biomédica da Universidade Federal do Rio Grande do Norte para obtenção do título de Graduada em Engenharia Biomédica.

Graduanda: Natália Emília Pereira de Freitas.

Orientadora: Prof. Dra. Beatriz Stransky Ferreira.

Natal/RN

2019

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA BIOMÉDICA

ANÁLISE DE DADOS EPIDEMIOLÓGICOS E CLÍNICOS EM PACIENTES COM DIABETES MELLITUS TIPO 2

Banca examinadora do trabalho de conclusão de curso:

Prof. Dra. Beatriz Stransky Ferreira _____

UFRN - Orientadora

Prof. Dr. Tetsu Sakamoto _____

UFRN - Avaliador Interno

Prof. Me. Patrick Cesar Alves Terrematte _____

UFERSA - Avaliador Externo

Natal/RN

2019

AGRADECIMENTOS

Primeiramente, agradeço a Deus, por todas as bênçãos, graças e conquistas da minha vida. Por ser meu pai do céu, meu guia e protetor em todos os momentos, por me iluminar e me dar forças para seguir em frente com meus sonhos.

Aos meus pais, que me ensinaram desde criança que a educação é o melhor caminho para mudar o mundo e a vida das pessoas para melhor. Que me educação, me deram amor, atenção, carinho, me apoiaram e estiveram ao meu lado em todos os momentos da minha vida. Me deram forças, me incentivaram, me ajudaram durante toda minha trajetória e foram os primeiros a acreditarem em mim e nos meus sonhos.

À toda minha família, meus tios, primos, padrinhos e amigos. Por todo apoio e incentivo. Em especial a minha tia avó Tête, por todo carinho, incentivo, preocupação e por estar sempre ao meu lado, acreditando nos meus sonhos e sempre acompanhando cada passo e etapa da minha vida.

À meu namorado Leonardo, por todas as vezes que me ajudou em disciplinas e trabalhos do curso, por ter me apoiado e me incentivado em todos os momentos. Por todo carinho, amor e compreensão. Por acreditar em mim e nos meus sonhos.

À minha professora orientadora Beatriz Stransky, por ter me orientado, me ensinado conteúdos novos que foram fundamentais para o desenvolvimento deste trabalho, por toda disponibilidade, dedicação e paciência. Por todo conhecimento compartilhado e toda ajuda em todas as etapas deste trabalho. Por ter acreditado em mim, e me apresentado o mundo da análise estatística de dados.

Aos meus amigos da reitoria, por serem sempre gentis, compreensivos, responsáveis, éticos e ótimas companhia de trabalho. Por terem me apresentado o mundo da propriedade intelectual e me ensinaram a cada dia um pouco mais sobre o assunto. Por me mostrarem novas possibilidades e por todo apoio e incentivo durante meu curso.

À todos os professores que fizeram parte da minha vida, por cada conhecimento compartilhado, incentivo e exemplo. Por todos os incentivos e por terem me ajudado a trilhar todo caminho da educação.

SUMÁRIO

1. INTRODUÇÃO……...……...……..11

2. OBJETIVO……...……..……..13

2.1. Objetivos Específicos……..……..13

3. REFERENCIAL TEÓRICO……..……14

3.1. Diabetes Mellitus……..……14

3.2. Epidemiologia da Diabetes……...……...….15

3.3. Fisiopatologia……..….....18

3.3.1. Fatores Genéticos……..……..…..…..21

3.4. Diagnóstico e Tratamento……..……...24

3.5. Estatística em Saúde……...……...26

3.5.1 Estatística e Epidemiologia…….....27

4. METODOLOGIA…….…….…...29

4.1. Banco de dados……...…...……...…….....29

4.2. Ambiente e Pacotes para Análise de dados…….…….....…...…....29

4.3. Testes Estatísticos…….…….…...……..30

4.4. Desenvolvimento do Modelo de Regressão Logístico Multivariável…….….....….31

5. RESULTADOS E DISCUSSÃO…….....……..33

5.1. Análise Exploratória dos Dados…….……..…....33

5.2. Correlação e teste de hipóteses……..…...…….…….....……..40

5.3. Modelo de Regressão Logístico……...…....45

6. CONCLUSÃO…….……..…...……....51

7. BIBLIOGRAFIA…….…….…...……....52

LISTA DE FIGURAS

Figura 1: Estimativa de diabetes no mundo……...…….…..…..15
Figura 2: Incidência de diabetes no Brasil (2006-2016)…….…...…….....16
Figura 3: Incidência de excesso de peso de acordo com a faixa etária e anos de estudo……....17
Figura 4: Incidência de obesidade entre brasileiros (2006-2016)……....18
Figura 5: Fisiopatologia da Diabetes Mellitus…….……...……...…….....19
Figura 6: Fisiopatologia da insulina e do glucagon……..……...…….....20
Figura 7: Fatores de risco do diabetes……...……..23
Figura 8: Critérios laboratoriais para o diagnóstico de diabetes……...……...24
Figura 9: Cuidado, tratamento e prevenção do diabetes……...……...25
Figura 10: Ambiente de trabalho do RStudio……...……..…….30
Figura 11: Forma compactada do arquivo data.frame, após a criação e modificação das variáveis……...…….……...….35
Figura 12: Boxplots das variáveis numéricas...…….……..…...……...38
Figura 13: Gráficos de dispersão das variáveis contínuas……...…...….42
Figura 14: Valores para avaliar as correlações…….……....43
Figura 15: Relação entre as variáveis e a chance de diabetes (log(odds)).....47
Figura 16: Resumo das informações do modelo de regressão logística…….....48
Figura 17: Intervalo de confiança (95%).....49

LISTA DE TABELAS

Tabela 1: Informações e unidades de cada variável……..……...34

Tabela 2: Análise das medidas de tendências central das variáveis numéricas….....36

Tabela 3: Frequência das variáveis categóricas…….……..……39

Tabela 4: Tabela de contingência das variáveis categóricas x DM……..40

Tabela 5: Teste X^2 entre as variáveis categóricas e a variável DM2…….....41

Tabela 6: Correlação de Spearman para as variáveis contínuas……..…44

Tabela 7: Teste T entre as variáveis numéricas e a variável categórica DM2...……..45

LISTA DE ABREVIATURAS

DATASUS - Departamento de Informática do Sistema Único de Saúde

DCNT - Doenças Crônicas não Transmissíveis

DM - Diabetes Mellitus

DM1 - Diabetes Mellitus tipo 1

DM2 - Diabetes Mellitus tipo 2

EDA - Análise exploratória dos dados

IDE - Ambiente de desenvolvimento integrado

IMC - Índice de massa corporal

OMS - Organização Mundial da Saúde

ONU - Organização Mundial da Saúde

RI - Resistência à Insulina

SBD - Sociedade Brasileira de Diabetes

SIH - Sistema de Informações Hospitalares

SIM - Sistema de Informações sobre Mortalidade

VIGITEL - Sistema de Vigilância de Fatores de Risco

FREITAS, Natália Emília Pereira. **Análise de dados epidemiológicos e clínicos em pacientes com Diabetes Mellitus tipo 2.** Trabalho de Conclusão de Curso, Graduação em Engenharia Biomédica, Universidade Federal do Rio Grande do Norte, 55p., 2019.

RESUMO

O Diabetes Mellitus é uma doença causada tanto pela ausência de secreção de insulina (tipo 1), quanto pela diminuição da sensibilidade dos tecidos-alvo ao efeito metabólico da insulina (tipo 2). O diabetes tipo 2, corresponde a aproximadamente 90% dos casos da doença. De acordo com os dados da Organização Mundial da Saúde (OMS) acredita-se que 1 em cada 11 pessoas no mundo têm diabetes. O diabetes é responsável por várias complicações como a doença cardiovascular, a diálise por insuficiência renal crônica e as cirurgias para amputações dos membros inferiores. Diante disso, o objetivo deste trabalho foi analisar as informações de um banco de dados denominado "Diabetes", utilizando o RStudio e a linguagem de programação R, para fazer uma análise exploratória, buscar correlações entre as variáveis e desenvolver um modelo de regressão para explicar os principais fatores de risco e causas do Diabetes Mellitus. Este trabalho serviu também como estudo de caso para mostrar a importância da integração e análise de dados epidemiológicos, clínicos e genéticos para o entendimento de processos e planejamentos em saúde.

Palavras-chave: Diabetes Mellitus tipo 2, análise de dados, correlação, teste de hipóteses, análise de regressão, regressão logística.

FREITAS, Natália Emília Pereira. **Analysis of epidemiological and clinical data in patients with type 2 Diabetes Mellitus.** Conclusion Work Project, Biomedical Engineering Bachelor Degree, Federal University of Rio Grande do Norte, 55p., 2019.

ABSTRACT

Diabetes Mellitus is a disease caused by both lack of insulin secretion (type 1) and decreased sensitivity of target tissues to the metabolic effect of insulin (type 2). Type 2 diabetes accounts for approximately 90% of cases of the disease. According to data from the World Health Organization (WHO) it is believed that 1 in 11 people worldwide have diabetes. Diabetes is responsible for various complications such as cardiovascular disease, dialysis for chronic renal failure and surgeries for lower limb amputations. Given this, the objective of this work was to analyze the information from a database called "Diabetes", using RStudio and the R programming language, to make an exploratory analysis, look for correlations between variables and develop a regression model to explain the main risk factors and causes of diabetes mellitus. This work also served as a case study to show the importance of integrating and analyzing epidemiological, clinical and genetic data for understanding health processes and planning.

Keywords: Diabetes Mellitus type 2, data analysis, correlation, hypothesis testing, regression analysis, logistic regression.

1. INTRODUÇÃO

O Diabetes Mellitus (DM) é a síndrome do metabolismo defeituoso de carboidratos, lipídios e proteínas, causado tanto pela ausência de secreção de insulina como pela diminuição da sensibilidade dos tecidos à insulina. É uma doença caracterizada pela elevação da glicose no sangue (GUYTON & HALL, 2011). Atualmente é considerado uma das principais doenças crônicas que afetam o homem contemporâneo, acometendo populações de países em todos os estágios de desenvolvimento econômico-social (ORTIZ, 2001). Porém, vale ressaltar que algumas classes sociais e econômicas são mais afetadas que outras.

Existem dois subtipos principais de DM: os tipos 1 e 2. O Diabetes Mellitus tipo 1, tem etiologia autoimune ou idiopática. O Diabetes Mellitus tipo 2 (DM2) é o mais comum, sendo responsável por cerca de 90% de todos os casos (IDF, 2017), e é uma condição resultante da produção inadequada de insulina pelas células beta pancreáticas ou pela incapacidade do organismo em responder totalmente à insulina, definida como resistência à insulina (RI), ou ambas. O DM2 é uma condição heterogênea do ponto de vista etiológico, sendo uma desordem poligênica resultante da interação entre suscetibilidade em genes alvo, estilo de vida e ambiente (GOLDSTEIN & MULLER-WIELAND, 2016).

O DM2 é mais comumente visto em adultos mais velhos, mas é cada vez mais frequente em crianças, adolescentes e adultos jovens devido a fatores predisponentes como a obesidade e comportamentos como a inatividade física e má alimentação (IDF, 2017). O diabetes, de uma forma geral, é uma importante causa de cegueira, insuficiência renal, amputação de membros inferiores dentre outras consequências como dobrar o risco de óbito por causa cardiovascular em comparação a população geral (SELVIN et al., 2004; SIQUEIRA et al., 2007).

Segundo os dados da Sociedade Brasileira de Diabetes (SBD), em 2016, mais de 12 milhões de brasileiros tinham diabetes. Só no Brasil, entre 2006 e 2016, segundo o Ministério da Saúde, houve um aumento de 60% no diagnóstico da doença, e seu custo deve dobrar até 2030. O Brasil ocupa o 4º lugar entre os 10 países com maior número de pessoas com diabetes. De acordo com os dados da Organização Mundial da Saúde (OMS) acredita-se que 1 em cada 11 pessoas no mundo têm diabetes.

A análise de dados consiste em um estudo detalhado de certo objeto ou assunto, ou seja, é uma técnica minuciosa para obtenção de informações sobre algum tópico, a fim de entender e responder um problema. Na área da saúde, a análise estatística de dados permite uma melhor avaliação acerca das incidências, prevalências, causas e efeitos de uma doença em populações, regiões ou etnias. Além disso, permite de avaliar os fatores de risco e quantificar a eficácia de medidas de prevenção e tratamentos. Nos últimos anos, a análise de dados se tornou uma ferramenta importante para o estabelecimento de políticas públicas e como suporte para o desenvolvimento da medicina de precisão.

Desta forma, este trabalho representa o início de uma linha de pesquisa sobre fatores epidemiológicos, clínicos e genéticos ligados a Diabetes Mellitus tipo 2, a partir de dados públicos disponíveis em repositórios de órgãos de saúde governamentais e consórcios ligados a institutos de pesquisa, como por exemplo o que foi utilizado neste trabalho, que foi disponibilizado pelo departamento de bioestatística da Universidade de Vanderbilt, nos Estados Unidos. Com isso pretende-se investigar as correlações entre estas variáveis e o desenvolvimento ou agravamento da doença, tomando como referência, estudos epidemiológicos e/ou clínicos anteriores. A análise de dados tem sido uma ferramenta poderosa para transformar os números dos resultados encontrados em informações, significados, solução de problemas e em um entendimento maior sobre o caso estudado.

2. OBJETIVO

Neste trabalho teve como objetivo a exploração das informações do banco de dados "Diabetes", para avaliação de correlações, fatores de risco e preditivos associados ao desenvolvimento da Diabetes Mellitus.

2.1 Objetivos Específicos

1) Fazer uma análise exploratória do banco de dados Diabetes.
2) Avaliar as correlações entre as variáveis e a presença da doença.
3) Desenvolver um modelo de regressão logístico múltiplo, com o melhor conjunto de variáveis explicativas para fenótipo estudado.
4) Descrever a relação entre as variáveis explicativas (independentes ou preditoras) e a variável dependente (Diabetes Mellitus).

3. REFERENCIAL TEÓRICO

3.1. Diabetes Mellitus

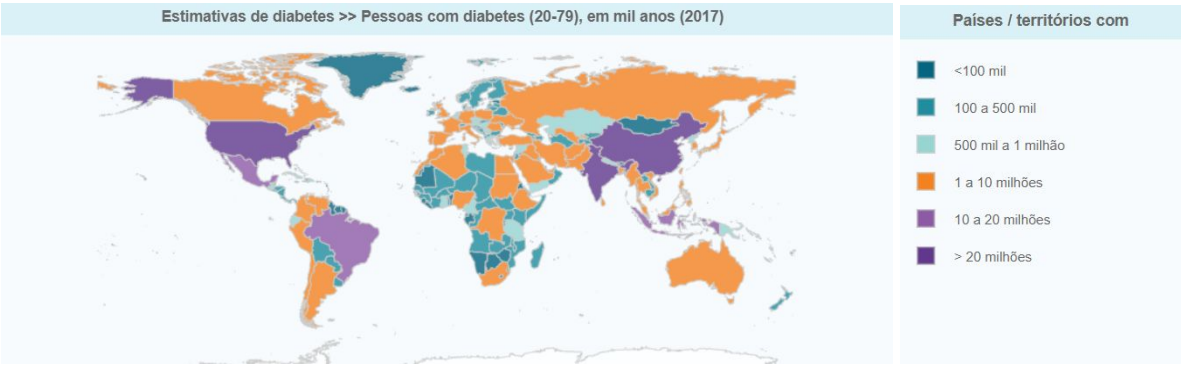
O Diabetes Mellitus (DM) é a síndrome do metabolismo defeituoso de carboidratos, lipídios e proteínas, causado tanto pela ausência de secreção de insulina como pela diminuição da sensibilidade dos tecidos à insulina. É uma doença caracterizada pela elevação da glicose no sangue (GUYTON & HALL, 2011). Quando não tratada corretamente, o diabetes pode resultar em várias consequências e complicações, como por exemplo: doenças renais, doenças cardíacas, retinopatia diabética, neuropatia periférica e disfunção erétil.

Há dois tipos de Diabetes Mellitus, o tipo 1 e o tipo 2. O Diabetes Mellitus tipo 1 corresponde a aproximadamente 10% dos casos de diabetes, e é causado pela ausência de produção de insulina pelas células beta do pâncreas, essa ausência pode ser justificada por diversos fatores, como infecções virais, distúrbios autoimunes ou hereditariedade que podem estar envolvidos na destruição das células beta do pâncreas. Esse tipo de diabetes pode ocorrer em qualquer idade, incluindo na fase adulta, após problemas que levem a destruir as células beta. O diabetes tipo 2, corresponde a aproximadamente 90% dos casos de diabetes e é inicialmente causado pela diminuição da sensibilidade dos tecidos que produzem insulina. Essa redução da sensibilidade à insulina é também chamada de resistência à insulina e prejudica a utilidade e armazenamento dos carboidratos de maneira adequada, elevando a glicose sanguínea e estimulando o aumento de compensação da secreção da insulina. Essa resistência à insulina adquirida por um indivíduo é um processo gradativo, iniciando-se principalmente com o ganho de peso, principalmente abdominal e obesidade, mas também pode ser causada por outros fatores, como doença do ovário policístico, gestação, autoanticorpos ou mutações aos receptores de insulina, excesso de hormônio de crescimento (acromegalia), excesso de glicocorticóides e acúmulo de lipídios no fígado (GUYTON & HALL, 2011).

3.2. Epidemiologia da Diabetes

De acordo com os dados da Organização Mundial da Saúde (OMS) acredita-se que 1 em cada 11 pessoas no mundo têm diabetes. Esse número só cresce. Em 2014, a estatística apontava para 422 milhões de diabéticos, um salto em relação aos 108 milhões de 1980. Segundo os dados da Sociedade Brasileira de Diabetes (SBD), em 2016, mais de 12 milhões de brasileiros tinham diabetes. Só no Brasil, entre 2006 e 2016, segundo o Ministério da Saúde, houve um aumento de 60% no diagnóstico da doença, e seu custo deve dobrar até 2030. O Brasil ocupa o 4º lugar entre os 10 países com maior número de pessoas com diabetes. Em 2017, o número de brasileiros com o diagnóstico de diabetes era 12,5 milhões de pessoas. A prevalência de diabetes no Brasil está entre 8 e 9% da população, na população acima de 65 anos de idade, a prevalência de diabetes é de 19%. Em 2017, foram 4 milhões de mortes por diabetes no mundo (IDF Atlas, 2017). No Brasil, essa enfermidade foi responsável por 5,3% dos óbitos ocorridos em 2011, com taxa de mortalidade de 33,7 óbitos a cada 100 mil habitantes, apesar da redução de 1,7% ao ano verificada no período 2000-2011 (ISER, 2015). Na figura 1 podemos observar a estimativa de diabetes no mundo, segundo os dados do IDF Atlas de 2017.

Figura 1: Estimativa de diabetes no mundo

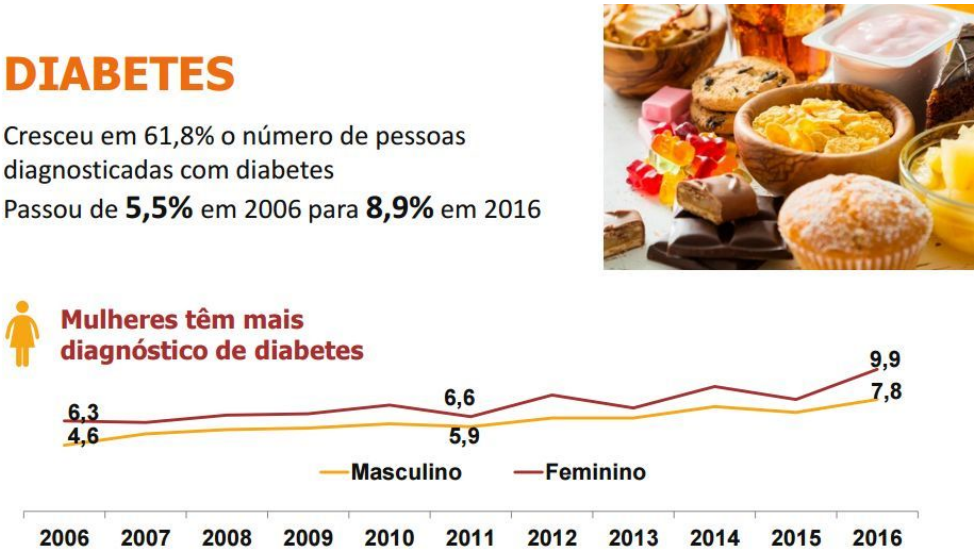


Fonte: IDF (2017)

De acordo com a Pesquisa Vigitel 2017, sistema de Vigilância de Fatores de Risco para doenças crônicas não transmissíveis (DCNT) do Ministério da Saúde, 7,6% da população das capitais brasileiras são portadores de diabetes. Ainda de acordo com o levantamento, o indicador de diabetes aumenta com a idade, principalmente entre idosos com

mais de 65 anos (24%) e é maior entre os com menor escolaridade, que frequentam a escola por até oito anos (14,8%). Já entre as capitais, a frequência do diagnóstico médico de diabetes variou entre 4,5% em Palmas e 8,8% no Rio de Janeiro. Entre 2010 e 2016, o diabetes já vitimou com óbitos 406.452 pessoas no Brasil. De acordo com o Sistema de Informações sobre Mortalidade (SIM), o número cresceu 11,8% no período, saindo de 54.877 mortes para 61.398 no ano de 2016. A figura 2 mostra o aumento de casos de diabetes entre 2006 e 2016, de acordo com os dados da pesquisa da Vigitel 2016. Dados do Sistema de Informações Hospitalares (SIH) apontam que a quantidade de internações teve queda de 8,7%: foram 148.384 em 2010 e 135.364, em 2016. O diabetes é responsável por complicações, como a doença cardiovascular, a diálise por insuficiência renal crônica e as cirurgias para amputações dos membros inferiores (MINISTÉRIO DA SAÚDE, 2018).

Figura 2: Incidência de diabetes no Brasil (2006-2016)

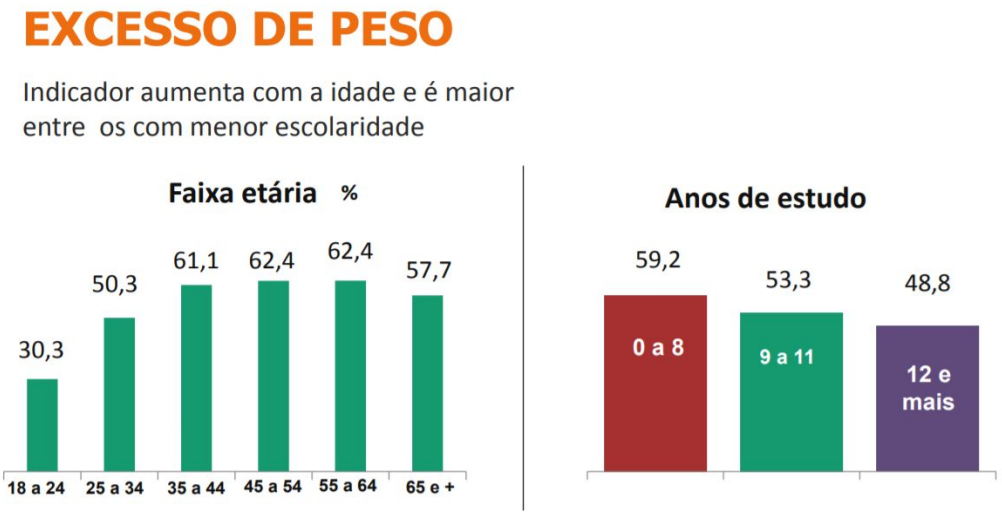


Fonte: Vigitel (2016)

A obesidade está frequentemente associada com a Síndrome Metabólica e é um importante fator de risco de evolução para o Diabetes Mellitus tipo 2 (DM2) clínico. A presença de sobrepeso e obesidade exerce uma influência considerável na elevada morbidade e mortalidade da doença decorrente principalmente da associação com a doença cardiovascular, que é a principal causa de mortalidade em pacientes com DM2 (GOMES, 2006).

De acordo com a pesquisa da Vigitel 2016, a obesidade aumenta com o avanço da idade. Mas mesmo entre os mais jovens, de 25 a 44 anos, atinge indicador alto: 17%. Além disso, o consumo regular de feijão diminuiu, em 2012 era de 67,5%, enquanto em 2016 foi de 61,3%. Também foi observado que mais da metade da população está com peso acima do recomendado e aproximadamente 18,9% dos brasileiros estão obesos.

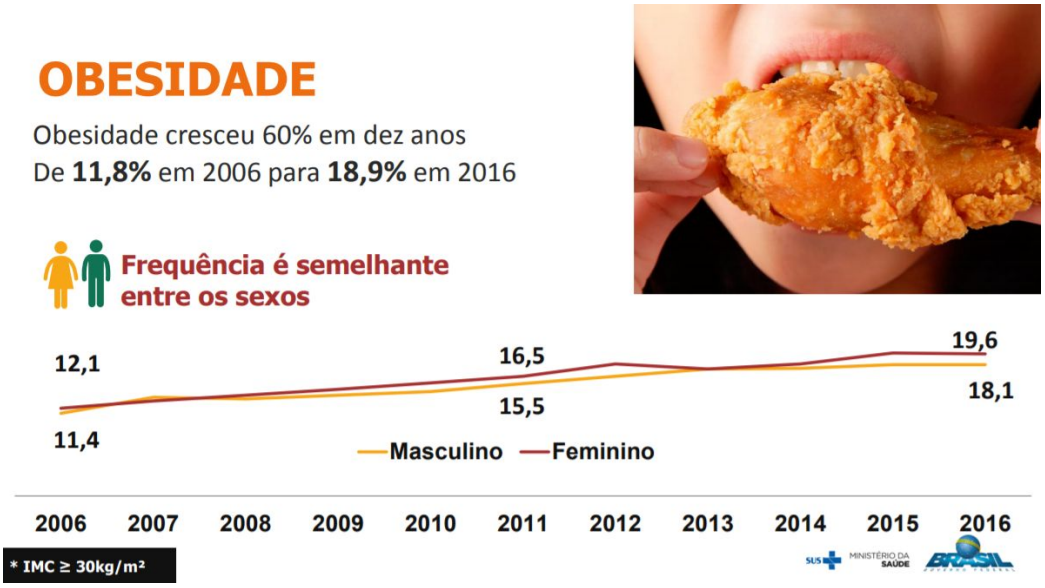
Figura 3: Incidência de excesso de peso de acordo com a faixa etária e anos de estudo



Fonte: Vigitel (2016)

Também foi observado que houve um aumento da prática de exercícios físicos durante o tempo livre. Em 2009, 30,3% da população fazia exercícios por pelo menos 150 minutos por semana, já em 2016 a prevalência foi de 37,6%. Porém, a prevalência diminui com a idade, sendo mais frequente entre os jovens de 18 a 24 anos. Também foi verificado e pode ser observado na figura 3, que a prevalência de obesidade duplica a partir dos 25 anos e que ela é maior entre os que têm menor escolaridade. Além disso, como mostrado na figura 4, a obesidade cresceu 60% em dez anos, passando de 11,8% em 2006 para 18,9% em 2016.

Figura 4: Incidência de obesidade entre brasileiros (2006-2016)



Fonte: Vigitel (2016)

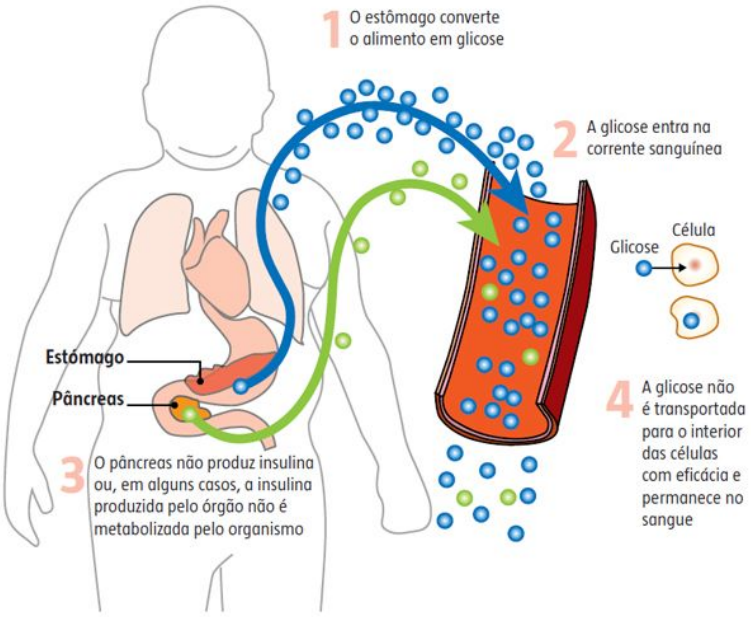
3.3. Fisiopatologia

A insulina é um hormônio secretado pelo pâncreas, com importante função no metabolismo dos carboidratos no sangue. Sua principal função é promover a entrada de glicose para as células do organismo de forma que ela possa ser aproveitada para as diversas atividades celulares. A figura 5 mostra resumidamente o que ocorre no organismo de pessoas com diabetes, ou seja, como a falta da insulina ou um defeito na sua ação resulta em acúmulo de glicose no sangue.

Existem várias condições que podem levar ao desenvolvimento do Diabetes Mellitus, mas a grande maioria dos casos está dividida em dois tipos gerais de diabetes, o tipo 1 e o tipo 2. Conforme mencionado anteriormente, o diabetes tipo 1, também chamado de diabetes mellitus dependente de insulina, é causado pela ausência de secreção da insulina. Essa forma de diabetes têm origem predominantemente genética e hereditária, sendo na grande maioria das vezes, resultado da destruição das células beta pancreáticas por um processo imunológico, ou seja, pela formação de anticorpos pelo próprio organismo contra as células beta, levando a deficiência de insulina. O diabetes tipo 2, também chamado de Diabetes Mellitus não dependente de insulina, é inicialmente causado pela diminuição da sensibilidade dos tecidos-alvo ao efeito metabólico da insulina. Essa sensibilidade reduzida à insulina é

frequentemente chamada de resistência insulínica. Fatores hereditários e genéticos são determinantes para o desenvolvimento do Diabetes Mellitus tipo 2, mas esse tipo de diabetes também é profundamente influenciado por hábitos de vida, como alimentação rica em açúcares e gorduras, além do sedentarismo (GUYTON & HALL, 2011).

Figura 5: Fisiopatologia da Diabetes Mellitus



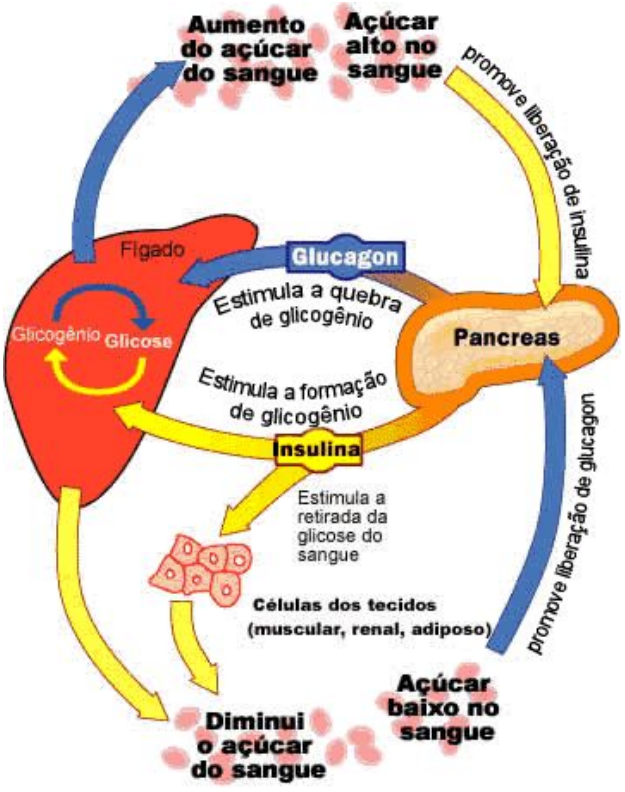
Fonte: Sociedade Brasileira de Diabetes (2014)

O pâncreas, além de suas funções digestivas, secreta dois hormônios importantes, insulina e glucagon. Esses hormônios são cruciais para a regulação normal do metabolismo da glicose, dos lipídios e das proteínas. O pâncreas é formado por dois tipos principais de tecidos, os ácinos e as ilhotas de Langherans. O primeiro secreta o suco digestivo no duodeno e o segundo secreta insulina e glucagon diretamente no sangue.

A insulina desempenha papel importante no armazenamento do excesso de energia. No caso de excesso de carboidratos, a insulina faz com que sejam armazenados sob a forma de glicogênio, principalmente no fígado e nos músculos. Além disso, todo o excesso de carboidratos que não pode ser armazenado sob a forma de glicogênio é convertido sob o estímulo da insulina, em gordura e armazenado no tecido adiposo. Para começar a exercer seus efeitos sobre as células-alvo, a insulina, se liga e ativa o receptor protéico da membrana. É o receptor ativado que causa os efeitos subsequentes. A insulina dirige a maquinaria

metabólica intracelular, de modo a produzir os efeitos desejados sobre o metabolismo de carboidratos, lipídios e proteínas. Um dos efeitos mais importantes da insulina é o de fazer com que a maioria da glicose absorvida após refeições, seja armazenada quase imediatamente no fígado sob a forma de glicogênio. Então, quando a concentração de glicose sanguínea começa a cair, a secreção de insulina diminui rapidamente e o glicogênio hepático é de novo convertido em glicose, que é liberada de volta ao sangue, para impedir que a concentração de glicose caia a níveis muito baixos. Quando a quantidade de glicose que penetra as células hepáticas é maior que a quantidade que pode ser armazenada sob a forma de glicogênio ou da que pode ser utilizada para o metabolismo local dos hepatócitos, a insulina promove a conversão de todo esse excesso de glicose em ácidos graxos. A insulina também inibe a gliconeogênese, que ocorre, em sua maior parte, por meio da redução das quantidades e atividades que as enzimas hepáticas precisam para a gliconeogênese (GUYTON & HALL, 2011).

Figura 6: Fisiopatologia da insulina e do glucagon



Fonte: Só Biologia (2019)

O glucagon, que é o hormônio secretado pelas células alfa das ilhotas de Langerhans quando a concentração da glicose sanguínea cai, tem diversas funções que são

diametralmente opostas às da insulina. A mais importante dessas funções é a de aumentar a concentração da glicose sanguínea, efeito que é exatamente oposto ao da insulina. Os principais efeitos do glucagon sobre o metabolismo da glicose são: A quebra do glicogênio hepático e o aumento da gliconeogênese no fígado (GUYTON & HALL, 2011). Esses dois efeitos aumentam enormemente a disponibilidade da glicose para os outros órgãos do organismo. Apenas uns poucos microgramas de glucagon podem fazer com que o nível de glicose sanguínea duplique ou aumente ainda mais, dentro de uns poucos minutos. O fígado funciona como importante sistema tampão da glicose sanguínea. A figura 6 mostra como esses dois hormônios atuam regulando a concentração da glicose no sangue. Tanto a insulina como o glucagon funcionam como importantes sistemas de controle por feedback para manter a concentração de glicose sanguínea normal.

3.3.1. Fatores Genéticos

Apesar dos avanços dos estudos genômicos das últimas décadas, a arquitetura genética detalhada do risco de Diabetes Mellitus tipo 2 ainda não foi definida com precisão. Polimorfismos em genes envolvidos em formas monogênicas de diabetes também desempenham um papel no DM2 poligênico. Pensa-se que o próprio DM2 seja um distúrbio poligênico que se desenvolve devido à interação complexa entre múltiplos genes e fatores ambientais. Ainda é pouco conhecido como esses genes interagem entre si e com o meio ambiente para produzir DM2. Ao contrário do DM1, onde o risco genético está principalmente concentrado na região HLA, o componente genético do risco do DM2 não está concentrado em uma região e parece ser o resultado da interação de vários genes espalhados por todo o genoma (ALI, 2013).

Muitos estudos buscam identificar os genes associados ao desenvolvimento de DM2. Existem alguns genes que já foram identificados de forma confiável, como por exemplo, o CAPN10 e o TCF7L2. Esses dois genes estão relacionados ao DM2, o primeiro que é o CAPN10, que codifica uma protease de cisteína que faz parte da família calpaína, uma grande família de genes ubiquamente expressos que desempenham vários papéis no remodelamento intracelular, sinalização pós-receptor e outras funções intracelulares. Atualmente, a função desse gene no metabolismo da glicose permanece desconhecida e sua ligação com o DM2, embora confirmada em várias populações, nem sempre é consistente. O segundo é o TCF7L2,

a associação entre DM2 e vários polimorfismos de nucleotídeo único nesse gene foi fortemente confirmada em vários estudos de associação genômica ampla em diferentes grupos étnicos e esse gene continua sendo o mais replicado e mais fortemente associado gene de risco para o DM2 no momento (ALI, 2013).

Pensa-se que tanto a resistência à insulina quanto a falha das células β resultam da interação complexa de muitas vias diferentes sob o controle combinado de fatores ambientais e genéticos. O papel da genética na Diabetes Mellitus tipo 2 (DM2) é indicado pelo agrupamento familiar de sensibilidade à insulina e secreção de insulina, a maior taxa de concordância da DM2 em gêmeos monozigóticos versus dizigóticos e a alta prevalência de diabetes tipo 2 em certos grupos étnicos. Estima-se que 30% a 70% do risco de DM2 podem ser atribuídos à genética. Os padrões de herança sugerem que o diabetes tipo 2 é poligênico e heterogêneo - isto é, múltiplos genes estão envolvidos e diferentes combinações de genes desempenham um papel em diferentes subconjuntos de indivíduos (DORIA, 2008).

Nas últimas duas décadas, quatro abordagens foram usadas para desvendar a genética do diabetes tipo 2, cada uma com algum sucesso. A primeira abordagem foi focar nas formas de diabetes tipo 2 transmitidas com um padrão dominante de herança mendeliana e / ou outras características clínicas específicas. Isso levou à descoberta de genes envolvidos no diabetes de maturidade dos jovens (MODY); os genes envolvidos em várias síndromes de grave resistência à insulina, os genes envolvidos no diabetes neonatal; e os genes envolvidos no diabetes mitocondrial e outras síndromes genéticas raras. Juntas, essas formas monogênicas do diabetes tipo 2 são responsáveis por menos de 2% a 5% desta doença. A definição desses genes relativamente raros do diabetes, no entanto, forneceu novas idéias sobre as vias envolvidas na regulação da homeostase da glicose (DORIA, 2008).

Por exemplo, elucidar a família de genes MODY destacou uma complexa rede hierárquica de fatores de transcrição que modulam o desenvolvimento e a função das células β (DUNCAN et al, 1998). Da mesma forma, a descoberta de mutações no receptor de insulina Akt2 e PPARγ em síndromes de resistência severa à insulina confirma o importante papel dessas proteínas sinalizadoras in vivo em humanos (O'RAHILLY, 2007). Do ponto de vista terapêutico, a identificação de mutações no canal de potássio sensível ao ATP (Kir6.2 / KCNJ11) em lactentes com diabetes neonatal tem sido muito importante, permitindo que essas crianças sejam tratadas com sulfonilureias orais em vez de injeções de insulina ao longo da vida (GLOYN et al; 2004, PEARSON et al; 2006). Mais interessante, do ponto de vista

mecanicista, tem sido a descoberta de mutações em genes não previamente vinculados ao diabetes, como os genes para a proteína A da membrana nuclear ou a proteína seipina (BSCL2) em certas formas de lipodistrofia e resistência à insulina (SHACKLETON et al. 2000, MAGRE et al., 2001). No entanto, exatamente como as mutações nesses genes resultam em lipodistrofia ou resistência severa à insulina permanece incerto (DORIA, 2008).

O diabetes tipo 1 resulta da destruição autoimune das células b produtoras de insulina das ilhotas pancreáticas. Essa destruição é observada no decorrer de meses ou anos, e os sintomas só aparecem quando a maior parte das ilhotas pancreáticas é destruída. A incidência é maior em crianças de ancestralidade europeia, seguidas de afro-americanas e hispânicas; as taxas são baixas em crianças asiáticas e americanas nativas (HAY et al, 2014).

Figura 7: Fatores de risco do diabetes

Fonte: 121doc (2017)

O diabetes tipo 2 tem um forte componente genético, embora exista uma variação nos defeitos herdados em diferentes famílias. A obesidade, em particular, e a falta de exercício físico constituem causas importantes, porém raramente são suficientes por si só para causar diabetes na juventude. A maioria dos pacientes pediátricos provém de baixo nível socioeconômico e de famílias disfuncionais, e alguns apresentam transtornos psiquiátricos (HAY et al, 2014). A figura 7 apresenta as principais causas de diabetes tipo 1 e tipo 2.

3.4. Diagnóstico e Tratamento

Os métodos usuais para estabelecer o diagnóstico do diabetes se baseiam em diversos testes químicos realizados na urina e no sangue, como por exemplo, o da glicose urinária e o da glicose sanguínea de jejum e níveis de insulina. Os sintomas clássicos de diabetes são: poliúria, polidipsia, polifagia e perda involuntária de peso. Outros sintomas que levantam a suspeita clínica são: fadiga, fraqueza, letargia, prurido cutâneo e vulvar, balanopostite e infecções de repetição. Algumas vezes o diagnóstico é feito a partir de complicações crônicas como neuropatia, retinopatia ou doença cardiovascular aterosclerótica.

Resumidamente, os testes laboratoriais mais comumente utilizados para suspeita de diabetes ou regulação glicêmica alterada são:

• Glicemia de jejum: Nível de glicose sanguínea após um jejum de 8 a 12 horas;

• Teste oral de tolerância à glicose (TTG-75g): O paciente recebe uma carga de 75 g de glicose, em jejum, e a glicemia é medida antes e 120 minutos após a ingestão;

• Glicemia casual: Tomada sem padronização do tempo desde a última refeição.

Pessoas cuja glicemia de jejum situa-se entre 110 e 125 mg/dL (glicemia de jejum alterada), por apresentarem alta probabilidade de ter diabetes, podem requerer avaliação por TTG-75g em 2h. Mesmo quando a glicemia de jejum for normal (< 110 mg/dL), pacientes com alto risco para diabetes ou doença cardiovascular podem merecer avaliação por TTG (MINISTÉRIO DA SAÚDE, 2006).

Os critérios clínicos e laboratoriais para o diagnóstico de diabetes são resumidos na figura 8.

Figura 8: Critérios laboratoriais para o diagnóstico de diabetes

Sintomas de diabetes (poliúria, polidipsia, polifagia ou perda de peso inexplicada)

+ glicemia casual e > 200 mg/dL (realizada a qualquer hora do dia, independentemente do horário das refeições);

= OU =

Glicemia de jejum e > 126 mg/dL*;

= OU =

Glicemia de 2 horas e > 200 mg/dL no teste de tolerância à glicose*.

* Devem ser confirmados com nova glicemia.

Fonte: Ministério da Saúde (2006)

Considerando a elevada carga de morbi-mortalidade associada, a prevenção do diabetes e de suas complicações é hoje prioridade de saúde pública. Na atenção básica, ela pode ser efetuada por meio da prevenção de fatores de risco para diabetes como sedentarismo, obesidade e hábitos alimentares não saudáveis; da identificação e tratamento de indivíduos de alto risco para diabetes (prevenção primária); da identificação de casos não diagnosticados de diabetes para tratamento (prevenção secundária); e intensificação do controle de pacientes já diagnosticados visando prevenir complicações agudas e crônicas (prevenção terciária). O cuidado integral ao paciente com diabetes e sua família é um desafio para a equipe de saúde, especialmente para poder ajudar o paciente a mudar seu modo de viver, o que estará diretamente ligado à vida de seus familiares e amigos. Aos poucos, ele deverá aprender a gerenciar sua vida com diabetes em um processo que vise qualidade de vida e autonomia (MINISTÉRIO DA SAÚDE, 2006).

Figura 9: Cuidado, tratamento e prevenção do diabetes

Fonte: 121doc (2017)

O tratamento efetivo do Diabetes Mellitus tipo 1 requer administração de insulina suficiente, de modo que o paciente possa ter metabolismo de carboidratos, lipídios e proteínas tão normal quanto possível. Para as pessoas com diabetes tipo 2, dieta e exercícios são geralmente recomendados na tentativa de induzir a perda de peso e de reverter a resistência à

insulina. Se o método falhar, pode-se administrar medicamentos para aumentar a sensibilidade à insulina ou para estimular o aumento da produção da insulina pelo pâncreas. Contudo, em muitas pessoas, deve-se empregar insulina exógena para regular a glicose sanguínea. A figura 9 mostra de forma resumida, alguns tratamentos do diabetes e algumas formas de prevenção. Um tratamento adequado em conjunto com uma alimentação saudável e a prática de exercícios físicos, são essenciais tanto para a prevenção do Diabetes Mellitus tipo 2, quanto para evitar possíveis complicações nos pacientes com diabetes.

3.5. Estatística em Saúde

A estatística é um ramo da matemática que lida com a organização, análise, interpretação e apresentação de dados, e seus usos são numerosos, desde prever o monitoramento do clima, verificar a qualidade de produtos na linha de montagem ou julgar se a nova política do governo funcionou ou não. O uso da estatística na área da saúde remonta há pelo menos 150 anos e o fundador da moderna epidemiologia, dr. John Snow, é o ator de um caso clássico de aplicação da estatística na saúde pública.

Em Londres, a cólera já havia eclodido duas vezes, em 1849 e 1954, e havia duas teorias concorrentes sobre a origem da doença. Uma das teorias apontava para o miasma, devido ao ar poluído e a outra para o contágio, onde coisas muito pequenas que se espalhavam entre as pessoas. Em 1855, o médico John Snow, publica sua análise sobre a relação entre os distritos de Londres e o número de mortes por cólera no ensaio "Sobre a Maneira de Transmissão da Cólera". Ele defendia a teoria do contágio e sugeria que os distritos mais atingidos seriam os mais próximos do rio Tamisa, para o qual uma grande quantidade de esgoto era bombeada. William Farr, secretário geral, defendia a teoria do miasma e mostrava associações entre o número de mortes e fatores como elevação acima do nível do mar e o número médio de pessoas. Ele estava convencido de que a elevação acima do nível do mar era o principal fator. Ambos usavam dados e cálculos manuais para tentar apoiar suas teorias. Infelizmente, o status social superior de William Farr ajudou a convencer as autoridades de que o ar, e a água, não era o culpado pelo surto (WALDMAN, 1998). Hoje sabemos que a teoria do miasma está errada e uma análise dos dados de 2004 e vários métodos estatísticos sugerem que é realmente o suprimento de água e não a elevação acima do nível do mar que traria a doença.

Nos anos 2000 a Organização das Nações Unidas estabeleceu os Objetivos de Desenvolvimento do Milênio, que variam desde taxas de pobreza extrema até a interrupção da disseminação de HIV/Aids e o trabalho continua com a Agenda de 2030 para o Desenvolvimento Sustentável. O relatório da ONU de 2019 afirma que aproximadamente 2 bilhões de pessoas sofrem de insegurança alimentar e 820 milhões de pessoas estão desnutridas. Ao mesmo tempo, as taxas de excesso de peso estão crescendo em quase todas as regiões do mundo, com números globais atingindo 2 bilhões de adultos com excesso de peso e 40 milhões de crianças com menos de cinco anos (GSDR, 2019) . Mas como averiguar estes dados? De acordo com o próprio relatório, o que é medido pode ser avaliado e as estatísticas ajudam a estabelecer metas e impulsionar o progresso. A Agenda de 2030 ONU não pretende apenas tornar o mundo melhor e em breve. Eles escolhem coisas que podem medir e estabelecem um cronograma para o monitoramento do progresso de ações ao longo do caminho. Para que isso seja possível, cada região de cada país precisa de dados confiáveis para medir seu progresso em relação às metas, e para obter dados confiáveis é necessário sistemas de informações robustos e analistas competentes. Atualmente, a estatística é parte essencial na tomada de decisões na área da saúde, pela sua aplicação que vão desde decisões políticas em saúde pública à medicina personalizada, passando pelas práticas da medicina baseada em evidências.

3.5.1 Estatística e Epidemiologia

O principal objetivo da epidemiologia é prevenir doenças e promover saúde. Para que isso seja possível, é necessário conhecer as causas das doenças e agravos à saúde e as maneiras pelas quais podem ser modificados. Para determinar quais fatores afetam os resultados de saúde no nível da população, os epidemiologistas empregam vários modelos de estudo diferentes, com diferentes capacidades explicativas/preditivas. Esses projetos são usados para examinar a relação entre exposições (ou determinantes) e resultados de saúde. Um resultado de saúde pode ser uma doença, condição, morte, evento ou uma alteração no estado ou comportamento de saúde. Por exemplo, além de doenças, pode-se estudar eventos de saúde, como lesões ou a ocorrência de um "evento", como número de mortes por armas de fogo.

Um dos pontos de muita controvérsia em epidemiologia é o conceito de causa. O entendimento das causas das doenças e agravos à saúde é importante não apenas para a prevenção, mas também para o correto diagnóstico e tratamento. O processo pelo qual se faz inferência causal – julgamento ligando possíveis causas e seus desfechos – é o principal tema da filosofia geral da ciência, e o conceito de causa tem diferentes significados em diferentes contextos (BONITA et al, 2010). A causa de uma doença ou agravo à saúde é um evento, condição, característica ou uma combinação desses fatores que desempenham um papel importante no desenvolvimento de um desfecho em saúde. Logicamente, uma causa deve preceder o efeito (desfecho). Uma causa é dita como sendo suficiente quando ela inevitavelmente produz ou inicia um desfecho, e é dita necessária se o desfecho não pode acontecer na sua ausência. Algumas doenças são causadas por características genéticas individuais, enquanto outras causas interagem com fatores genéticos para tornar certos indivíduos mais vulneráveis do que outros. O termo causa ambiental é frequentemente usado para distinguir essas outras causas dos fatores genéticos. Geralmente, existe algum componente genético e ambiental em cada mecanismo causal.

A epidemiologia, como qualquer outra ciência, se baseia no método científico. A realização de estudos epidemiológicos requer o uso de amostras para que sejam feitas inferências sobre uma população, e a aplicação de conceitos e ferramentas da bioestatística para a síntese e análise dos dados. Apoiada grandemente em uma base estatística, e a abordagem epidemiológica consiste basicamente em:

i) observação dos fenômenos de saúde e doença na população;
ii) quantificação dos mesmos em frequências e distribuições;
iii) análise das frequências e distribuições de saúde e de seus determinantes;
iv) definição de cursos de ação apropriados.

Esse processo de observar-quantificar-comparar-propor serve também para avaliar a efetividade e o impacto das intervenções em saúde, para construir novos modelos que descrevam e expliquem as observações e para utilizá-los na predição de novos fenômenos (BONITA et al, 2010).

4. METODOLOGIA

4.1. Banco de dados

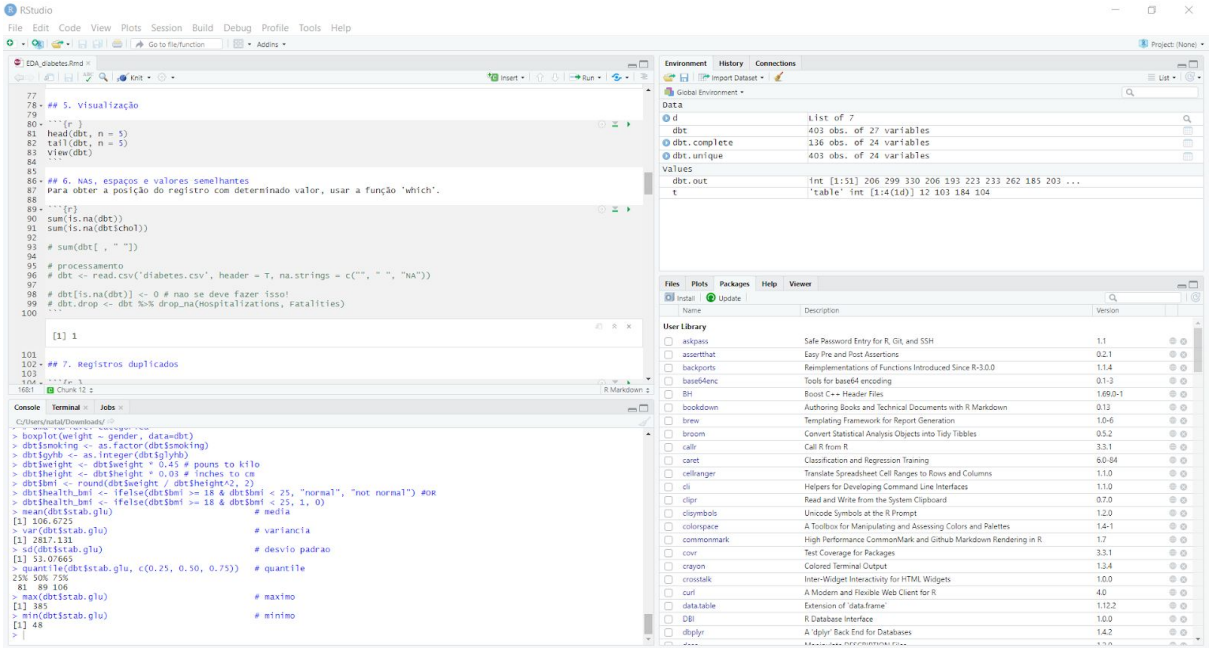
Para este trabalho foi utilizado o banco de dados "Diabetes" disponibilizado pelo departamento de bioestatística da Universidade de Vanderbilt, Inglaterra (http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets), com permissão de "acesso aberto" (open access). O banco de dados consiste em informações sobre 403 indivíduos selecionados, que foram rastreados para diabetes, em uma amostra de 1046 indivíduos. Estas informações consistem de dados pessoais, clínicos e fisiológicos, num total de 24 variáveis. As informações foram coletados em uma pesquisa sobre a prevalência de obesidade, diabetes e outros fatores de risco cardiovascular, desenvolvida na região central da Virgínia para afro-americanos (WILLEMS et al, 1997).

4.2. Ambiente e Pacotes para Análise de dados

O processamento e análise dos dados foi desenvolvida utilizando o ambiente de desenvolvimento integrado (IDE) RStudio e a linguagem de programação R. O Rstudio (https://rstudio.com/) é um software livre que inclui um console, editor de realce de sintaxe que suporta execução direta de código, além de ferramentas para plotagem, histórico, depuração e gerenciamento de espaço de trabalho (Figura 10). O R é uma linguagem de programação, voltada para computação e gráficos estatísticos que vem progressivamente se especializando em manipulação, análise e visualização gráfica de dados e tem sido considerado como uma das melhores linguagens computacionais para essa finalidade (https://www.r-project.org/).

Os pacotes são bibliotecas com dados e funções para diferentes áreas do conhecimento relacionado a estatística e áreas afins. Neste trabalho foram utilizadas os pacotes Tidyverse (https://www.tidyverse.org/), uma coleção de pacotes R projetados para ciência de dados. Todos os pacotes compartilham uma filosofia de design, gramática e estruturas de dados subjacentes. Os scripts para análise de dados foram desenvolvidos localmente.

Figura 10: Ambiente de trabalho do RStudio



Fonte: Autoral

4.3. Testes Estatísticos

Os testes de hipóteses são métodos científicos para a tomada de decisão baseado em dados de uma amostra. Nos permite verificar se os dados amostrais trazem evidência que apoiem ou não uma hipótese estatística formulada. Sendo assim, utilizados para determinar quais resultados de um estudo científico podem levar à rejeição da hipótese nula a um nível de significância pré-estabelecido. O caminho utilizado neste trabalho, foi através do p-valor, por ser um caminho mais rápido para fazer o teste de hipótese. Em vez de se construir uma região crítica, calcula-se diretamente o p-valor. O p-valor é uma estatística muito usada para sintetizar o resultado de um teste de hipóteses.

Para configurar um teste de hipótese, é necessário assumir um valor para o parâmetro de sua população, ou seja, a hipótese nula. Em seguida, usando uma amostra e um teste estatístico apropriado, é determinada a probabilidade de obter o valor para essa amostra se o valor da população for o que você supôs. Para valores de baixa probabilidade (convencionalmente, $p \leq 0{,}05$), você conclui que há evidências de que seu valor populacional

hipotético está incorreto, ou seja, você rejeita a hipótese nula, pois é improvável que você obtenha seus dados se, de fato, sua hipótese nula for verdadeira.

Os testes estatísticos utilizados para verificar a correlação entre variáveis foram Qui-quadrado para as categóricas, e o de Spearman para variáveis contínuas com distribuições desconhecidas ou diferentes da distribuição normal.

4.4. Desenvolvimento do Modelo de Regressão Logístico Multivariável

A regressão logística é um método aplicado aos problemas de classificação e predição, e é amplamente utilizada para os chamados resultados binários, que podem assumir apenas dois valores. Na medicina, os exemplos clássicos são a morte, pois o indivíduo pode estar morto ou vivo, ou se uma pessoa tem ou não uma determinada doença / condição em um determinado momento no tempo.

Para um resultado binário, geralmente estamos mais interessados em modelar a proporção de indivíduos com o resultado de interesse, ou seja, a proporção de indivíduos com diabetes em nosso exemplo. Isso é equivalente à probabilidade de um indivíduo ter diabetes. A regressão logística tem duas grandes vantagens: é fácil de executar em software padrão e você pode examinar e descrever os relacionamentos entre as variáveis com muito mais facilidade do que algumas técnicas de aprendizado de máquina.

Classificamos que os pacientes são chamados para a variável binária e, em seguida, analisamos o restante do conjunto de dados para ver quais características dos pacientes estão mais associadas a ser um do que a ser um zero. Por exemplo, a idade e o sexo de um paciente afetam sua chance de ter diabetes.

Para executar um modelo logístico, é importante verificar algumas suposições: a variável de resultado deve ser categórica, com distribuição binomial e os preditores podem ser categóricos ou contínuos. Se um categórico, não é necessário haver de um número igual de observações em cada categoria, mas categorias com números muito pequenos podem causar problemas. Se contínuas, elas não precisam ser distribuídas normalmente, mas deve-se assumir que a relação entre a variável e o resultado é linear. A partir destas premissas, pode-se desenvolver um modelo de regressão logístico.

Apesar de não haver uma única metodologia para desenvolver um modelo de regressão, as etapas descritas abaixo são boas práticas a serem seguidas para qualquer tipo de regressão:

1. Inspecionar as variáveis usando estatísticas resumidas e histogramas para variáveis contínuas e tabulações para variáveis categóricas. Isso ajuda a identificar e quantificar a quantidade de informações ausentes e quaisquer observações estranhas ou distantes. Se encontrar algum valor estranho, será necessário decidir o que fazer com ele antes de prosseguir.

2. Examinar o relacionamento entre suas variáveis preditoras candidatas usando tabulações cruzadas para variáveis categóricas e correlações em pares e matrizes de gráfico de dispersão para variáveis contínuas. Isso ajuda a identificar associações em potencial nas variáveis preditoras candidatas que poderiam ser problemáticas se incluídas no seu modelo multivariável.

3. Ajustar um modelo de regressão logístico entre o resultado e cada uma das variáveis preditoras candidatas. Isso permite avaliar o relacionamento entre cada uma das variáveis preditoras candidatas e a variável de resultado. Novamente, isso ajuda a detectar erros nos dados ou na codificação e também permite que se preveja o que espera que aconteça ao ajustar o modelo multivariável.

5. RESULTADOS E DISCUSSÃO

5.1. Análise Exploratória dos Dados

A partir da análise exploratória dos dados, foi possível resumir e visualizar as características importantes do conjunto de dados, pois antes de construir modelos estatísticos ou fazer qualquer outra análise, é importante examinar as variáveis e obter informações sobre o conjunto de dados. Dessa forma, a análise exploratória de dados possibilita identificar distribuições, tendências e padrões, identificar erros, variações e anomalias, identificar correlações entre variáveis e sugerir hipóteses a serem testadas.

Para fazer a análise exploratória dos dados, inicialmente foi examinado o arquivo *data.frame*. Esse arquivo possui 403 observações (linhas) e 24 variáveis (colunas). Entre as 24 variáveis, 18 são do tipo inteira, 2 do tipo numérica e 4 do tipo factor (categórica). A tabela 1 mostra o nome dado a cada variável, o significado, a unidade das medidas, o tipo e a quantidade de NA's, que são os dados que não foram avaliados ("*not available*"), ou seja, o banco de dados não possui a informação de uma variável, para um determinado registro.

As variáveis com maior quantidade de NA's foram a bp.2s e a bp.2d, que representam a segunda pressão arterial sistólica e diastólica. Essas variáveis possuem 65% dos seus dados contendo NA's, o que mostra a ausência dessas informações em mais da metade dos dados do arquivo. Todas as outras variáveis possuem uma pequena quantidade, ou até mesmo, a ausência de NA's. As variáveis *glyhb* e dm, são o segundo grupo de variáveis com apenas 3,2% de NA's. Com isso, foi observado que o banco de dados "Diabetes", possui uma quantidade significativa de dados para serem estudados de forma confiável, pois a maioria das variáveis possuem informações completas ou em grande quantidade, para poder realizar este trabalho. Nesta inspeção também verificou-se que não havia nenhum registro duplicado.

Tabela 1: Informações e unidades de cada variável

Informação	Nome	Legenda	Unidade	Tipo	NAs
Dados pessoais	x	Ordem dos dados		int	0
	id	Identificação da pessoa		int	0
	location	Local em que reside		factor	0
	age	Idade	anos	int	0
	gender	Gênero		factor	0
	time.ppn	Tempo pós-prandial	minutos	int	3
	insurance	Seguro		int	0
	smoking	Fumante		int	0
	fh	Histórico familiar de DM		int	0
	dm	Diabetes Mellitus		factor	13
Dados físicos	height	Altura	polegadas	int	5
	weight	Peso	libras	int	1
	frame	Estrutura corporal		int	12
	waist	Cintura	polegadas	int	2
	hip	Quadril	polegadas	int	2
Dados clínicos	chol	Colesterol total		int	1
	stab.glu	Glicose estabilizada		int	0
	hdl	Lipoproteína de alta densidade		int	1
	ratio	Proporção de colesterol / HDL		num	1
	glyhb	Hemoglobina Glicosilada (HbA1C)		num	13
	bp.1s	Primeira pressão arterial sistólica		int	5
	bp.1d	Primeira pressão arterial diastólica		int	5
	bp.2s	Segunda pressão arterial sistólica		int	262
	bp.2d	Segunda pressão arterial diastólica		int	262

Fonte: Autoral

Após a inspeção do data.frame, iniciamos a fase de pré-processamento dos dados, com a modificação e a criação de novas variáveis, para obter novas possibilidades de análise.

A variável smoking, foi transformada em factor e foram criadas três novas variáveis, a *gyhb*, bmi e a *health_bmi*. A primeira é uma medida da porcentagem de hemoglobina glicada em uma variável inteira, a segunda é o cálculo do índice de massa corporal e a terceira avalia se o índice de massa corporal está normal ou não. Com isso, houve um aumento na quantidade de objetos do arquivo *data.frame,* passando de 24 variáveis para 27 variáveis, a figura 11 mostra como ficou os dados de forma compactada.

Figura 11: Forma compactada do arquivo *data.frame*, após a criação e modificação das variáveis

```
> str(dbt)
'data.frame':   403 obs. of  27 variables:
 $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ id        : int  1000 1001 1002 1003 1005 1008 1011 1015 1016 1022 ...
 $ chol      : int  203 165 228 78 249 248 195 227 177 263 ...
 $ stab.glu  : int  82 97 92 93 90 94 92 75 87 89 ...
 $ hdl       : int  56 24 37 12 28 69 41 44 49 40 ...
 $ ratio     : num  3.6 6.9 6.2 6.5 8.9 ...
 $ glyhb     : num  4.31 4.44 4.64 4.63 7.72 ...
 $ location  : Factor w/ 2 levels "Buckingham","Louisa": 1 1 1 1 1 1 1 1 1 1 ...
 $ age       : int  46 29 58 67 64 34 30 37 45 55 ...
 $ gender    : Factor w/ 2 levels "female","male": 1 1 1 2 2 2 2 2 2 1 ...
 $ height    : num  1.86 1.92 1.83 2.01 2.04 2.13 2.07 1.77 2.07 1.89 ...
 $ weight    : num  54.5 98.1 115.2 53.6 82.4 ...
 $ frame     : Factor w/ 4 levels "","large","medium",..: 3 2 2 2 3 2 3 3 2 4 ...
 $ bp.1s     : int  118 112 190 110 138 132 161 NA 160 108 ...
 $ bp.1d     : int  59 68 92 50 80 86 112 NA 80 72 ...
 $ bp.2s     : int  NA NA 185 NA NA NA 161 NA 128 NA ...
 $ bp.2d     : int  NA NA 92 NA NA NA 112 NA 86 NA ...
 $ waist     : int  29 46 49 33 44 36 46 34 34 45 ...
 $ hip       : int  38 48 57 38 41 42 49 39 40 50 ...
 $ time.ppn  : int  720 360 180 480 300 195 720 1020 300 240 ...
 $ insurance : int  1 0 2 1 0 1 2 0 2 2 ...
 $ fh        : int  0 0 0 0 0 0 1 0 1 0 ...
 $ smoking   : Factor w/ 3 levels "1","2","3": 3 2 2 3 3 1 2 2 1 2 ...
 $ dm        : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
 $ gyhb      : int  4 4 4 4 7 4 4 3 4 5 ...
 $ bmi       : num  15.7 26.6 34.4 13.2 19.8 ...
 $ health_bmi: num  0 0 0 0 1 1 1 1 0 0 ...
```

Fonte: Autoral

Após finalizar o pré-processamento, foi iniciada a análise da distribuição das variáveis. Foi feita a análise das medidas de tendências central e variância das variáveis numéricas - média, a variância, o desvio padrão, os quartis, os valores máximo, mínimo, mediana e a distância interquartil. A tabela 2 gerada no RStudio, mostra de forma resumida o resultado para cada uma das variáveis contínuas.

Tabela 2: Análise das medidas de tendências central das variáveis numéricas

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
X	1	101.5	202	202	302.5	403	0
id	1000	4792	15766	15978	20336	41756	0
chol	78	179	204	207.8	230	443	1
stab.glu	48	81	89	106.7	106	385	0
hdl	12	38	46	50.45	59	120	1
ratio	1.5	3.2	4.2	4.522	5.4	19.3	1
glyhb	2.68	4.38	4.84	5.59	5.6	16.11	13
age	19	34	45	46.85	60	92	0
height	1.56	1.89	1.98	1.981	2.07	2.28	5
weight	44.55	67.95	77.62	79.92	90	146.25	1
bp.1s	90	121.2	136	136.9	146.8	250	5
bp.1d	48	75	82	83.32	90	124	5
bp.2s	110	138	149	152.4	161	238	262
bp.2d	60	84	92	92.52	100	124	262
waist	26	33	37	37.9	41	56	2
hip	30	39	42	43.04	46	64	2
time.ppn	5	90	240	341.2	517.5	1560	3
insurance	0	0	1	1.055	2	2	0
fh	0	0	0	0.1886	0	1	0
gyhb	2	4	4	5.092	5	16	13
bmi	10.81	17.16	19.77	20.47	22.93	39.67	6
health_bmi	0	0	1	0.5214	1	1	6

Fonte: Autoral

A distribuição das variáveis contínuas pode ser verificada também através de gráficos univariados, como em um histograma ou boxplot. Para variáveis contínuas que apresentam uma distribuição normal, o conhecimento da média e desvio-padrão é o suficiente para descrever todos os valores da distribuição. Para aquelas que não seguem exatamente esta distribuição é importante saber a média e o intervalo interquartil. Os gráficos de box-plot são

mais informativos quanto a distribuição das variáveis numéricas, uma vez que eles explicitam os limites inferiores e superiores dos valores, além dos 1º, 2º (mediana) e 3º quartis, e *outliers*, valores que ultrapassam o limite inferior (max{min(dados); Q1 - 1,5(Q3-Q1)}) ou superior (min{max(dados); Q3 + 1,5(Q3 - Q1)}) da distribuição .

A distribuição dos valores de algumas variáveis numéricas são mostradas na figura 12 e na tabela 2. Conhecendo os limites estabelecidos como normais em relação às variáveis clínicas (*chol, stab.glu, hdl, glybh*) e índice de massa corporal (*bmi*), conforme estabelecido pelo Ministério da Saúde, podemos verificar como estes dados se comportam em relação aos valores de referência.

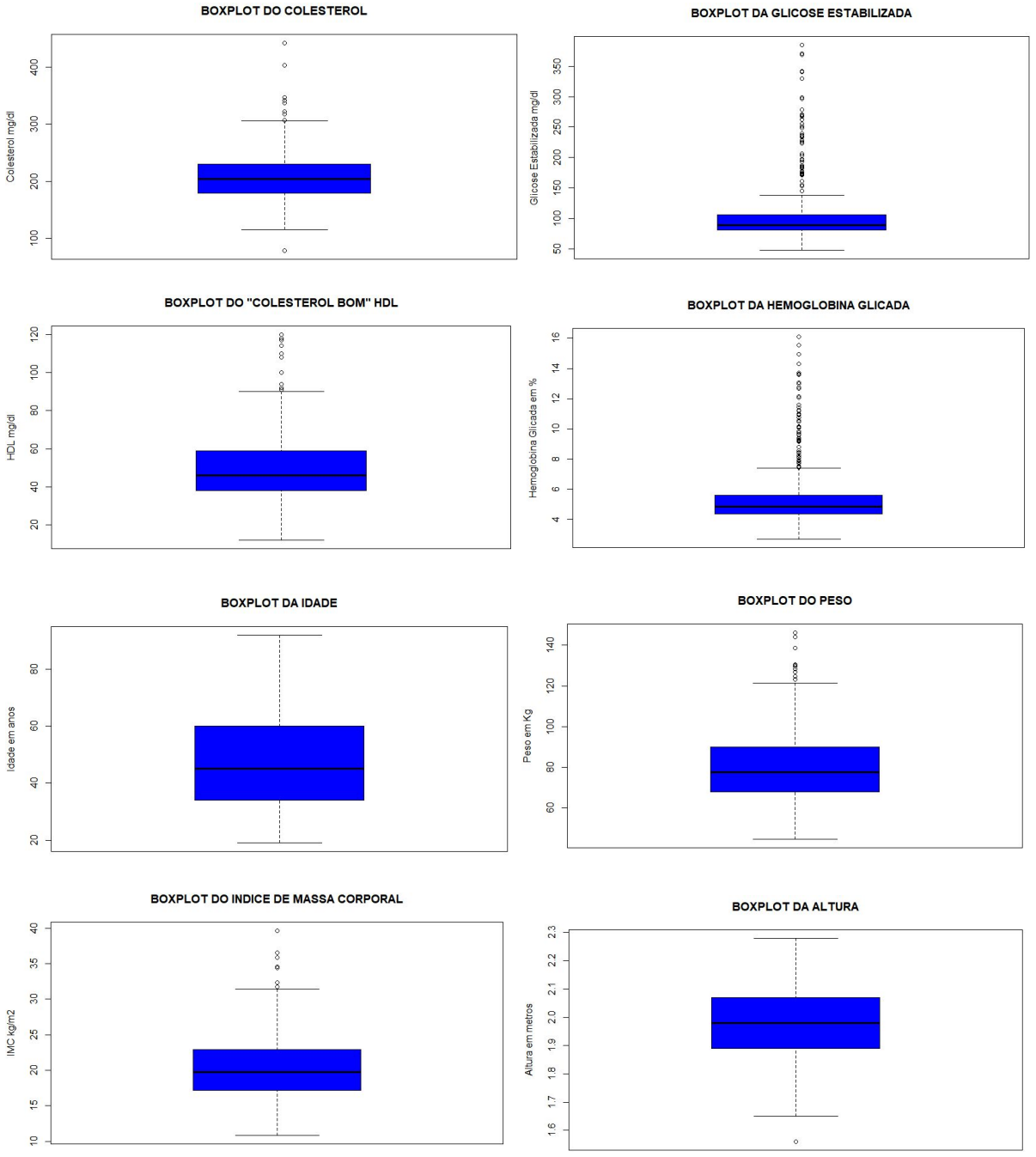
Para a variável *chol,* o centro da distribuição dos dados (mediana) foi de 204 mg/dl e a dispersão (máximo valor, 306 mg/dl - mínimo valor, 115 mg/dl) foi de 115 mg/dl. Estes dados apresentam uma distribuição simétrica em torno da média e 9 *outliers*, que são os valores acima ou abaixo dos máximos e mínimos estabelecidos para essa amostra. O limite superior para que uma taxa de colesterol total possa ser considerada normal é 190 mg/dl, portanto a maioria dos dados da amostra possuem valores acima do considerado normal.

A variável *stab.glu*, apresenta mediana de 89 mg/dl, o mínimo valor de 48 mg/dl e o máximo valor de 138 mg/dl. A distribuição dos dados é positivamente assimétrica, com 51 outliers. Novamente mais de 50% das amostras apresentam valores acima do que é considerado normal para glicose estabilizada (≥ 99 mg/dl).

A variável *hdl,* apresenta mediana de 46 mg/dl, mínimo valor de 12 mg/dl e o máximo valor de 90 mg/dl. A distribuição dos dados é positivamente assimétrico e foram identificados 13 outliers, com valores acima do limite. Considerado que valores maiores que 40 mg/dl são considerados normais, então para esse caso, os outliers não representam uma anomalia para a amostra.

Em relação a variável *glyhb,* a mediana foi de 4,84%, com mínimo valor de 2,68% e máximo de 7,40%. A distribuição dos dados é positivamente assimétrica e observa-se 57 outliers. Considerando os valores de referência, que indicam a presença de diabetes, como iguais ou superiores a 6,5%, mais de 75% da amostra encontra-se dentro do limite de normalidade. Para a variável *bmi*, a mediana foi de 19,77 kg/m2, o mínimo valor foi de 10,81 kg/m2 e o máximo valor foi de 31,41 kg/m2. Esta variável apresenta uma distribuição positivamente assimétrico e 8 outliers, todos com valores do IMC acima do valor normal, que é entre 18,5 kg/m2 e 25,0 kg/m2.

Figura 12: Boxplots das variáveis numéricas



Fonte: Autoral

Para as variáveis categóricas foi feita a análise da frequência das mesmas, como mostrado na tabela 3.

Tabela 3: Frequência das variáveis categóricas

……...……...……...FRAME……...……..……..				
	Large	Medium	Small	NAs
Porcentagem (%)	25,6	45,7	25,8	3,0
……..SMOKING….……..…….……..				
	1 (current)	2 (never)	3 (ex)	NAs
Porcentagem (%)	30,0	55,3	14,6	0
……...……..DM……...……				
	No (0)	Yes (1)	NAs	
Porcentagem (%)	81,9	14,9	3,2	
……..…...GENDER……....…....…....……				
	Female (0)	Male (1)	NAs	
Porcentagem (%)	58,1	41,9	0	
……..…...LOCATION……..……..…..				
	Buckingham	Louisa	NAs	
Porcentagem (%)	49,6	50,4	0	

Fonte: Autoral

Analisando os resultados da tabela 3, podemos observar que a variável *location* tem os dados distribuídos equitativamente, indicando uma quantidade semelhante de indivíduos em cada localização. As outras variáveis porém não se encontram uniformemente distribuídas. Na variável *gender* é possível observar uma predominância de mulheres em relação aos homens. Na variável *dm* foi observada uma grande diferença na distribuição dos valores, pois aproximadamente 80% dos dados são de indivíduos sem diabetes. Na variável *frame* um pouco menos da metade dos indivíduos tinham estatura média, e na variável *smoking*, mais da metade dos dados eram de indivíduos que nunca fumaram.

5.2. Correlação e teste de hipóteses

Correlação é qualquer relação estatística (causal ou não causal) entre duas variáveis. É uma semelhança ou equivalência que existe entre duas hipóteses, situações ou objetos diferentes. Os coeficientes de correlação auxiliam os pesquisadores a mensurar a relação entre variáveis distintas. Os coeficientes de correlação são métodos estatísticos para se medir as relações entre variáveis e o que elas representam.

Para a análise das correlações das variáveis categóricas, foram feitas tabelas cruzadas, também conhecida como tabela de contingência. Essas tabelas são usadas para avaliar o relacionamento das categorias de acordo com a independência ou homogeneidade. A tabela 4 mostra a frequência das variáveis categóricas em relação a variável *dm*.

Tabela 4: Tabela de contingência das variáveis categóricas x DM

DM x bmi_categorized			
bmi_categorised	**Diabetes Mellitus (Não)**	**Diabetes Mellitus (Sim)**	**Total (linhas)**
Normal	100 (92)	9 (16,9)	109
Obeso	118 (124,1)	29 (22,8)	147
Acima do peso	99 (100,5)	20 (18,5)	119
Abaixo do peso	9 (9,3)	2 (1,7)	11
Total (colunas)	326	60	386
DM x gender			
gender	**Diabetes Mellitus (Não)**	**Diabetes Mellitus (Sim)**	**Total (linhas)**
Feminino	194 (192,9)	34 (35,07)	228
Masculino	136 (137,07)	26 (24,9)	162
Total (colunas)	330	60	390
DM x smoking			
smoking	**Diabetes Mellitus (Não)**	**Diabetes Mellitus (Sim)**	**Total (linhas)**
1 (current)	95 (97,3)	20 (17,7)	115
2 (never)	187 (184,5)	31 (33,5)	218
3 (ex)	48 (48,2)	9 (8,8)	57
Total (colunas)	330	60	390

DM x location			
location	**Diabetes Mellitus (Não)**	**Diabetes Mellitus (Sim)**	**Total (linhas)**
Buckingham	159 (160,8)	31 (29,2)	190
Louisa	171 (169,2)	29 (30,8)	200
Total (colunas)	330	60	390
DM x frame			
frame	**Diabetes Mellitus (Não)**	**Diabetes Mellitus (Sim)**	**Total (linhas)**
Largo	76 (83,8)	23 (15,1)	99
Médio	152 (150,7)	26 (27,2)	178
Pequeno	93 (86,4)	9 (15,6)	102
Total (colunas)	321	58	379

* O valor esperado é apresentado entre parênteses

Fonte: Autoral

Para se avaliar a dispersão das variáveis categóricas nominais e avaliar a associação existente entre variáveis qualitativas foi utilizado o teste Qui-Quadrado. O princípio básico deste teste é comparar proporções, ou seja, possíveis divergências entre as frequências observadas e esperadas para um certo evento.

Tabela 5: Teste X^2 entre as variáveis categóricas e a variável DM2

…….....……Teste Qui-Quadrado (X^2) entre as variáveis do *data.frame* x DM2…….……....			
	X^2	G.L. (Graus de Liberdade)	P-valor
Health_bmi	5.5948	1	0.01801
Gender	0.026997	1	0.8695
Smoking	0.58997	2	0.7445
Location	0.12701	1	0.7216
Frame	8.2055	3	0.04195

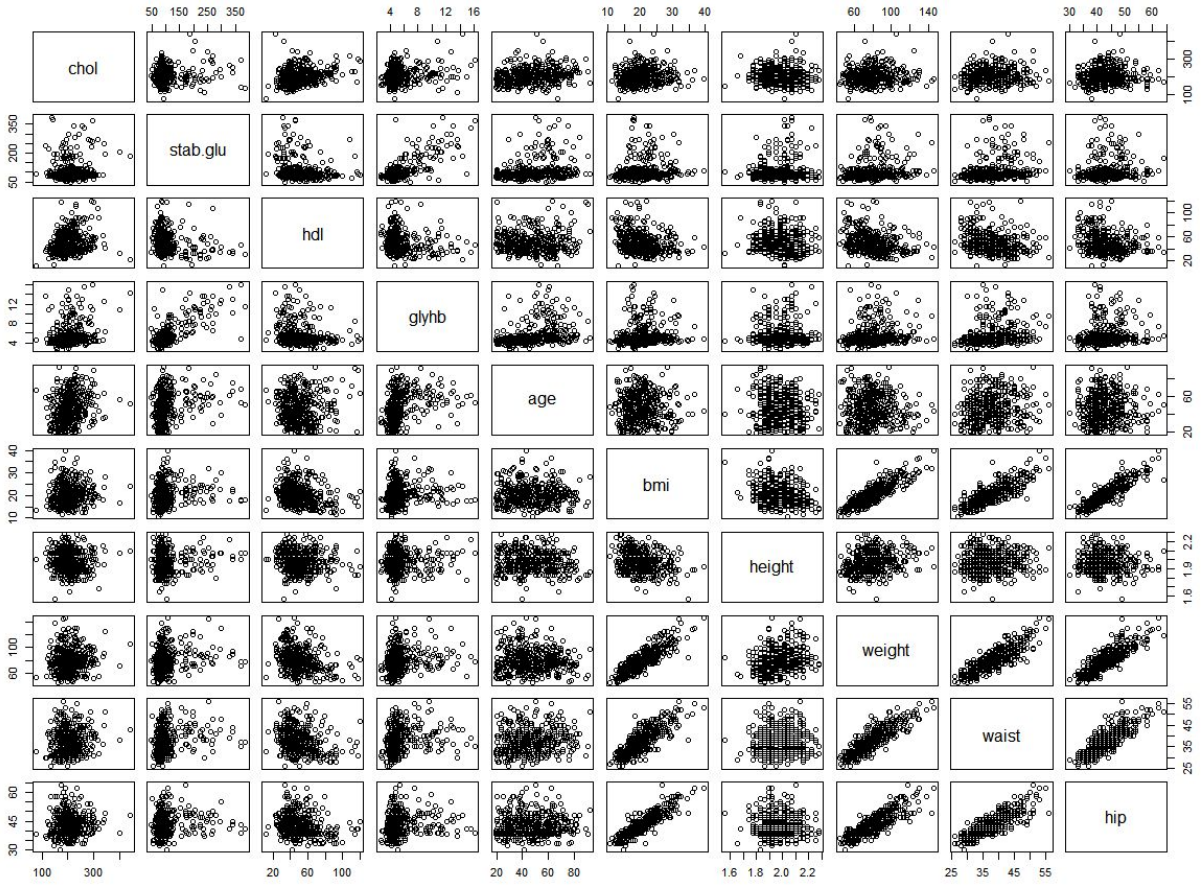
Fonte: Autoral

De acordo com a análise do X^2 da tabela 5, é preciso aceitar a hipótese nula para as variáveis: gender, smoking e location (p-valor ≥ 0.05). Ou seja, não é possível afirmar que a

existência de associação entre essas variáveis não seja devido ao acaso, pois a probabilidade de errar caso a hipótese nula fosse rejeitada seria muito alta. Com isso, a partir das análises feitas por esse teste foi possível afirmar que, para essa amostra de dados, não existe uma relação entre os gêneros e ter Diabetes Mellitus tipo 2, assim como também não existe uma relação entre fumar e ter DM2, nem a localização de cada pessoa com o fato de ter DM2. Entretanto para as variáveis *health_bmi* e frame é possível rejeitar a hipótese nula ($p < 0.05$), ou seja é possível inferir que existe uma relação entre o IMC e ter DM2, assim como também existe uma relação entre a estrutura corporal e o fato de ter DM2, que não é determinada pelo acaso.

Para a verificação das correlações das variáveis contínuas - inteiras, numéricas, ou categóricas com uma distribuição contínua, foram plotados gráficos bivariados de dispersão. Os gráficos de dispersão são utilizados quando se deseja expor o quanto uma variável é afetada por outra, conforme mostrado na figura 13.

Figura 13: Gráficos de dispersão das variáveis contínuas

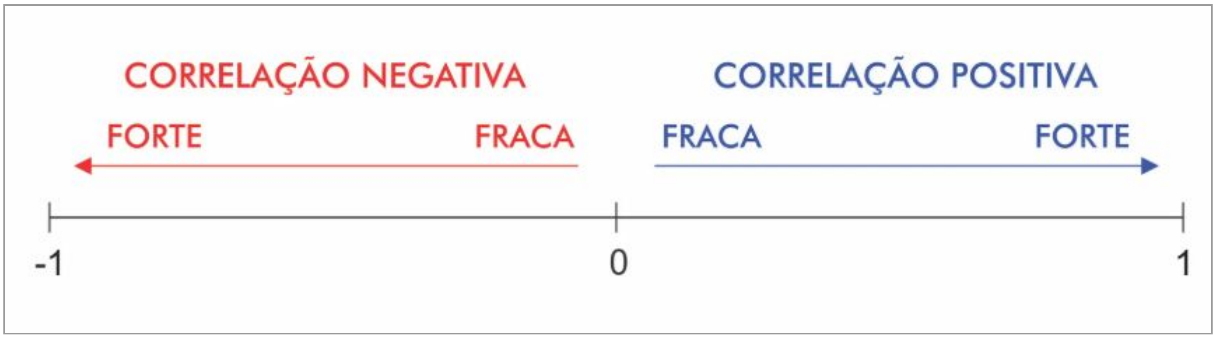


Fonte: Autoral

Observando a figura 13, é possível notar o resultado das dispersões e escolher entre as variáveis que se correlacionam, qual utilizar no modelo de regressão, levando em conta que existem correlações mais fortes e outras mais fracas, assim como também percebemos a falta de correlação entre algumas. A variável bmi, possui correlações positivas fortes. É possível observar esse tipo de correlação entre a variável *bmi* e as variáveis *weight, waist e hip*. Também podemos notar esse tipo de correlação entre as variáveis *weight* e *waist, weight* e *hip*, *hip* e *waist,* e *stab.glu* e *glyhb*. Alternativamente, as correlações negativas estão em menor quantidade e são correlações bem mais fracas. Um exemplo de correlação negativa na figura 13, é a correlação existente entre a variável *hdl* e *weight*. Também foi observada a presença de algumas correlações nulas, ou seja, não há correlação aparente entre as variáveis. Esse exemplo foi visto para as correlações entre as variáveis *stab.glu* e *height, age* e *bmi, height* e *glyhb, hdl* e *age,* entre outras que também apresentaram dados dispersos para essa amostra do banco de dados.

Estas relações podem ser quantificadas pelo método de Spearman e valores entre -1 e +1 são relacionados à classificação de correlação positivas ou negativas e correlação forte, fraca ou nula (figura 14). De forma geral, uma correlação acima de 0,7 ou abaixo de -0,7 é considerada forte.

Figura 14: Valores para avaliar as correlações



Fonte: Oper (2019)

A tabela 6 foi gerada no RStudio e mostra os valores dos coeficientes de correlação entre as variáveis contínuas calculados por este método. Com isso, percebemos que as observações feitas a partir dos gráficos da figura 13, são compatíveis com as informações da tabela, pois os dados numéricos deixam as informações mais claras.

Tabela 6: Correlação de Spearman para as variáveis contínuas do banco de dados

	chol	stab.glu	hdl	glyhb	age	bmi	height	weight	waist	hip
chol	1.00	0.14	0.14	0.23	0.29	0.12	-0.12	0.05	0.10	0.10
stab.glu	0.14	1.00	-0.20	0.52	0.33	0.18	0.01	0.19	0.25	0.20
hdl	0.14	-0.20	1.00	-0.19	-0.05	-0.25	-0.13	-0.33	-0.32	-0.24
glyhb	0.23	0.52	-0.19	1.00	0.43	0.21	0.02	0.20	0.31	0.22
age	0.29	0.33	-0.05	0.43	1.00	0.01	-0.08	-0.02	0.17	0.04
bmi	0.12	0.18	-0.25	0.21	0.01	1.00	-0.24	0.84	0.81	0.88
height	-0.12	0.01	-0.13	0.02	-0.08	-0.24	1.00	0.28	0.07	-0.11
weight	0.05	0.19	-0.33	0.20	-0.02	0.84	0.28	1.00	0.85	0.81
waist	0.10	0.25	-0.32	0.31	0.17	0.81	0.07	0.85	1.00	0.83
hip	0.10	0.20	-0.24	0.22	0.04	0.88	-0.11	0.81	0.83	1.00

Fonte: Autoral

Nesta etapa da análise verificamos as variáveis que apresentam uma forte correlação positiva e portanto, devem explicar de forma redundante a resposta da variável dependente. Neste caso, para a construção do modelo de regressão múltipla, é selecionada apenas a variável que melhor explica a distribuição da variável dependente.

O teste t foi usado para testar a hipótese nula sobre a diferença das médias de duas amostras, no caso (com ou sem diabetes), assumindo que a variância da população não seja conhecida. De acordo com as informações contidas na tabela 7, é preciso aceitar a hipótese nula para as variáveis: *height*, *bp.1d*, *bp.2s*, *bp.2d*, *time.ppn* e *insurance*. Pois de acordo com o p-valor encontrado, não foi possível relacionar essas variáveis com o fato de ter DM2. Em relação às variáveis: *chol*, *stab.glu*, *hdl*, *ratio*, *glyhb*, *age*, *weight*, *bp.1s*, *waist*, *hip*, *fh*, *gyhb*, *bmi* foi preciso rejeitar a hipótese nula, pois de acordo com o p-valor encontrado nessas variáveis, foi possível afirmar que existe uma relação entre essas variáveis e o fato de ter DM2.

Tabela 7: Teste t entre as variáveis numéricas e a variável categórica DM2

……..…..……..……....Teste t entre as variáveis do *data.frame* x DM2……..…...……..…….					
	t	G.L. (Graus de Liberdade)	P-valor	Sem diabetes (média)	Com diabetes (média)
Chol*	-4.0978	387	5.084e-05	203.386	228.600
Stab.glu*	-18.73	388	2.2e-16	91.55152	194.16667
Hdl*	2.4418	387	0.01506	51.17629	45.28333
Ratio*	-5.5749	387	4.652e-08	4.324316	5.635000

Glyhb*	-34.183	388	2.2e-16	4.762303	10.140833
Age*	-6.24	388	1.147e-09	44.66061	58.40000
Height	-0.3931	383	0.6945	1.978528	1.985085
Weight*	-3.2054	387	0.001461	78.57136	86.72034
Bp.1s*	-3.9667	383	8.699e-05	135.1877	147.7667
Bp.1d	-0.9709	383	0.3322	82.99692	84.85000
Bp.2s	-0.67457	136	0.5011	152.1682	155.1613
Bp.2d	1.4091	136	0.1611	93.47664	90.16129
Waist*	-4.4718	386	1.021e-05	37.35061	40.88333
Hip*	-2.8082	386	0.005235	42.69207	44.90000
Time.ppn	-0.88294	385	0.3778	330.1835	368.5000
Gyhb*	-33.568	388	2.2e-16	4.260606	9.666667
Bmi*	-2.7759	382	0.005777	20.18871	22.03828

* Variáveis com p-valor ≤ 0.05

Fonte: Autoral

5.3. Modelo de Regressão Logístico

Para o desenvolvimento do modelo de regressão logístico, utilizamos as variáveis *age, gender, cholesterol, BMI* e *HDL,* alguns destes fatores já reconhecidamente relacionados ao desenvolvimento da diabetes. O objetivo desta etapa não foi construir um modelo de regressão otimizado a partir da análise e combinação de cada uma das variáveis preditoras, mas sim selecionar algumas destas variáveis para mostrar a utilidade desta abordagem para a análise de dados em saúde. Conforme descrito na metodologia, o desenvolvimento de um modelo de regressão se inicia com a verificação das distribuição dos valores das variáveis a serem utilizadas como preditores.

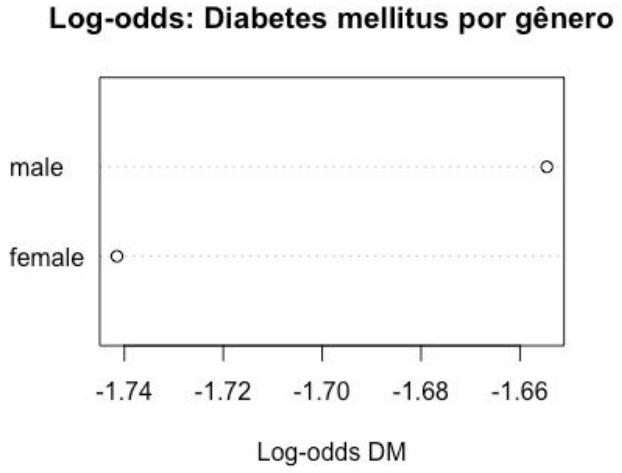
Após a análise exploratória inicial verificou-se, separadamente, a relação de cada uma das variáveis selecionadas e a variável dependente, DM, medida em termos de chances (odds) de desenvolvimento de diabetes. A chance é definida pela probabilidade de um evento ocorrer dividida pela probabilidade o evento não ocorrer (chance = p/1-p). Embora as probabilidades sejam a maneira mais intuitiva de modelar variáveis binárias, se modelarmos a

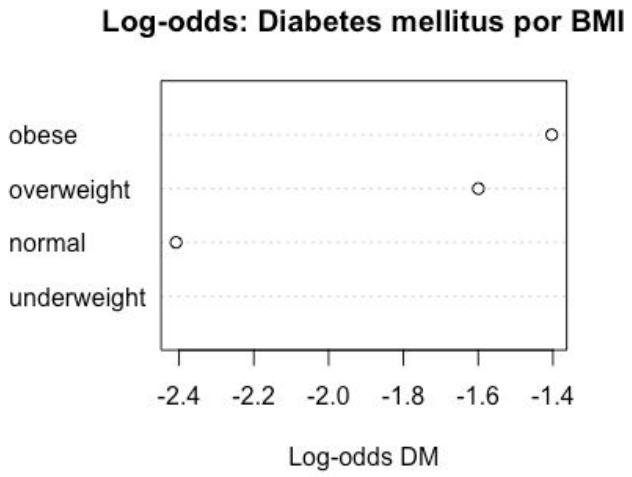
probabilidade como variável de resultado, teremos previsões impossíveis do modelo (ou seja, valores previstos abaixo de 0 ou acima de 1). A razão pela qual modelamos o log (odds) em vez das probabilidades como variável de resultado é porque as probabilidades só podem assumir valores positivos e, portanto, esse modelo também daria previsões impossíveis. O log (odds) pode assumir valores positivos ou negativos e, portanto, todas as previsões do modelo são possíveis.

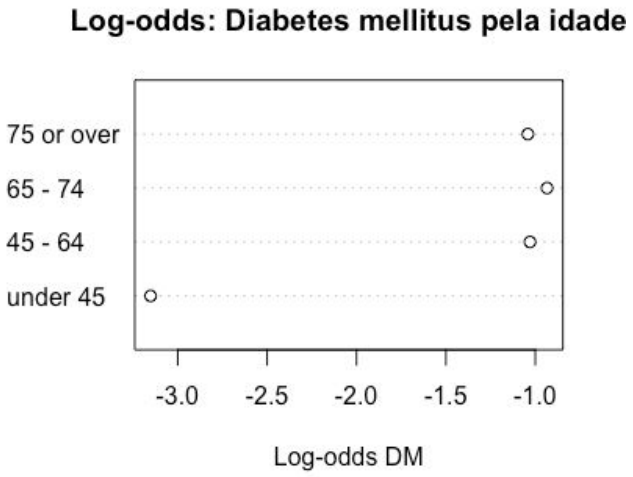
Geralmente, estudos de saúde pública costumam apresentar o desfecho de interesse em função da Razão de Chance (odds ratio), comparando a chance (odds) daquele evento dada a variável independente, ajustado para os outros fatores (no caso de regressão múltipla). É necessário certa cautela para interpretar estudos de saúde pública que apresentam grandes razões de chances. À medida que a probabilidade aumenta, as chances aumentam de forma não linear, isto é, um aumento relativamente pequeno na probabilidade pode levar a um grande aumento nas chances. Um outro motivo para lidar com chances em vez de probabilidades é que quando se duplica uma probabilidade, pode-se obter valores acima de 1, o que é impossível, mas as chances acima de 1 são completamente razoáveis.

A relação individual das variáveis *age, gender* e *BMI* e a chance de diabetes (log(odds)) é mostrado na figura 15. Pode-se observar que a chance de desenvolver diabetes é maior para indivíduos do sexo masculino, aumenta com a idade e o fato de ter sobrepeso ou ser obeso. Entretanto o nível de significância varia de acordo com a variável avaliada.

Figura 15: Relação entre as variáveis e a chance de diabetes







Fonte: Autoral

Após verificar novamente o coeficiente de correlação entre as variáveis, executamos o modelo, por meio da chamada da função **glm** do pacote estatístico básico do R. As informações mais relevantes são mostradas na figura 16, a partir da execução da função **summary** executada sobre o resultado da regressão.

A partir deste resumo podemos verificar que 19 pacientes foram excluídos da análise devido a valores ausentes. Este é um número relativamente pequeno, que não inviabiliza a análise. Da mesma forma, os erros-padrão (Std. Error) são pequenos e estão dentro de valores esperados, indicando que não há problemas relativos ao funcionamento do algoritmo ou categorias com poucos pacientes.

Figura 16: Resumo das informações referentes ao modelo de regressão logística

```
Call:
glm(formula = dm ~ age + gender + bmi, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6843  -0.5763  -0.3885  -0.2575   2.6991

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.892669   1.027388  -6.709 1.96e-11 ***
age          0.055454   0.009884   5.611 2.02e-08 ***
gendermale   0.244852   0.322817   0.758  0.44816
bmi          0.073879   0.023310   3.169  0.00153 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 326.02  on 383  degrees of freedom
Residual deviance: 281.62  on 380  degrees of freedom
  (19 observations deleted due to missingness)
AIC: 289.62

Number of Fisher Scoring iterations: 5
```

Fonte: Autoral

Verificando os coeficientes de regressão para cada variável, podemos observar que o log(odds) para a idade é 0,055, e pode-se concluir que a idade está significativa - e positivamente - associada ao risco de contrair diabetes (p-value = 2,02e-8). O R imprime todos os coeficientes na escala logarítmica, no caso de regressão logística, pois estamos modelando as chances logarítmicas de ter diabetes. Transformando o logaritmo por meio da exponenciação, obtém-se 1,057 (1,06 com duas casas decimais). Conforme explicado anteriormente, esta representação é chamada de odds ratio (razão de chance) e significa que um aumento de um ano na idade está associado a seis por cento mais chances de ser diagnosticado com diabetes. Neste caso, o efeito da idade é ajustado para os efeitos do sexo e do IMC. No entanto este efeito pode ser devido, pelo menos em parte, a fatores ainda não considerados no modelo.

Além dos coeficientes e p-valores é necessário calcular o intervalo de confiança (IC) de 95% usando o erro padrão. Sabemos que há alguma incerteza sobre essa estimativa de chances 6% mais altas, porque é baseada em uma amostra de pacientes. Para expressar essa incerteza, podemos executar a função **confint** no R.

Figura 17: Intervalo de confiança (95%)

```
> exp(confint(m))
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) 0.0001226733 0.006986056
age         1.0373532463 1.078493131
gendermale  0.6762611041 2.409679250
bmi         1.0287121260 1.127738696
>
```

Fonte: Autoral

Arredondando os ICs para duas casa decimais, podemos verificar que o *odds ratio* é de 1,06 situa-se dentro do IC 95% de 1,04 a 1,08, e portanto com o nível de significância de menor que 0,05. Em relação ao gênero, o p-valor de 0,448 está bem acima do limite padrão de 0,05. O IC de 95% é amplo, de 0,68 a 2,41, portanto, esse conjunto de dados não informa muito sobre a relação entre gênero e risco de diabetes. Podemos somente afirmar que não há fortes evidências de uma diferença de risco entre os sexos. Em relação ao IMC, a razão de chances para um aumento unitário no IMC é exp (0,073879) = 1,08, com IC95% 1,03 a 1,13, p = 0,002. Esse é um p-valor bastante baixo, e podemos concluir que pessoas com IMC mais alto correm mais risco de diabetes. Isso é exatamente o que se esperava, dado que 90% dos casos de diabetes são do tipo 2, que está associado ao estilo de vida, que afeta o IMC.

A partir dos resultados encontrados, foi possível observar e comparar os dados deste trabalho com o de outros trabalhos publicados anteriormente. O artigo "Type 2 diabetes in Brazil: epidemiology and management" teve como objetivo fazer uma revisão bibliográfica de artigos, publicações e documentos oficiais do governo brasileiro, contendo informações sobre o DM2 no Brasil. Os resultados apresentam informações semelhantes aos obtidos neste trabalho, mostrando que existe uma relação entre a obesidade e o DM2, como também um forte relação entre o envelhecimento da população com o aumento dos casos de DM2. Os autores concluem que a prevalência de diabetes no Brasil vem crescendo junto com o aumento da prevalência dos fatores de risco (ALMEIDA-PITITTO et al., 2015). A nível global, o artigo "Global aetiology and epidemiology of type 2 diabetes mellitus and its complications", publicado por Zheng e colaboradores teve como objetivo fazer uma revisão

sobre os casos de Diabetes Mellitus no mundo, observando os principais fatores de risco, as tendências globais da doença e suas complicações. Os resultados obtidos também comprovam os dados encontrados na análise feita neste trabalho, mostrando que a prevalência de diabetes vem crescendo paralelamente com o aumento de peso tanto nos países desenvolvidos como o EUA, quanto nos países em desenvolvimento, como a China. Além disso, foi observado que a crescente epidemia de diabetes tem múltiplas razões, como o envelhecimento da população, desenvolvimento, urbanização, hábitos alimentares pouco saudáveis e sedentarismo, corroborando alguns dos resultados observados no presente trabalho (ZHENG et al., 2018).

6. CONCLUSÃO

Este trabalho teve como objetivo a exploração das informações do banco de dados "Diabetes", para avaliação de correlações, fatores de risco e preditivos associados ao desenvolvimento da Diabetes Mellitus. A partir dos resultados encontrados, podemos concluir que a análise de dados para o banco de dados “Diabetes” chegou a resultados promissores sobre os fatores de risco do DM2. Observamos como o desenvolvimento do DM2 está relacionado com pacientes com maior índice de colesterol, IMC, idade e peso. Além disso, foi possível observar as estatísticas com os valores das correlações entre cada variável, mostrando o quanto cada objeto influencia no desenvolvimento do DM2. Este trabalho serve também como estudo de caso para mostrar a importância da integração e análise de dados epidemiológicos, clínicos e genéticos para o entendimento de processos e planejamentos em saúde.

Em trabalhos futuros, pretende-se otimizar e testar o modelo, considerando as outras variáveis disponíveis. Além disso, pretendemos fazer uma análise de dados sobre DM2 utilizando dados do Departamento de Informática do Sistema Único de Saúde (DATASUS - http://datasus.saude.gov.br/), avaliando as regiões com maior ocorrência de DM, os principais fatores de risco e as complicações mais presentes para cada região do Brasil. Abordando o mesmo problema a partir de outros dados, pretende-se também avaliar o impacto de variantes e/ou mutações genéticas no desenvolvimento de Diabetes mellitus, investigando as associações entre estas variantes e fenótipos, dados clínicos e sobrevida, utilizando dados do *Type 2 Diabetes Knowledge Portal* (T2D Knowledge Portal - http://www.type2diabetesgenetics.org/home/portalHome).

7. BIBLIOGRAFIA

ALVES,C. *et al.* **Distribuição e freqüência de alelos e haplótipos HLA em brasileiros com diabetes melito tipo 1.** Arq Bras Endocrinol Metab vol.50 no.3 São Paulo Jun. 2006.

ALI,O. **Genetics of type 2 diabetes.** World J Diabetes, 2013.

ALMEIDA-PITITTO,B. *et al.* **Type 2 diabetes in Brazil: epidemiology and management.** São Paulo: Deve Press, 2015.

ARAÚJO,L.M.B.; BRITTO,M.M.S.; CRUZ,T.R.P. **Tratamento do Diabetes Mellitus do Tipo 2: Novas Opções.** Arq Bras Endocrinol Metab vol.44 no.6 São Paulo Dec. 2000.

BATISTA,A.P. **Ancestralidade genética e marcadores de risco para a doença cardiovascular em distritos do município de Ouro Preto, Minas Gerais**. 2019. 206 F. Tese (Doutorado em Ciências Biológicas) - Núcleo de Pesquisa em Ciências Biológicas, Universidade Federal de Ouro Preto, Ouro Preto, 2019.

BAMSHAD,J.C. **Genética Médica.** 4ª ed. Elsevier, 2010.

Caderno de atenção básica diabetes mellitus. Disponível em:<http://bvsms.saude.gov.br/bvs/publicacoes/diabetes_mellitus.PDF>. Acesso em: 18 set. 2019.

CAROLINO,I.D.R. *et al.* **Fatores de risco em pacientes com diabetes mellitus tipo 2.** Rev Latino-am Enfermagem Mar. 2008.

DORIA,A.; PATTI,M-E.; KAHN,C.R. **The Emerging Genetic Architecture of Type 2 Diabetes.** Cell Metabolism Volume 8, Issue 3, Set. 2008.

121doc. Diabetes. Disponível em: <https://www.121doc.com/br/diabetes>. Acesso em: 28 de nov. 2019.

R. Bonita, R. Beaglehole, T. Kjellström; Epidemiologia básica [tradução e revisão científica Juraci A. Cesar]. - 2.ed. - São Paulo, Santos. 2010.

GERCHMAN,F. **Raça, etnia e as complicações crônicas do diabetes melito tipo 2.** 2004. 75 F. Dissertação (Mestrado em Ciências Médicas: Endocrinologia) - Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

GOMES,M.B. *et al.* **Prevalência de Sobrepeso e Obesidade em Pacientes Com Diabetes Mellitus do Tipo 2 no Brasil: Estudo Multicêntrico Nacional.** Arq Bras Endocrinol Metab vol 50 nº 1. 2006.

GRILLO,M.F.F.; GORINI,M.I.P. **Caracterização de Pessoas com Diabetes Mellitus tipo 2.** Rev Bras Enferm, Brasília Jan. 2007.

GUYTON, A.C.; HALL, J.E. **Tratado De Fisiologia Médica.** 12ª ed. Rio de Janeiro: Elsevier, 2011.

HAY,W.W. *et al.* **Current Pediatria: Diagnóstico e Tratamento.** 22ª ed. Porto Alegre: AMGH, 2016.

Independent Group of Scientists appointed by the Secretary-General, Global Sustainable Development Report 2019: The Future is Now – Science for Achieving Sustainable Development, (United Nations, New York, 2019).

International Diabetes Federation - IDF Diabetes Atlas, Brazil Country report 2017 & 2045. Disponível em:<https://reports.instantatlas.com/report/view/704ee0e6475b4af885051bcec15f0e2c/BRA>. Acesso em: 28 out. 2019.

ISER,B.P.M.*et al.* **Prevalência de diabetes autorreferido no Brasil: resultados da Pesquisa Nacional de Saúde 2013.** Brasília: Epidemiol.Serv.Saúde, 2015.

LYRA,R. *et al.* **Prevenção do diabetes mellitus tipo 2.** Arq Bras Endocrinol Metab vol.50 no.2 São Paulo Abr. 2006.

Ministério da Saúde. Alimentação adequada é grande aliada no combate ao pré-diabetes. Disponível em:<http://www.saude.gov.br/noticias/agencia-saude/44682-saiba-como-uma-alimentacao-adequada-pode-ser-aliada-no-combate-ao-pre-diabetes>. Acesso em: 25 set. 2019.

Ministério da Saúde. Diabetes (diabetes mellitus): Sintomas, Causas e Tratamentos. Disponível em:<http://www.saude.gov.br/saude-de-a-z/diabetes>. Acesso em: 03 out. 2019.

MOLENA-FERNANDES,C.A. *et al.* **A importância da associação de dieta e de atividade física na prevenção e controle do Diabetes mellitus Diabetes mellitus tipo 2**. Acta Sci. Health Sci. Maringá, v. 27, n. 2, p. 195-205, 2005.

ORTIZ,M.C.A.; ZANETTI,M.L. **Levantamento dos fatores de risco para diabetes mellitus tipo 2 em uma instituição de ensino superior.** Rev Latino-am Enfermagem Mai. 2001.

Press release. Available at. Disponível em: <https://sustainabledevelopment.un.org/content/documents/24791GSDR_press_release_for_11_Sept.pdf>. Acesso em: 20 de nov. 2019.

Portal Action. Introdução às Tabelas Cruzadas. Disponível em: <http://www.portalaction.com.br/tabela-de-contingencia/introducao-tabelas-de-contingencia>. Acesso em: 26 de nov. 2019.

Portal da Educação. Diabetes mellitus do tipo 2: Síndrome metabólica e modificação no estilo de vida. Disponível em: <https://www.portaleducacao.com.br/conteudo/artigos/farmacia/diabetes-mellitus-do-tipo-2-sindrome-metabolica-e-modificacao-no-estilo-de-vida/549>. Acesso em: 04 nov. 2019.

REIS,A.F.; VELHO,G. **Bases Genéticas do Diabetes Mellitus Tipo 2.** Arq Bras Endocrinol Metab vol 46 nº 4 Ago. 2002.

SARTORELLI,D.S.; FRANCO,L.J.; CARDOSO,M.A. **Intervenção nutricional e prevenção primária do diabetes mellitus tipo 2: uma revisão sistemática.** São Paulo: Scielo Saúde Pública, 2005.

SILVA,C.A.; LIMA,W.C. **Efeito Benéfico do Exercício Físico no Controle Metabólico do Diabetes Mellitus Tipo 2 à Curto Prazo.** Arq Bras Endocrinol Metab vol 46 nº 5 Out. 2002.

Só Biologia. Só ciências > Corpo humano. Disponível em: <https://www.sobiologia.com.br/conteudos/Corpo/sistemaendocrino2.php>. Acesso em: 28 out. 2019.

Sociedade Brasileira de Diabetes. Diabetes já atinge um milhão de crianças no Brasil. Disponível em: <https://www.diabetes.org.br/publico/para-voces/sbd-na-imprensa/902-diabetes-ja-atinge-um-milhao-de-criancas-no-brasil>. Acesso em: 28 out. 2019.

Sociedade Brasileira de Diabetes. Viva saudável com Diabetes!. Disponível em: <https://www.diabetes.org.br/publico/viva-saudavel-com-diabetes>. Acesso em: 29 out. 2019.

Sociedade Brasileira de Endocrinologia e Metabologia. Ministério da Saúde: Vigitel 2016. Disponível em: <https://www.endocrino.org.br/minsterio-da-saude-divulga-dados-do-vigitel-2016/>. Acesso em: 29 out. 2019.

TELO,G.H.*et al.* **Prevalence of diabetes in Brazil over time: a systematic review with meta-analysis.** Porto Alegre: BioMed Central, 2016.

TORRES,H.C.; PEREIRA,F.R.L.; ALEXANDRE,L.R. **Avaliação das ações educativas na promoção do autogerenciamento dos cuidados em diabetes mellitus tipo 2*.** Rev Esc Enferm USP Dez. 2010.

Waldman, Eliseu Alves Vigilância em Saúde Pública, volume 7 / Eliseu Alves Waldman ; colaboração de Tereza Etsuko da Costa Rosa. – – São Paulo : Faculdade de Saúde Pública da Universidade de São Paulo, 1998. – – (Série Saúde & Cidadania).

ZHENG,Y.; LEY,S.H.; HU,F.B. **Global aetiology and epidemiology of type 2 diabetes mellitus and its complications.** Nature Reviews vol 14 Feb. 2018.