UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

FELIPE MARTINS PEREIRA

Estudo utilizando Machine Learning da Relação dos Dados Coletados pelo SINASC e o Peso do Recém Nascido

Guaratinguetá
2021

Felipe Martins Pereira

Estudo utilizando Machine Learning da Relação dos Dados Coletados pelo SINASC e o Peso do Recém Nascido

Trabalho de Graduação apresentado ao Conselho de Curso de Graduação em Engenharia Elétrica da Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estatual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia Elétrica .

Orientador: Prof. Dr. Daniel Julien Barros da Silva Sampaio

Guaratinguetá
2021

Pereira, Felipe Martins
P436e Estudo utilizando Machine Learning da relação dos dados coletados pelo SINASC e o peso do recém nascido / Felipe Martins Pereira – Guaratinguetá, 2021.
35 f : il.
Bibliografia: f. 35

Trabalho de Graduação em Engenharia Elétrica – Universidade Estadual Paulista, Faculdade de Engenharia de Guaratinguetá, 2021.
Orientador: Prof. Dr. Daniel Julien Barros da Silva Sampaio

1. Inteligência artificial. 2. Redes neurais (Computação). 3. Cuidado Pré-Natal. I. Título.

CDU 681.3

Luciana Máximo
Bibliotecária CRB-8-3595

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

FELIPE MARTINS PEREIRA

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE **"GRADUANDO EM ENGENHARIA ELÉTRICA "**

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Prof. Dr. DANIEL JULIEN BARROS DA SILVA SAMPAIO
Coordenador

BANCA EXAMINADORA:

Prof. Dr. Daniel Julien Barros da Silva Sampaio
Orientador/UNESP-FEG

Prof. Dr. Luiz Fernando Costa Nascimento
UNESP-FEG

Prof. Dr. Fernando Ribeiro Filadelfo
Membro Externo

Março , 2021

Dedico este trabalho de modo especial, à milha família.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus, fonte da vida e da graça. Agradeço a minha vida, minha inteligência, minha família e meus amigos; aos meus orientadores, Prof. Dr. Daniel Julien Barros da Silva Sampaio e Prof. Dr. Luiz Fernando Costa Nascimento, que possibilitaram um desafio incrível e de grande aprendizado, além de me orientarem quando estava perdido; aos meus pais Valéria e Danilo, que apesar das dificuldades enfrentadas, sempre incentivaram meus estudos; às funcionárias da Biblioteca do Campus de Guaratinguetá pela dedicação, presteza e principalmente pela vontade de ajudar; aos funcionários da Faculdade de Engenharia do Campos de Guaratinguetá pela dedicação e alegria no atendimento.

“O insucesso é apenas uma oportunidade para recomeçar com mais inteligência.“
(Henry Ford)

RESUMO

O baixo peso ao nascer (BPN) é uma condição que impacta uma parcela considerável da população mundial e pode influenciar o surgimento de doenças durante a vida adulta, além de estar associado com a mortalidade neonatal. O conceito de BPN é apresentado pela OMS como crianças que nasceram com um peso inferior a 2500g e está associado a condição de vida da família da criança. O Sistema Único de Saúde (SUS) realiza a coleta de um conjunto de informações no momento do nascimento da criança e consolida no sistema SINASC, informações que estão relacionadas a condição da criança e a características da família. Com essas informações é possível desenvolver um modelo com o uso de técnicas de ciência de dados e de redes neurais artificiais para encontrar uma correlação desses dados ao BPN e possibilitar um melhor acompanhamento pré-natal para diminuir a incidência da condição nas crianças. A rede neural artificial construída foi por meio da biblioteca Keras, constituída de cinco camadas e que demonstrou bons resultados na em relação a classificação de baixo peso, demonstrando uma correlação dos dados do SINASC com o peso do recém nascido.

PALAVRAS-CHAVE: Baixo Peso ao Nascer. Ciência de Dados. Inteligência Artificial. Redes Neurais Artificiais. Dados do SINASC.

ABSTRACT

Low birth weight (LBW) is a condition that impacts a considerable portion of the world population and can influence the onset of diseases during adulthood, in addition to being associated with neonatal mortality. The concept of LBW is presented by WHO as children who were born weighing less than 2500g and is associated with the child's family's condition of life. The Sistema Único de Saúde (SUS) collects a set of information at the time of the child's birth and consolidates in the SINASC system, information that is related to the child's condition and family characteristics. With this information, it is possible to develop a model using data science techniques and artificial neural networks to find a correlation of these data to LBW and to enable better prenatal care to reduce the incidence of the condition in children. The artificial neural network constructed was through the Keras library, consisting of five layers and which demonstrated good results in relation to the classification of low weight, demonstrating a correlation of the SINASC data with the weight of the newborn.

KEYWORDS: Low birth weight. Data Science. Artificial Intelligence. Artificial neural networks. SINASC data.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de máquina
BPN	Baixo peso ao nascer
DATASUS	Departamento de Informática do Sistema Único de Saúde
IA	Inteligência Artificial
RNA	Rede neurais artificiais
SINASC	Sistema de Informações sobre Nascidos Vivos
SUS	Sistema Único de Saúde

SUMÁRIO

1 INTRODUÇÃO

Aprender é uma das palavras que desde os primeiros passos de uma criança na escola, até o último dia da vida dessa pessoa, está sempre por perto, mesmo que de forma discreta. Tudo que uma pessoa faz vem por meio de um aprendizado ou gera algum aprendizado. Desde a primeira ferramenta feita pelo homem, a primeira fogueira, as primeiras letras de um alfabeto e até mesmo as primeiras máquinas para transporte, tudo começa com um ciclo de observar algo e testar diversas hipóteses sobre aquilo.

Com o passar do desenvolvimento tecnológico, cada vez mais se produz informações sobre diversas observações e essas informações têm se tornado cada vez mais importantes para continuar o desenvolvimento tecnológico. Isso fez com que novas técnicas para abordar essas informações fossem criadas com a intenção de aproveitar tudo que está explícito e implícito no conjunto de dados coletados.

Algumas das primeiras abordagens para compreender um grande conjunto de informações sobre algo foram as técnicas estatísticas de média, mediana e moda, em que o conjunto de dados é observado de acordo com o valor que representa o meio daquele conjunto, o valor que está no meio e o valor que mais se repete, porém não traziam grandes resultados para avaliar tudo que estava acontecendo naquele processo.

Com o passar do tempo, surgiram mais técnicas como distribuição de frequência, distribuição normal, entre outras. Todas com a intenção de tentar entender o comportamento daquele objeto em estudo e tentar prever algum resultado para o processo.

Com o desenvolvimento da tecnologia computacional, novas abordagens surgiram, entre elas estão os conjuntos de técnicas denominadas aprendizado de máquina (AM) do inglês, Machine Learning (ML), na qual se utilizam da capacidade analítica dos computadores para tentar ensinar sobre o processo em análise, e a com esse ensinamento, prever valores ou situações sobre o conteúdo abordado.

O aprendizado de máquina consiste em possibilitar aos computadores a capacidade de aprender, na qual são construídos algoritmos que operam a partir de um modelo que possui um conjunto de informações amostrais de entrada e com isso é possível realizar previsões ou tomar decisões a partir dos dados apresentados anteriormente.

Existem duas formas de abordar o AM na construção de um algoritmo de aprendizado: o supervisionado e o não supervisionado. O primeiro se baseia na construção de um modelo na qual os dados são fornecidos com rótulos para cada agrupamento de atributos, isso faz com o modelo seja conduzido a interpretar as informações com uma classificação pré-estabelecida. Já a técnica de AM não-supervisionado, é apresentado ao modelo um conjunto de dados sem uma classificação prévia, e o modelo agrupa essas informações de acordo com correlações que ele encontra entre cada conjunto de atributos, sendo assim, impossível de uma pessoa prever quais relações o modelo criou.

O estudo em questão abordou a primeira técnica, aprendizado de máquina supervisionado, para tentar encontrar relações no conjunto de dados fornecidos pelo SUS no sistema SINASC, com a intenção de tentar prever por meio de uma técnica de aprendizado profundo de máquina do inglês, Deep Learning, qual seria o peso, categorizado em baixo e normal, da criança ao nascer de acordo com o conjunto de informações já armazenadas pelo sistema SINASC.

Em dezesseis de maio de 2019, A Organização Pan-Americana da Saúde (OPAS, 2019) trouxe em seu site uma publicação sobre o tema na qual apresenta que um em cada sete bebês em todo o mundo nascem com baixo peso, além de pontuar que 80% dos casos de mortes dos recém-nascidos do mundo, possuem baixo peso. Demonstrou também que os que sobrevivem têm um risco maior de desnutrição e problemas físicos e de possível desenvolvimento de diabetes, doenças cardiovasculares, entre outras durante a sua vida.

Com esse cenário, se utilizar de novas técnicas para aproveitar dados que já são coletados pelos sistemas de saúde do Brasil para tentar achar uma correlação entre o baixo peso e o conjunto de informações, podem trazer grandes desenvolvimentos sociais em relação a saúde dos recém-nascidos e melhores práticas para o acompanhamento de todo o processo de gestação da criança.

2 FUNDAMENTAÇÃO TEÓRICA

O capítulo tem por objetivo apresentar os conceitos relacionados ao objeto em estudo. No subcapítulo 2.1 temos uma contextualização do Sistema Único de Saúde (SUS) e do seu Sistema de Informações de Nascidos Vivos (SINASC). Subcapitulo 2.2 uma contextualização da situação brasileira e mundial em relação a crianças com baixo peso. Em seguida uma apresentação sobre conceitos relacionados a Inteligência Artificial, no subcapitulo 2.3, com um pouco sobre definições relacionadas a IA na seção 2.3.1. Na seção 2.3.2 um pouco sobre o conceito Machine Learning e na seção 2.3.3 apresentações sobre Redes Neurais Artificias. Na seção 2.4 uma apresentação sobre as ferramentas utilizadas.

2.1 SISTEMA ÚNICO DE SAÚDE E O SINASC

O SUS é um sistema reconhecido mundialmente pelo seu atendimento público à saúde. É considerado um dos maiores sistemas de saúde pública do mundo e proporcionou grandes desenvolvimentos sociais para a população brasileira, permitindo que uma grande parcela da população pudesse ter acesso ao tratamento médico.

Como apresentado pelo Ministério da Saúde (2015), o SUS teve a sua criação em 1988 pela Constituição Federal Brasileira, todos os brasileiros e brasileiras passaram, desde o nascimento, a ter direito aos serviços de saúde, na qual a assistência é integral e gratuita. Além da assistência gratuita, o SUS foi responsável por diversos programas que proporcionaram uma melhora de vida da população, entre eles estão o Programa Nacional de Imunização (PNI) – responsável por 98% das vacinas aplicadas no país, o atendimento ao público para transplantes de órgãos e ao atendimento de forma gratuita para a população de portadores de HIV.

Entre as iniciativas que auxiliam o SUS em suas operações, encontra-se o Departamento de Informática do Sistema Único de Saúde (DATASUS, 2020), que surgiu em 1991, e entre as suas competências está o papel de definir padrões para a captação de informações relacionadas à saúde.

Em quase 30 anos de atuação, o DATASUS desenvolveu diversos sistemas para tratamento de dados e entre eles está o Sistema de Informações de Nascidos Vivos (SINASC).

Com a intenção de reunir informações epidemiológicas referentes aos nascimentos no território nacional, o SINASC permite com que as secretarias de saúde possam acompanhar o desenvolvimento de doenças relacionadas ao nascimento, além de criar campanhas e melhorias no atendimento e acompanhamento da gestação da criança.

2.2 SITUAÇÃO MUNDIAL E BRASILEIRA EM RELAÇÃO AO BAIXO PESO

O conceito de baixo peso ao nascer (BPN) é apresentado pela OMS como crianças que nasceram com um peso inferior a 2,5kg. O BPN é um ponto de extrema importância durante a gestação da criança pois se associa com a mortalidade neonatal. Como demonstrado no artigo "Baixo peso ao nascer e seus fatores associados".

> O BPN é um importante problema de saúde pública, pois se associa com a mortalidade neonatal. Revisão sistemática da literatura até 2011 e metanálise avaliaram em 8,5 a razão de chances associada à mortalidade neonatal, em recém-nascidos a termo (37 semanas de gestação) com peso ao nascer <2,5kg.(3) No Brasil, estudo de corte sobre mortalidade neonatal entre 2011 e 2012 mostrou também que o BPN é um de seus fatores associados.(4) Além da mortalidade neonatal, o BPN se associa com morbidades, como asma(5) e hipertensão.(6) (MOREIRA; SOUSA; SARNO, 2018, p.2)

Como abordado por Tourinho e Reis, o peso da criança ao nascer está muito associado ao condicionamento ao qual aquela família vive e a alimentação da mãe durante todo o período de gestação. "Ele reflete as condições nutricionais da gestante e do neonato e tem influência direta no crescimento e desenvolvimento da criança e nas condições de saúde do indivíduo na vida adulta53,54."(TOURINHO; REIS, 2013)

Apresentado no dia de 16 de maio de 2019 no site OPAS Brasil (OPAS, 2019), em 2015 mais de 20 milhões de crianças tiveram BPN em todo o mundo. Além da informação de que mais de 80% dos recém nascidos que vem a óbito, entram na classificação de peso baixo.

Todo esse cenário, demonstra a necessidade do cuidado e acompanhamento adequados durante toda a gestação e como o auxílio de ferramentas para o acompanhamento do pré-natal da criança, pode gerar grandes impactos na vida dessas.

Em relação ao cenário global sobre baixo peso, no relatório "Global Nutrition Targets 2025" (WHO, 2014) é apresentado um comparativo sobre os percentuais dos nascimentos com baixo peso pelo mundo. Um dos pontos evidenciados, no quadro a seguir, é a relação da qualidade de vida da população na região e como isso impacta no percentual de BPN.

No Brasil, comparando os dados obtidos pelo SINASC nos períodos de 1996 e 2011, a distribuição fica com maior percentual nas Regiões Sudeste e Sul, e com menor expressão nas regiões Norte e Centro-Oeste, conforme demonstrado no trecho do artigo "Baixo Peso ao nascer e seus fatores associados".

Tabela 1 – Percentual de bebês com baixo peso em relação as regiões no mundo

Regiões	% com baixo peso ao nascer	% não pesados ao nascer
África Subsaariana	13	54
África Oriental e do Sul	11	46
África Ocidental e Central	14	60
Oriente Médio e Norte da África	-	-
Sul da Asia	28	66
Leste Asiático e Pacífico	6	22
América Latina e Caribe	9	10
Países menos desenvolvidos	13	46
Mundo	15	48

Fonte: WHO (2014).

> No Brasil, a avaliação dos dados do Sistema de Informações sobre Nascidos Vivos (SINASC) entre 1996 e 2011 mostrou ser de 8,0% a proporção de BPN nas 26 capitais dos Estados e em Brasília, tendo sido maior nas Regiões Sudeste (8,4%) e Sul (8,0%) e menor nas Regiões Norte (7,2%), Nordeste (7,6%) e Centro-Oeste (7,4%).(2) (MOREIRA; SOUSA; SARNO, 2018, p.2)

Outro fator importante apresentado no artigo foi o levantamento de que comparando os dados do SINASC entre 1996 e 2011, o grau de escolaridade da mãe e a melhoria na cobertura de cuidados durante o pré-natal se associaram com a redução do risco de BPN, demonstrando uma grande importância no acompanhamento de informações relacionadas ao cotidiano da gestante.

2.3 INTELIGÊNCIA ARTIFICIAL

Identificar padrões foi o que trouxe o homem ao mundo que ele vive hoje. Todo desenvolvimento existente partiu da análise de um conjunto de informações decorrentes de acontecimentos, seja esses gerados de forma proposital ou simplesmente aleatória. E com essa análise, o surgimento de teorias e produtos fundamentados no conhecimento adquirido.

Todo esse aspecto de observar fenômenos e consolidar padrões em relação ao mesmo é o que vem fundamentar o conceito de aprendizado, um dos valores que torna o ser humano uma das espécies diferenciada em relação as demais existentes no planeta terra.

Aprender é poder reconhecer algo e classificá-lo em um certo grupo, porém não para em simplesmente em uma única classificação, mas sim com a exposição dos eventos, melhorar cada vez mais a classificação realizada, a fim de tornar-se mais próximo da realidade. Todo esse processo é o que permitiu o homem alcançar o conceito de inteligência.

Devido ao desenvolvimento da área de tecnologia computacional, uma das grandes questões que se tornou um dos principais pilares dessa área, foi a aplicação de métodos e metodologias para tentar implementar o conceito de inteligência nas aplicações fornecidas para o mercado.

Todos os esforços para implementar ferramentas que conseguissem prever e aprender com o comportamento dos consumidores, trouxeram grandes resultados para as organizações e passaram a ser o modelo de negócio de grandes impérios da tecnologia de hoje.

2.3.1 Inteligência Artificial

> Um bom médico consegue, dado o conjunto de sintomas e de resultados de exames clínicos, diagnosticar se um paciente está com problemas cardíacos. Para isso o médico utiliza conhecimento adquirido durante sua formação e experiência proveniente do exercício da profissão. Como escrever um programa de computador que, dados os sintomas e os resultados de exames clínicos, apresente um diagnóstico que seja tão bom quanto o de um médico experiente?(FACELI et al., 2011, p.2)

Um exemplo muito claro da utilização do conceito de inteligência artificial é apresentado no livro dos autores Katti Faceli, Ana Carolina Lorena, João Gama e André C. P. L. F. de Carvalho, na qual a inteligência artificial é um conjunto de técnicas que tentam aproximar os resultados obtidos por uma ferramenta ao diagnóstico de um profissional experiente na área, ou seja, é por meio de um programa alcançar resultados que precisariam de um grande conhecimento e experiência, pontos que normalmente levam anos para serem adquiridos.

Essas abordagens não apenas permitem trazer resultados assertivos em relação ao que está sendo analisado, como no caso de um diagnóstico médico, mas também descobrir padrões que seriam de certa forma impossíveis de serem classificados por um ser humano devido ao grande volume de informações necessários para reconhecer esse padrão, como identificar qual seria um produto interessante para uma determinada pessoa a partir dos dados de navegação *web* deste indivíduo.

A inteligência artificial pode ser conceituada em duas dimensões e que se dividem em quatro categorias menores. As categorias superiores ao quadro apresentado no livro Inteligência Artificial (RUSSEL; PETER, 2013) representam a busca pelo pensamento e raciocínio, enquanto as inferiores referem a comportamento. As definições do lado esquerdo medem à aproximação ao desempenho humano, enquanto ao lado direito a aproximação a uma inteligência ideal.

Essas diferentes formas de pensar em relação a Inteligência Artificial (IA) se reflete as inúmeras técnicas de IA que surgiram para criar programas. Algumas se utilizam de estudos da ciência cognitiva para criar modelos que se aproximem a forma de como a mente humana funciona, desde a criação do pensamento até a ação tomada. Um conjunto de técnicas utiliza-se do teste de Turing para tentar alcançar o conceito de IA, na qual o modelo precisa ter uma visão computacional e capacidades robóticas para poder manipular objetos e se movimentar. Outras se baseiam no puro conceito de lógica, na qual um determinado conjunto de ações leva a uma conclusão correta. E por fim temos técnicas que se baseiam e encontrar o melhor resultado

Figura 1 – Algumas definições de inteligência artificial

Pensando como um humano	**Pensando racionalmente**
"O novo e interessante esforço para fazer os computadores pensarem (...) *máquinas com mentes*, no sentido total e literal." (Haugeland, 1985) "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985) "O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992)
Agindo como seres humanos	**Agindo racionalmente**
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzweil, 1990) "O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas." (Rich and Knight, 1991)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes." (Poole *et al.*, 1998) "AI... está relacionada a um desempenho inteligente de artefatos." (Nilsson, 1998)

Fonte: Russell; Peter (2013).

possível, na qual analisa o conjunto de informações e a partir das possíveis decisões, assume aquela com o melhor resultado.

2.3.2 Aprendizado de Máquina

Entre as diversas Técnicas de Inteligência Artificial (IA) encontra-se o Aprendizado de Máquina (AM). O AM nada mais é do que um processo de indução do resultado (ou aproximação) a partir de um conjunto informações fornecidas. Uma definição que engloba mais aspectos em relação a essa técnica é apresentada por Mitchel (MITCHEL, 1997) como "A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência.", ou seja, aprendizado de máquina é aprender com as experiências passadas.

Para que esse processo de aprendizado aconteça, o conceito de indução é empregado e ele permite obter conclusões genéricas a partir de um conjunto de informações que demonstram como o problema pode ser resolvido

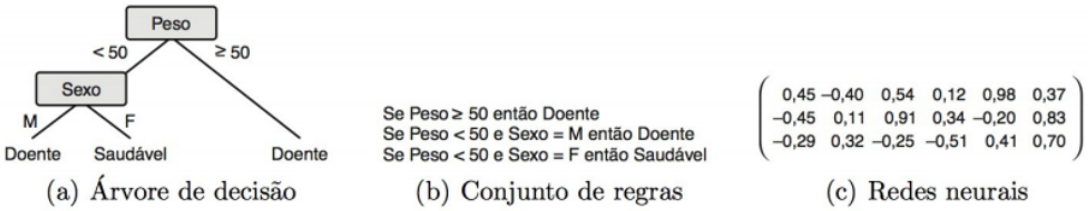
2.3.2.1 Vieses de um Aprendizado de Máquina

Um algoritmo de AM está sempre buscando a melhor hipótese para o problema, em um conjunto de hipóteses possíveis e essa busca se baseia em uma estrutura dos possíveis resultados, forma que é denominada como "Viés Indutivo".

Existem diversas formas de conseguir comparar os principais possíveis resultados, entre elas encontram-se a Árvore de decisão, regras de decisão e redes neurais. A Árvore de decisão é baseada em uma estrutura de árvore na qual cada ponto de decisão interno (nó) é representado por uma pergunta referente a alguma característica do objeto em análise e o nó externo representa

o resultado referente ao objeto (Classificação). As redes neurais artificiais representam uma hipótese a partir de um conjunto de valores reais, associados aos pesos das conexões da rede. Na Figura 2 esses conceitos são apresentados de forma visual.

Figura 2 – Diferentes vieses de representação



fonte: Faceli et al (2011).

O Viés de busca é a forma com que o algoritmo irá interpretar a representação a fim de encontrar o que melhor se ajusta aos dados de treinamento. A combinação desses dois vieses é o que permite um algoritmo de AM restringir os resultados e manter o aprendizado generalizado durante o treinamento.

Quando trabalhamos com algoritmos de AM, um dos critérios que precisa ser definido para a construção é o paradigma de aprendizado, na qual pode ser divido em duas categorias, preditivo ou descritivo.

2.3.2.2 Paradigmas de Aprendizado

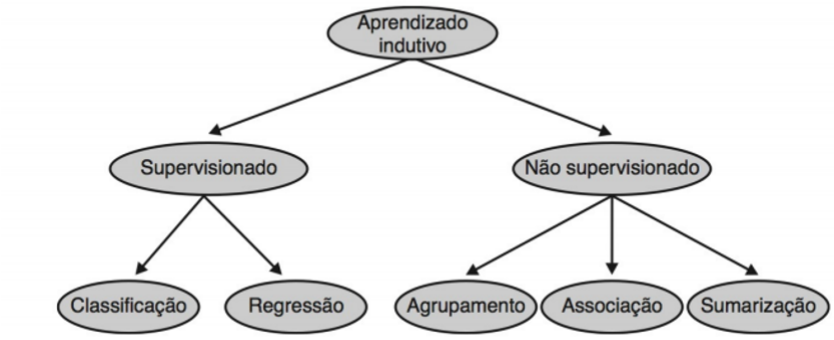
Em tarefas na qual tem a necessidade de realizar uma previsão, o objetivo é determinar um modelo que a partir dos dados fornecidos para treinamento, dados fornecidos por um supervisor (entidade que conhece a saída), consiga determinar rótulos corretos para novas situações. Como essas informações são fornecidas já com o conhecimento do resultado, esse paradigma é denominado como aprendizado supervisionado. Os rótulos são a classificação de saída dos dados e podem ser encontrados na literatura como classe do dado ou atributo de saída.

No algoritmo de aprendizado supervisionado, ter um conjunto de dados com valores nominais em seus resultados é considerado como um algoritmo de classificação. No caso de os dados representar um conjunto infinito e ordenado, um algoritmo de regressão.

Para tarefas na qual a meta é explorar um conjunto de dados ou descrever, os algoritmos de AM não se utilizam de uma referência de saída e devido a essa característica, são denominados paradigma de aprendizado não supervisionado. Nessa situação o algoritmo tenta descobrir semelhanças entre os dados e agrupá-los em grupos. Os algoritmos podem ser de sumarização, associação e agrupamento. A figura 3 traz de forma visual esses conceitos.

Um dos algoritmos de classificação que será abordado na aplicação do trabalho é o algoritmo de supervisionado de classificação com redes neurais artificiais.

Figura 3 – Hierarquia de aprendizado



Fonte: Faceli et al (2011).

2.3.3 Redes Neurais Artificiais

Ensinar uma máquina a realizar tarefas simples como pegar um objeto e colocar em outro lugar, ou percorrer um certo caminho, depende da análise de diversas informações e que pode trazer uma grande dificuldade na hora de programar um robô para realizá-las.

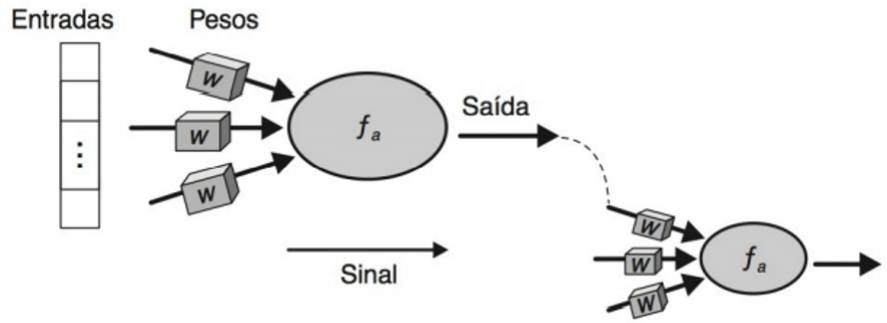
Para o homem, depois de um certo estágio do desenvolvimento da criança, essas tarefas passam a ser simples e realizadas de forma natural e sem muitas dificuldades, porém o número de informação analisada para realizá-las não diminuiu, apenas o cérebro da criança aprendeu com experiencias passadas a identificar quais informações são importantes para cada situação.

O funcionamento do cérebro humano é o que trouxe o conceito de uma rede neural artificial, na qual um conjunto de células determinam a resposta para um conjunto de estímulos. Cada célula recebe os estímulos, combina, processa e dependendo da intensidade e frequência desses estímulos, gera novos impulsos que podem impactar novas células. Essa forma de processar a informação é o conceito do neurônio em redes neurais.

O neurônio é a unidade de processamento de uma RNA e a função dele é basicamente gerar um estímulo de acordo com a informação fornecida no terminal de entrada. Ele recebe uma informação em seus terminais, essa informação sofre um certo ajuste de acordo com o peso determinado anteriormente para aquele tipo de informação e então essas informações ajustadas são agregadas e combinadas de acordo com uma determinada função, gerando assim uma saída como estímulo do conjunto de informações apresentadas. A Figura 4 traz uma representação de todo o processo.

A função de ativação é o que determina se um determinado neurônio será ativado e existem diversas funções propostas na literatura. Entre elas existem as funções linear, limiar e sigmoidal, na qual cada uma gera uma resposta diferente ao estímulo de entrada. Na função linear o processamento retorna como saída o valor de entrada, enquanto na função limiar a saída será igual a 1 ou 0 dependendo se o valor de entrada atingiu um determinado valor. Na função sigmoidal ela se aproxima da função limiar gerando dois valores de saída, 1 ou 0, porém ela possui um comportamento diferencial, o que faz responder com uma certa faixa para o valor de

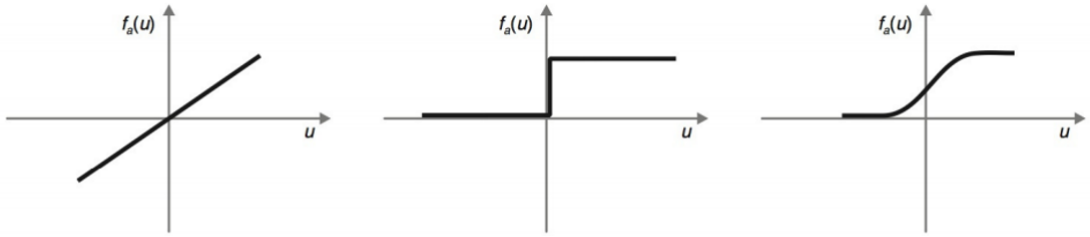
Figura 4 – Neurônio artificial



Fonte: Faceli et al (2011).

entrada. Alguns exemplos de funções são demonstrados na figura 5.

Figura 5 – Funções de ativação



(a) Linear (b) Limiar (c) Sigmoidal

Fonte: Faceli et al (2011).

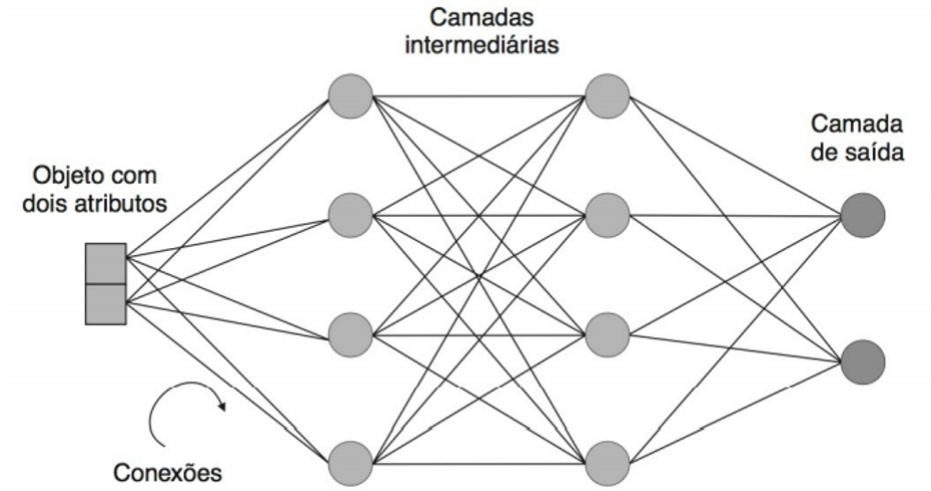
A construção de uma RNA pode ser com os neurônios dispostos em uma única camada ou em mais camadas. Na situação de multicamadas, uma parte dos neurônios estão ligados a outros neurônios e portando os valores de entrada deles serão os valores de saída do anterior, gerando assim uma rede de neurônios. Essa rede pode ser construída com diferentes padrões de relacionamentos entre os neurônios, classificados como completamente, parcialmente e localmente conectada. A figura 6 demonstra uma RNA de multicamadas.

Uma rede completamente conectada significa dizer que cada neurônio de uma camada está conectado a todos os outros neurônios da camada anterior. Da mesma forma, uma rede parcialmente conectada é quando o neurônio está conectado a apenas alguns dos neurônios da camada anterior. Uma rede localmente conecta significa que apenas uma parte dos neurônios estão conectados a um neurônio, e esse conjunto está localizado em uma região bem definida.

Um ponto importante em uma RNA é o ajuste dos parâmetros, pesos associados as conexões, e na literatura existem diversos algoritmos e arquiteturas. As principais arquiteturas e algoritmos são: a Rede Perceptron e Adaline, Perceptron Multicamadas e o algoritmo *Back-propagation*.

O algoritmo *Back-propagation* é constituído em duas fases com nome *forward* e *backward*. A fase *forward* o fluxo da informação acontece até o resultado de saída da RNA, onde cada neurônio recebe o conjunto de informações, processa e envia o resultado para a próxima camada. No final o valor obtido é comparado ao valor esperado e da comparação é gerado uma diferença

Figura 6 – Exemplo de RNA multicamadas típicas



Fonte: Faceli et al (2011).

entre os valores que indica o erro cometido pela rede. Com esse erro, se inicia a fase *backward*, na qual os pesos de entrada de cada neurônio são ajustados para se obter um valor próximo ao esperado, seguindo desde a camada de saída até a primeira camada intermediária.

2.4 FERRAMENTAS

Existem diversas abordagens e ferramentas em ciência de dados para compreender um conjunto de dados e identificar padrões. Abordagens que trazem linguagens como Python, R e Java como base para construção de modelos.

O Python possui uma grande popularidade para construção de modelos preditivos e o uso vem cada vez mais sendo prioritário em análises de modelos de dados e principalmente na utilização em RNA. Essa evolução do uso se deve a curta curva de aprendizagem necessária para utilizar a linguagem e o grande volume de bibliotecas e recursos que a comunidade da linguagem contribui com criações.

Entre as principais bibliotecas Python utilizadas em ciência de dados (ACADEMY, 2018), estão:

- Arrow: permite criar, alterar, remover e converter datas e horários com razoável simplicidade;
- Numpy: permite o trabalho facilitado para o processamento de matrizes e vetores, trazendo frameworks para manipular essas as informações contidas nas estruturas de forma rápida e eficiente;
- Matplotlib: voltada para a criação de análises estatísticas e visualização dos dados;

- Pandas: Uma das bibliotecas mais importantes no trabalho com ciência de dados devido ao amplo conjunto de ferramentas que disponibiliza e a possibilidade de trabalhar com diversas estruturas de dados, desde dados estruturados até séries temporais;
- Seaborn: biblioteca voltada a visualização de dados baseada no matplotlib que permite a criação de interfaces elaboradas de apresentação de dados de forma simplificada;
- Scikit Learn: voltada ao aprendizado de máquinas que possui diversos algoritmos voltado a classificação, regressão e agrupamento;
- Tensor Flow: biblioteca para aprendizado de máquina aplicável a uma ampla variedade de tarefas se utilizando de redes neurais;
- Keras: Api baseada no Tensor Flow que tenta facilitar a utilização e criação de redes neurais. Trazendo construções simplificadas para a criação de um modelo.

Duas ferramentas utilizadas na criação de códigos voltados para a ciência de dados são o Jupyter notebooks e o Google Colab. O Jupyter notebooks é um projeto open source que permite a criação de documentos que misturam programação, gráficos e textos para a criação de análises de dados. Nele é possível adicionar novas bibliotecas e até mesmo utilizar as bibliotecas fornecidas pela ferramenta. O Google Colab é uma ferramenta web baseada no Jupyter notebooks que permite a criação, da mesma forma que o Jupyter, de documentos voltados a ciência de dados. O grande diferencial do Google Colab é a disponibilização de uma máquina virtual para o processamento do arquivo, o que permite criar análises em diversos dispositivos.

Além da criação do código do modelo por meio de linguagens e bibliotecas específicas, algumas empresas fornecem aplicações para a criação de modelos de IA de forma visual como o IBM Watson Studio (IBM, 2021) e a Azure Machine Learning (AZURE, 2021).

3 APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS AO BANCO DO SINASC

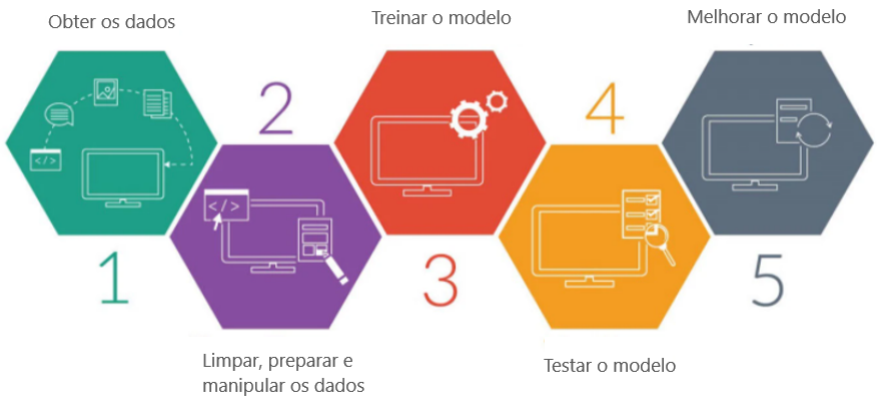
Neste capítulo é apresentado todo o processo de análise de dados e aplicação de aprendizado de máquina no conjunto de dados obtidos do SINASC. Composto por procedimentos metodológicos, o passo a passo para obter os dados, tratamento e análise e a aplicação de um algoritmo de redes neurais para processamento dos dados.

3.1 PROCEDIMENTOS METODOLÓGICOS

O trabalho se iniciou com a busca e aprendizado sobre o tema em diversas bibliografias relacionadas a ciência de dados e construção de modelos de dados, além da busca de informações em sites relacionados ao tema. Na avaliação e entendimento sobre a condição baixo peso a principal fonte de informação foram artigos produzidos por órgãos reconhecidos e pesquisas acadêmicas. Todos os trabalhos e bibliografias encontram-se referenciadas no final do trabalho.

Com a construção de uma base de conhecimento, se iniciou o processo de construção do modelo de RNA, na qual pode ser divido em cinco grandes etapas, conforme demonstrado na Figura 7.

Figura 7 – Etapas na construção de um aprendizado de máquina



Fonte: Machinelearning (2021)

O trabalho constitui-se em coletar: os dados no sistema SINASC (etapa 1); analisar e preparar os dados coletados (etapa 2); criar um modelo com a biblioteca Keras e treinar o modelo (etapa 3); testar o modelo com valores separados para teste (etapa 4); buscar formas de melhorar os resultados (etapa 5).

Em relação ao escopo do conjunto de dados analisados uma limitação sobre a abordagem de informações voltadas as regiões foi uma premissa do trabalho, na qual foram definidas as regiões da cidade de Guaratinguetá, Pindamonhangaba, São Jose dos Campos e Taubaté, além de só

ser considerado registros com o atributo de semanas de gestação maior que 35 semanas e com gravidez única.

A utilização da biblioteca Keras foi uma premissa para apresentar a análise sem necessariamente considerar o desempenho dela para os dados apresentados. Todos os processos foram construídos na ferramenta Google Colab e sempre se utilizando da linguagem Python.

3.1.1 Coleta dos dados

A coleta de dados foi realizada por meio do site do DATASUS , na qual é disponibilizado o acesso a diversos sistemas de informação. O sistema utilizado foi o SINASC que disponibiliza informações relacionadas a Nascidos Vivos.

O arquivo fornecido é no formato DBC e a sua utilização foi por meio da conversão em um formato CSV utilizando a ferramenta TABWIN, por meio da funcionalidade Comprime/Expande arquivos DBF. A ferramenta TABWIN é fornecida no próprio site do DATASUS.

O arquivo convertido CSV corresponde a dados dos anos 2017, 2018 e 2019. Composto por 1.801.140 registros e 60 variáveis inicialmente, sendo elas e suas descrições:

- DTNASC – Data do nascimento;
- HORANASC – Horário do nascimento do recém-nascido;
- SEXO – Sexo do recém-nascido;
- RACACOR – Cor ou raça do recém-nascido;
- PESO – Peso em gramas verificado até a 5ª hora após o nascimento;
- APGAR1 – Valor do índice de Apgar, medido no 1º minuto de vida;
- APGAR5 – Valor do índice Apgar, medido no 5º minuto de vida;
- IDANOMAL – Anomalia congênita identificada no momento do nascimento;
- LOCNASC – Local onde ocorreu o nascimento;
- CODESTAB – Código do Cadastro Nacional de estabelecimento onde ocorreu o nascimento;
- CODMUNNASC – Código de município de nascimento;
- ESCMAE – Escolaridade, em anos de estudo concluídos. Código agregado da escolaridade da mãe para o modelo antigo da DN;
- ESCMAE2010 – Escolaridade 2010;
- SERIESCMAE – Série escolar da mãe;

- CODOCUPMAE – Código de ocupação da mãe conforme tabela do CBO (Código Brasileiro de Ocupações;
- DTNASCMAE – Data de nascimento da mãe;
- IDADEMAE – Idade da mãe;
- NATURALMAE – Se a mãe for estrangeira, constará o código do país de nascimento;
- ESTCIVMAE – Situação conjugal da mãe;
- RACACORMAE – Tipo de raça e cor da mãe;
- CODMUNRES – Código do município de residência;
- IDADEPAI – Idade do pai;
- QTDGESTANT – Número de gestações anteriores;
- QTDPARTNOR – Número de partos vaginais;
- QTDPARTCES – Número de partos cesáreos;
- QTDFILVIVO – Número de filhos vivos;
- QTDFILMORT – Número de perdas fetais e abortos;
- DTULTMENST – Data da última menstruação (DUM);
- GESTACAO – Semanas de gestação agrupado;
- SEMAGESTAC – Número de semanas de gestação;
- TPMETESTIM – Método utilizado para verificar nº de semanas de gestação;
- CONSULTAS – Número de consultas de pré-natal agrupada;
- CONSPRENAT – Número de consultas pré-natal por semana de gestação;
- MESPRENAT – Mês de gestação em que iniciou o pré natal;
- GRAVIDEZ – Tipo de gravidez;
- TPAPRESENT – Tipo de apresentação do RN;
- STTRABPART – Questionamento sobre indução do trabalho de parto;
- PARTO – Tipo de parto;
- STCESPARTO – Cesárea ocorreu antes do trabalho de parto iniciar;

- TPNASCASSI – Formação da pessoa que assistiu ao parto. Em partos assistidos por equipe multiprofissional, formação da pessoa que coordenou os trabalhos;
- CODANOMAL – Descrição completa da anomalia congênita identificada no momento do nascimento, segundo a CID 10;
- DTDECLARAC – Data do preenchimento da declaração;
- TPFUNCRESP – Tipo de documento do responsável pelo preenchimento da declaração;
- TPDOCRESP – Tipo de documento do responsável pelo preenchimento da declaração;
- ESCMAEAGR1 – Escolaridade 2010 agregada;
- STDNEPIDEM – Status de DN Epidemiológica;
- STDNNOVA – Status de DN Nova;
- CODPAISRES – Código do país de residência;
- TPROBSON – Código do Grupo de Robson, gerado pelo sistema;
- PARIDADE – Variável calculada pelo sistema;
- KOTELCHUCK – Códigos de classificação de adequação ao pré natal, gerado pelo sistema;
- CODUFNATU – Código da UF de naturalidade da mãe;
- CODMUNNATU – Código do município de naturalidade da mãe;
- DTRECORIGA – Data de recebimento original calculado pelo sistema;
- DIFDATA – Diferença entre a data de óbito e data do recebimento original da DO ([DTNASC] – [DTRECORIG]);
- DTRECEBIM – Data do último recebimento do lote, dada pelo Sisnet;
- VERSAOSIST – Versão do sistema;
- NUMEROLOTE – Número do lote;
- DTCADASTRO – Data do cadastro da DN no sistema;
- ORIGEM – Banco de dados de origem;

Os bancos de dados são fornecidos pelo sistema em arquivos separados para cada ano e foi necessário a concatenação e ajuste dos índices por meio da biblioteca pandas.

3.1.2 Limpeza e adequação dos dados

Na segunda etapa, a compreensão e adequação dos dados é um ponto importante para que o modelo de RNA possa compreender essas informações. Isso se deve ao fato de o cadastro das informações não terem todos os campos como obrigatórios, gerando assim a falta de informações para algumas variáveis, além de diferentes padrões e formatos que necessitam de uma atenção para evitar distorções. A limpeza e adequação pode ser dividida em três grupos de atividades:

- Compreensão das categorias de dados e comportamentos;
- Exclusão dos dados irrelevantes;
- Ajustes de acordo com a categoria para treinamento.

3.1.2.1 Compreensão das categorias de dados e comportamentos

O primeiro passo é separar as variáveis sistêmicas de controle do conjunto de dados. O dicionário de dados fornecidos pelo Ministério da Saúde permite a rápida identificação dessas variáveis.

Com o conjunto de dados relacionados ao cadastro do recém-nascido, o próximo passo é a compreensão dos tipos de atributos apresentados no problema. Os atributos podem ser divididos em qualitativo, quantitativo discreto e quantitativo contínuo.

A diferenciação entre qualitativo e quantitativos é essencial pois para cada um é necessária uma abordagem diferente de preparo para o processamento no modelo. Os dados qualitativos são abordados como atributos que possuem valores binários e por isso é necessário a quebra desses atributos pelo número de categorias presentes nele, onde cada novo atributo representa uma categoria. Já os atributos quantitativos podem ser tratados como valores ajustados a uma determinada escala ou mesmo divididos em categorias e por consequência seguem o mesmo tratamento de atributos qualitativos.

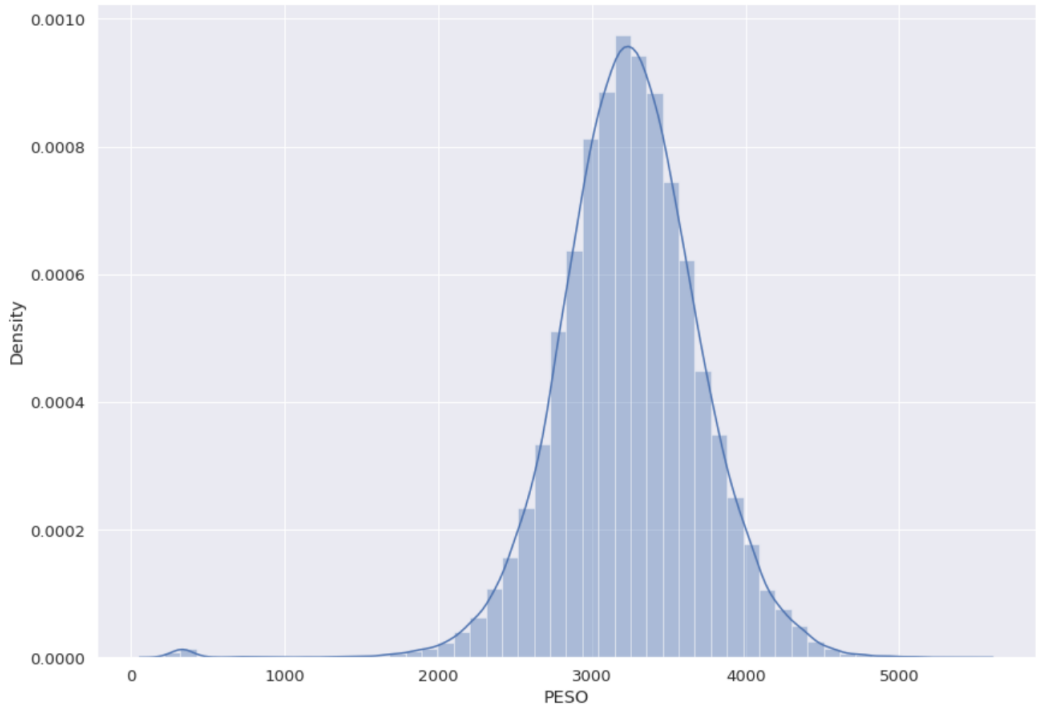
Com a divisão das categorias, inicia-se a compreensão dos atributos e os relacionamentos entre eles, principalmente para identificar se temos uma mesma informação sendo apresentado em mais um atributo.

Por fim, é realizado a compreensão do atributo “Target” que no caso do trabalho é o Peso, sobre a sua distribuição no conjunto de dados e a correlação com os demais atributos.

Quando observado a distribuição do Peso, observasse um desbalanceamento em relação ao número de recém-nascidos com peso baixo em relação aos que nascem com peso superior, conforme demonstrado na figura 8.

E que fica mais evidente o desbalanceamento entre os dados quando separados em categorias como “baixo peso” para valores inferiores a 2500g e “peso” para os demais valores. A distribuição demonstrou um percentual de aproximadamente 4,4% para valores na categoria “baixo peso” e 95,6% para valores na categoria “peso”.

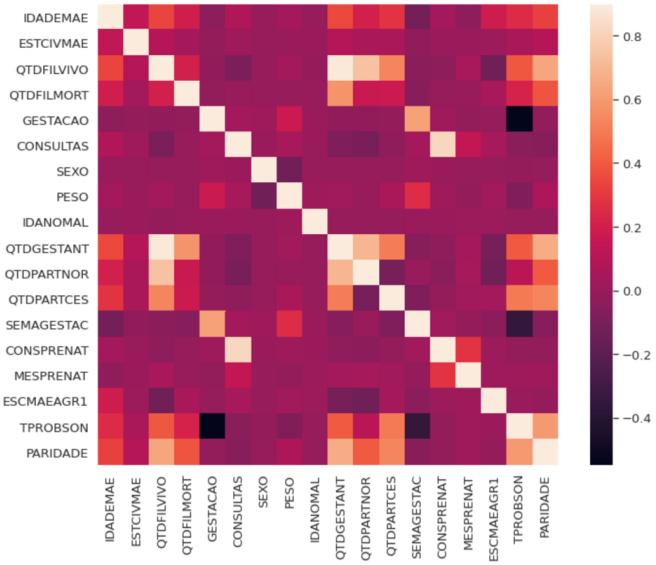
Figura 8 – Distribuição do Peso



Fonte: Produção do próprio autor

Comparando as variáveis que não são de controle do sistema por meio de um mapa de calor relacionado a correlação entre elas, é evidenciado alguns atributos que possuem uma certa influência em relação ao atributo PESO, entre elas a GESTACAO, CONSULTAS, SEXO. Conforme demonstrado na figura 9.

Figura 9 – Mapa de correlação entre as variáveis

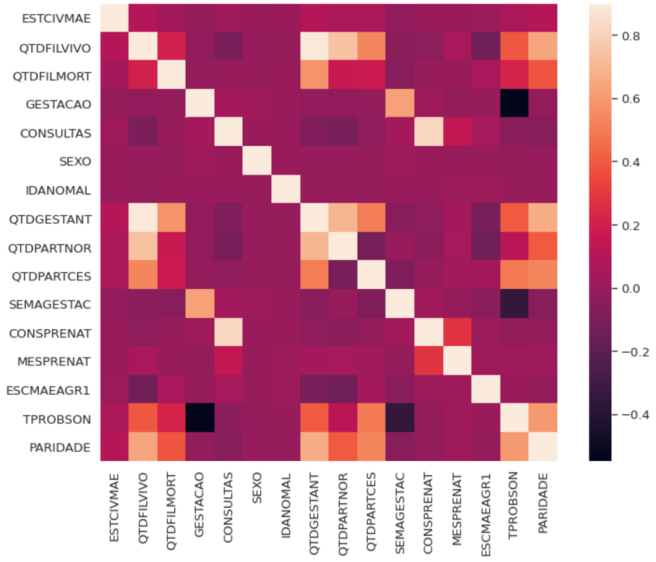


Fonte: Produção do próprio autor

3.1.2.2 Exclusão dos dados irrelevantes

Com a análise do mapa de correlação e considerando o dicionário de dados, alguns atributos foram eliminados. Com a eliminação dos dados, sobraram os atributos demonstrados na figura 10.

Figura 10 – Mapa de correlação entre as variáveis depois da exclusão



Fonte: Produção do próprio autor

3.1.2.3 Ajustes dos atributos

Nessa etapa o principal fator é entender os valores faltantes para cada atributo e ajustar o banco de dados. Todos os procedimentos são realizados com a biblioteca Pandas e funções desenvolvidas para ajuste dos valores em relação a categoria de dados envolvido em cada campo.

Em relação aos valores faltantes, alguns atributos demonstraram muitos dados faltantes. A figura 11 traz essa relação.

Para os registros ao qual algum dos atributos não estava com valor preenchido, a eliminação por completo da linha foi o caminho escolhido para evitar qualquer desvio na precisão do modelo. Com a eliminação das linhas o banco de dados passou a ter 1560 registros com classificação baixo peso e 34232 com classificação peso.

O processo para adequação dos atributos qualitativos foi por meio da criação de variáveis utilizando a função de criação de variáveis binárias da biblioteca pandas, na qual para cada nome de categoria de cada atributo, existe uma variável representando essa categoria. Os atributos que possuem essa característica são:

- ESTCIVMAE;
- CONSPRENAT;

Figura 11 – Distribuição de valores faltantes nos dados do modelo

	Missing Ratio
QTDFILMORT	18.218959
QTDPARTNOR	15.601660
QTDPARTCES	13.724864
QTDFILVIVO	7.502926
QTDGESTANT	5.560166
MESPRENAT	0.544739
CONSPRENAT	0.240451
ESTCIVMAE	0.125545
IDANOMAL	0.089371
ESCMAEAGR1	0.019151

Fonte: Produção do próprio autor

- ESCMAEAGR1;
- QTDFILVIVO;
- QTDFILMORT;
- GESTACAO;
- CONSULTAS;
- IDANOMAL;
- TPROBSON;
- PARIDADE;
- SEXO;
- MESPRENAT;
- IDADEMAE_binned.

Já o processo de adequação dos atributos quantitativos foi por meio da utilização da função MinMaxScaler da biblioteca Sklearn, na qual redimensiona os valores dos atributos dentro de uma escala entre 0 e 1 correspondente ao valor mínimo e máximo do atributo. Os atributos que possuem essa característica são:

- SEMAGESTAC;

- QTDGESTANT;
- QTDPARTNOR;
- QTDPARTCES.

O atributo PESO foi transformado em um atributo qualitativo representando 1 para pesos inferiores a 2500g e 0 para valores superiores a essa marca.

3.1.3 Limpeza e adequação dos dados

A criação da RNA foi por meio da biblioteca Keras (KERAS, 2020), na qual permite de maneira rápida criar um modelo com qualquer número de camadas, parâmetros e tipos de função de ativação e realizar diversos testes com mudança nessas características a fim de encontrar o melhor modelo para o projeto em questão.

A construção do modelo foi por meio de testes, em que a estrutura que trouxe o melhor resultado foi um modelo sequencial composto por cinco camadas, conforme demonstrado na figura 12.

Figura 12 – Modelo de RNA

```
model = Sequential()

model.add(Dense(116, input_shape=(116,),activation='relu'))
model.add(Dropout(0.3))

model.add(Dense(58, activation='relu'))
model.add(Dropout(0.3))

model.add(Dense(1, activation='sigmoid'))

model.compile(loss='binary_crossentropy', optimizer='adam', metrics='accuracy')
```

Fonte: Produção do próprio autor

Os parâmetros utilizados foram de três camadas construídas de forma totalmente conectadas, em que nas duas primeiras se utilizam de uma função de ativação de unidade linear retificada (relu) e a última uma função de ativação logística (*sigmoid*). Intercaladas por duas camadas de esquecimento de parcela do treinamento (*droupout*) com valores de 30%.

Os otimizadores utilizados foram escolhidos com base de recomendação da própria documentação da biblioteca, como para perdas o *binary_crossentropy* e o otimizador de estimação adaptativa do momento (*adam*).

Para a aplicação do modelo, o conjunto de dados foi divido de forma aleatória em um conjunto de treino e teste por meio da função *train_test_split* da biblioteca do Sklearn, na qual o conjunto de teste representa 30% do total volume de dados e são separados de forma a evitar a concentração de alguma característica apenas no conjunto de test.

O valor época utilizado foi correspondente a 50, na qual significa que o modelo passou por todos os dados um número de 50 vezes e realizou adequações em seus valores de peso.

Como o conjunto de dados apresenta um desbalanceamento, dois modelos de RNA foram construídos para o treinamento. O primeiro seguiu o treinamento com os dados preparados da forma descrita anteriormente e o segundo tevê um ajuste no conjunto de dados, onde a técnica SMOTE foi utilizada para diminuir a diferença de ocorrência entre as duas condições de peso observadas. A técnica SMOTE (BROWNLEE, 2020) tem como princípio a criação de casos sintéticos para a classe minoritária a fim de crescer o espaço de decisão desta classe.

4 RESULTADOS

4.1 RNA COM VALORES SEM AJUSTE

Durante o treinamento do modelo a acurácia dos valores ficaram com um resultado de 96,4%. E durante o teste o valor de acurácia se caiu para 95,1%, conforme demonstrado nas figuras 13 e 14.

Figura 13 – Treinamento com dados sem ajuste

```
Epoch 45/50
783/783 [==============================] - 1s 2ms/step - loss: 0.1101 - accuracy: 0.9655
Epoch 46/50
783/783 [==============================] - 1s 2ms/step - loss: 0.1083 - accuracy: 0.9672
Epoch 47/50
783/783 [==============================] - 1s 2ms/step - loss: 0.1154 - accuracy: 0.9657
Epoch 48/50
783/783 [==============================] - 1s 2ms/step - loss: 0.1121 - accuracy: 0.9655
Epoch 49/50
783/783 [==============================] - 1s 2ms/step - loss: 0.1077 - accuracy: 0.9662
Epoch 50/50
783/783 [==============================] - 1s 2ms/step - loss: 0.1172 - accuracy: 0.9640
<tensorflow.python.keras.callbacks.History at 0x7fdba460cb90>
```

Fonte: Produção do próprio autor

Figura 14 – Teste com dados sem ajuste

```
336/336 [==============================] - 0s 975us/step - loss: 0.2390 - accuracy: 0.9508
[0.23900866508483887, 0.9508288502693176]
```

Fonte: Produção do próprio autor

Quando observados os valores de precisão, recall e fi-score, encontramos respectivamente os valores de 0.24, 0.04 e 0.06 para o caso de peso considerado baixo peso. O valor de precisão relaciona o número de acertos em relação ao numero total de valores para uma classificação específica, na qual o modelo classificou 18 vezes de maneira correta baixo peso e errou 57 vezes a classificação. O Recall nos diz com que frequencia o classificador determinou a classe correta em relação ao total de representantes daquela classe, no modelo teve um valor de 18 vezes em relação ao total de 489 representates de baixo peso. O fi-score relaciona os dois primeiros valores a fim de trazer um resultado único.

Os resultados desmontram que o classificador não está conseguindo prever de uma forma adequada os valores para baixo peso, valores são apresentados nas figuras 15 e 16

4.2 RNA COM VALORES COM AJUSTE SMOTE

Durante o treinamento do modelo com o ajuste SMOTE nos dados a acurácia dos valores ficaram com um resultado de 93,0%. E durante o teste o valor de acurácia foi de 94,08%, conforme demonstrado nas figuras 17 e 18.

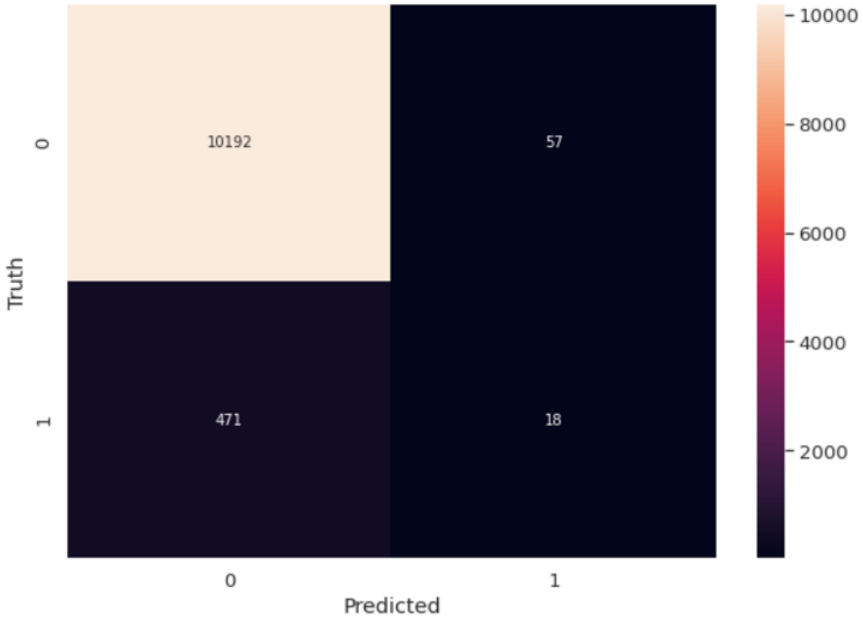
Quando observados os valores de precisão, recall e fi-score, encontramos respectivamente os valores de 0.98, 0.16 e 0.28 para o caso de peso considerado baixo. O valor de precisão

Figura 15 – Reporte de classificação com dados sem ajuste

	precision	recall	f1-score	support
0	0.96	0.99	0.97	10249
1	0.24	0.04	0.06	489
accuracy			0.95	10738
macro avg	0.60	0.52	0.52	10738
weighted avg	0.92	0.95	0.93	10738

Fonte: Produção do próprio autor

Figura 16 – Matriz de confusão modelo sem ajuste



Fonte: Produção do próprio autor

Figura 17 – Treinamento com dados com ajuste SMOTE

```
Epoch 45/50
1712/1712 [==============================] - 3s 2ms/step - loss: 0.1810 - accuracy: 0.9279
Epoch 46/50
1712/1712 [==============================] - 3s 2ms/step - loss: 0.1793 - accuracy: 0.9280
Epoch 47/50
1712/1712 [==============================] - 3s 2ms/step - loss: 0.1836 - accuracy: 0.9273
Epoch 48/50
1712/1712 [==============================] - 3s 2ms/step - loss: 0.1792 - accuracy: 0.9273
Epoch 49/50
1712/1712 [==============================] - 3s 2ms/step - loss: 0.1757 - accuracy: 0.9315
Epoch 50/50
1712/1712 [==============================] - 3s 2ms/step - loss: 0.1779 - accuracy: 0.9303
<tensorflow.python.keras.callbacks.History at 0x7fdba3cd6bd0>
```

Fonte: Produção do próprio autor

teve uma melhora passando para 1111 acertos de baixo baixo peso e 17 erros. O Recall passou a representar 16%. O que significa dizer que o modelo passou a prever de uma forma mais adequada os valores para baixo peso, porém ainda longe do ideal para atender de forma assertiva a população, já o número de falso negativos de 5736 classificados como baixo peso continua relativamente elevado. Os valores são apresentados nas figuras 19 e 20.

Figura 18 – Teste com dados com ajuste SMOTE

```
428/428 [==============================] - 1s 965us/step - loss: 0.1524 - accuracy: 0.9408
[0.15239855647087097, 0.9407726526260376]
```

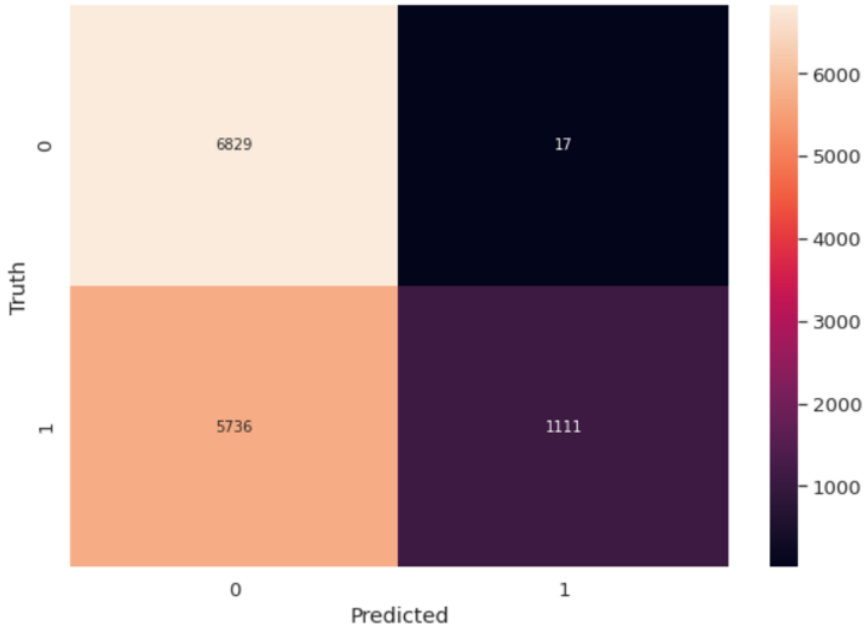
Fonte: Produção do próprio autor

Figura 19 – Reporte de classificação com dados com ajuste SMOTE

	precision	recall	f1-score	support
0	0.54	1.00	0.70	6846
1	0.98	0.16	0.28	6847
accuracy			0.58	13693
macro avg	0.76	0.58	0.49	13693
weighted avg	0.76	0.58	0.49	13693

Fonte: Produção do próprio autor

Figura 20 – Matriz de confusão com dados com ajuste SMOTE



Fonte: Produção do próprio autor

5 CONCLUSÃO E TRABALHOS FUTUROS

5.1 CONCLUSÃO

O trabalho teve como objetivo aplicar a técnica de machine learning de redes neurais artificiais na análise de dados relacionados a recém nascidos fornecidos pelo SINASC, sistema de responsabilidade do SUS. Apesar de os resultados provenientes dos modelos não serem satisfatórios para o emprego em ferramentas nos hospitais e centros de atendimentos de saúde, eles demonstram uma correlação entre os dados e o peso do recém-nascido o que possibilita melhorias nos modelos e melhorias nos dados coletados hoje pelo SUS a fim de possibilitar um produto que possa diminuir o número de BPN na população a partir do acompanhamento dessas informações durante todo o processo de pré-natal da mãe. A correlação fica evidente quando aplicamos a técnica SMOTE, pois o número de representantes de baixo peso aumenta e todos os indicadores demonstram uma um aumento. Com os resultados obtidos, temos um possível caminho para o desenvolvimento de um produto de mercado, o que traz como objetivo cumprido ao trabalho.

Todo o código e o dataset utilizado estão disponíveis para visualização no link no Google Drive [1] .

5.2 TRABALHOS FUTUROS

Com o desenvolvimento deste estudo e dos resultados obtidos pode-se perceber a oportunidade para o desenvolvimento de trabalhos futuros, entre ele:

- Ampliar o número de cidades abordadas como exemplos de dados. Para ter melhor generalização, balanceamento das classes e melhores classificadores;
- Obter mais variáveis para o conjunto de dados, entre elas variáveis relacionadas a situação geoeconômica da população;
- Aplicar outras técnicas para balancear os dados das classes para o aprendizado dos algoritmos, entre elas o ajuste de pesos e diminuir o tamanho do conjunto de treino para apenas apresentar dados obtidos pelo sistema do SINASC e com mesmo valor de classes;
- Tratar os dados eliminando os grupos que não representam um valor para o conjunto de dados, como por exemplo retirar as categorias categorizadas como ignorado.

[1] Código e documentação: https://drive.google.com/drive/folders/1rYybRIiftbr38L5BjpahiwNkhUnzmPT3?usp=sharing

REFERÊNCIAS

ACADEMY, D. S. **Top 20 bibliotecas Python para data science**. 2018. Disponível em: https://datascienceacademy.com.br/blog/top-20-bibliotecas-python-para-data-science. Acesso em: 20 fev. 2021.

AZURE. **Azure machine learning:** serviço de aprendizado de máquina empresarial para criar e implantar modelos rapidamente. 2021. Disponível em: https://azure.microsoft.com/pt-br/services/machine-learning. Acesso em: 20 fev. 2021.

BROWNLEE, J. **Smote for imbalanced classification with python**. 2020. Disponível em: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification. Acesso em: 20 fev. 2021.

DATASUS. **Portal da Saúde:** Datasus. 2020. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=28&pad=31655. Acesso em: 20 fev. 2020.

FACELI, K. *et al.* **Inteligência artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

IBM. **IBM Watson Studio:** Simplifique e escale a ciência de dados para prever e otimizar seus resultados de negócios. 2021. Disponível em: https://www.ibm.com/br-pt/cloud/watson-studio. Acesso em: 20 de fev. 2021.

KERAS. The Python deep learning libray. 2020. Disponível em: http://keras.io/. Acesso em: 25 out. 2020.

MITCHEL, T. M. **Machine learning**. Nova Iorque: McGraw-Hill, 1997.

MOREIRA, A. I. M.; SOUSA, P. R. M. d; SARNO, F. Baixo peso ao nascer e seus fatores associados. **Einstein**, v. 16, n. 4, 2018.

OPAS. **Um em cada sete bebês em todo o mundo nascem com baixo peso**. 2019. Disponível em: https://www.paho.org/bra/index.php?option=com_content&view=article&id=5935:um-em-cada-sete-bebes-em-todo-o-mundo-nascem-com-baixo-peso&Itemid=820. Acesso em: 15 fev. 2021.

RUSSEL, S.; PETER, N. **Inteligência artificial.** Rio de Janeiro: Elsevier, 2013.

Ministério da Saúde. **SUS:** 27 anos transformando a história da saúde no Brasil. 2015. Disponível em: http://www.blog.saude.gov.br/rh146e. Acesso em: 15 fev. 2021.

TOURINHO, A. B.; REIS, M. L. B. S. Peso ao nascer: uma abordagem nutricional. **Comunicação em Ciência da Saúde**, Brasília, v.23, n. 1, p. 19-30, ago. 2012. Disponível em: http://bvsms.saude.gov.br/bvs/periodicos/revista_ESCS_v23_n1_a02_peso_ao_nascer.pdf. Acesso em: 19 fev. 2021.

WHO. **Global nutrition targets 2025 low birth weight policy brief**. 2014. Disponível em: https://www.who.int/nutrition/publications/globaltargets2025_policybrief_lbw/en/#:~:text=The\%20goal\%20is\%20to\%20achieve,with\%20weight\%20at\%20birth. Acesso em: 10 fev. 2021.