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

Trabalho de Conclusão de Curso

DEEP LEARNING COM MODELOS BASEADOS EM LSTM PARA PREVISÃO DE CASOS DE SÍFILIS ADQUIRIDA NO BRASIL

Gabriel Vinícius Souza da Silva

Natal/RN

2022

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA BIOMÉDICA

DEEP LEARNING COM MODELOS BASEADOS EM LSTM PARA PREVISÃO DE CASOS DE SÍFILIS ADQUIRIDA NO BRASIL

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

Graduando: Gabriel Vinícius Souza da Silva

Orientador: Prof. Dr. Ignacio Sanchez Gendriz,

Coorientador: Ernano Arrais Junior

Natal/RN

2022

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA BIOMÉDICA

DEEP LEARNING COM MODELOS BASEADOS EM LSTM PARA PREVISÃO DE CASOS DE SÍFILIS ADQUIRIDA NO BRASIL

Banca Examinadora do Trabalho de Conclusão de Curso:

Prof. Dr. Ignacio Sánchez Gendriz UFRN - Orientador	_____
Prof. Ernano Arrais Junior UFRN Coorientador	_____
Prof. Dr. Ricardo Alexsandro Valentim LAIS/UFRN – Avaliador interno	_____
Prof. Dr(a) Daniele Montenegro da Silva Barros LAIS/UFRN – Avaliadora externa	_____

Natal/RN

2022

DEDICATÓRIA

Dedico este trabalho à minha família e amigos, que ao longo de toda a trajetória, sempre encontrei apoio e incentivo para realizá-lo.

AGRADECIMENTOS

Aos meus pais, Simone e Iranildo, que investiram na minha educação e me incentivaram, sendo meus pilares durante toda a graduação. O meu irmão Rafael o qual hoje sirvo de orientação e inspiração, aos meus avós e a minha namorada Sthefany que me deram coragem e apoio. A todos por compreenderem minha ausência durante o desenvolvimento deste trabalho.

O meu orientador e amigo professor Ignácio, que antes de tudo é um colega de pesquisa e trabalho. Agradeço pelos ensinamentos que contribuíram na minha formação, pelo empenho e dedicação nas orientações para elaboração deste trabalho, permitindo um bom desenvolvimento em conjunto com as minhas ideias.

Ao Laboratório de Inovação Tecnológica em Saúde (LAIS) da UFRN e a todos que fazem parte dele pela oportunidade. Este trabalho é fruto das pesquisas desenvolvidas pelo LAIS. Foi a minha primeira experiência profissional, os conhecimentos e a vivência adquirida foram de grande importância para me engajar no mercado de trabalho.

Aos professores e amigos que fiz durante o primeiro ciclo de CET e na Engenharia Biomédica, por todos os momentos bons e ruins que compartilhamos. Vocês foram essenciais no meu amadurecimento e formação acadêmica.

SUMÁRIO

LISTA DE FIGURAS viii

LISTA DE TABELAS ix

LISTA DE ABREVIAÇÕES, SIGLAS E SÍMBOLOS x

RESUMO xii

ABSTRACT xiii

1. INTRODUÇÃO 1

OBJETIVOS 3

1.1.1. OBJETIVO GERAL 3

1.1.2. OBJETIVOS ESPECÍFICOS 3

2. ESTADO DA ARTE 4

3. METODOLOGIA 6

3.1. BASE DE DADOS 6

3.2. PRÉ-PROCESSAMENTO 6

3.3. ANÁLISE EXPLORATÓRIA 7

3.4 REDE NEURAL LSTM 8

3.5. ARQUITETURA 11

3.5.1. PARÂMETROS DE DESEMPENHO 11

3.5.2. PARÂMETROS DA REDE 11

3.5.3. DIVISÃO DE TREINO, TESTE E VALIDAÇÃO 12

4. RESULTADOS E DISCUSSÕES 14

4.1. ANÁLISE DE DADOS 14

4.2 MODELAGEM 18

4.2.1 MODELO DIÁRIO 18

4.2.2. MODELO SEMANAL 19

5. CONCLUSÕES 21

6. REFERÊNCIAS 22

LISTA DE FIGURAS

Figura 1 – Números de casos de sífilis adquirida de 2010 a 2021…….....2
Figura 2 – Distribuição dos valores do número de casos na escala original e escalonado de zero a um……7
Figura 3 – Estrutura Rede Neural Recorrente……...8
Figura 4 – Termo de memória de longo prazo……...9
Figura 5 - Fluxo de funcionamento do LSTM……..10
Figura 6 – Estratégia de divisão de treino, teste e validação……13
Figura 7 – Taxa de detecção de sífilis adquirida (por 100.000 habitantes) no Brasil segundo a faixa etária, de 2010 a 2019, por semana e por ano…….14
Figura 8 – Taxa de detecção de sífilis adquirida (por 100.000 habitantes) no Brasil segundo a cor, de 2010 a 2019, por semana e por ano……..15
Figura 9 – Quantidade de casos de sífilis adquirida no Brasil segundo o Sexo, de 2010 a 2019, por semana e por ano……15
Figura 10 – Taxa de detecção de sífilis adquirida (por 100.000 habitantes) no Brasil segundo a escolaridade, de 2010 a 2019……16
Figura 11 – Mapa de calor da taxa de detecção de sífilis adquirida (por 100.000 habitantes) no Brasil normalizada por ano, de 2010 a 2019, por semana por ano……...17
Figura 12 – Comparação do número real de casos e das previsões para o modelo diário……...18
Figura 13 – RMSE para a previsão de cada um dos dias…….19
Figura 14 – Comparação do número real de casos e das previsões para o modelo semanal……20

LISTA DE TABELAS

Tabela 1 –Estrutura das Redes Neurais.....12
Tabela 2 – Parâmetros do modelo.....13
Tabela 3 – Tamanhos das bases de treino, teste e validação.....14

LISTA DE ABREVIAÇÕES, SIGLAS E SÍMBOLOS

CNN – Redes Neurais Convolucionais

DB-LSTM – *Bi-directional* LSTM

DOD – Decisão Orientada a Dados

ED-LSTM – *encoder-decoder* LSTM

IBGE – Instituto Brasileiro de Geografia e Pesquisa

IST – Infecção Sexualmente Transmissível

LR – *Learning Rate*

LSTM - *Long short-term memory*

MAE - *Mean Absolute Error*

ML – *Machine Learning*

MS – Ministério da Saúde

OMS – Organização Mundial da Saúde

RMSE – *Root Mean Square Error*

RMSprop – *Root Mean Squared Propagation*

RNN – Redes Neurais Recorrentes

SINAN – Infecção Sexualmente Transmissível

TI – Tecnologia da Informação

SILVA, Gabriel Vinícius. **Deep Learning com modelos baseados em LSTM para previsão de casos de Sífilis Adquirida no Brasil.** Trabalho de Conclusão de Curso, Graduação em Engenharia Biomédica, Universidade Federal do Rio Grande do Norte, 36p., 2022.

RESUMO

A sífilis adquirida é uma infecção sexualmente transmissível e um grave problema de saúde pública. A pandemia da Covid-19 mostrou a importância do monitoramento dos casos para apoiar as decisões e ações de combate a proliferação da doença através dos dados, e para a sífilis não é diferente. Com isso vários trabalhos foram desenvolvidos com o objetivo de prever essa quantidade de casos. Portanto, esse trabalho propõe dois modelos de *Deep Learning*, para a previsão do número de casos de sífilis adquirida no Brasil, um de previsão diária e outro semanal, como também a análise exploratória dos dados. Os dados foram analisados sob as diferentes visões presentes no Dataset, e mostrados através de gráficos variados ao longo do tempo. Foram desenvolvidas redes neurais baseadas em LSTM (do inglês *Long short-term memory*) para construir os modelos preditivos, estes mostram boa capacidade de prever a tendência do número de casos de sífilis. O modelo semanal apresentou um RMSE (do inglês *Root Mean Square Error*) de 450,03 e uma correlação de 0,72, já o diário apresentou um erro médio de 86,96 e uma correlação média de 0,90.

Palavras-chave: Sífilis adquirida. Deep Learning. LSTM.

SILVA, Gabriel Vinícius. **Deep Learning with LSTM-based models to predict Acquired Syphilis casees in Brasil** Conclusion Work Project, Biomedical Enginnering Bachelor Degree, Federal University of Rio Grande do Norte, 36p., 2022.

ABSTRACT

Acquired syphilis is a sexually transmitted infection and a serious public health problem. The Covid-19 pandemic showed the importance of monitoring cases to support decisions and actions to combat the spread of the disease through data, and syphilis is no different. Several works have been proposed in the literature to predict the evolution of the number of cases for different diseases. This work proposes two Deep Learning (DL) models to predict the number of cases of acquired syphilis in Brazil. The first DL model deals with daily case predictions and the second deals with weekly case predictions. The paper also presents an exploratory data analysis with the aim of contributing to a better understanding of the dynamics of syphilis cases in Brazil. Neural networks based on LSTM (Long short-term memory) were developed to build the predictive DL models, which show good ability to predict the trend of the number of syphilis cases. The weekly model presented an RMSE (Root Mean Square Error) of 450,03 and a correlation of 0,72, while the daily model presented an average RMSE of 86,96 and an average correlation of 0,90.

Key-words: Acquired syphilis. Deep Learning. LSTM

1. INTRODUÇÃO

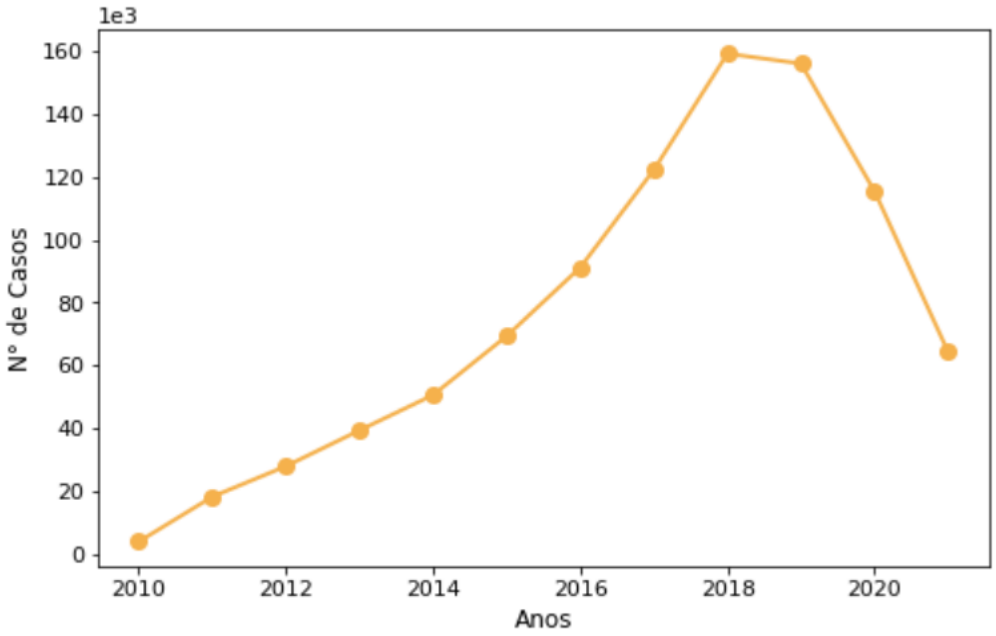
A sífilis é uma infecção de contato causada pela bactéria Treponema pallidum que pode ser caracterizada em três tipos diferentes. A sífilis adquirida pode ser transmitida sexualmente ou por transfusão de sangue, a sífilis congênita é transmitida da mãe para o feto durante a gestação, e a sífilis em gestantes. Essa infecção é um sério problema de saúde pública no contexto mundial (Stamm, 2013). Segundo a Organização Mundial da Saúde (OMS) são mais de 1 milhão de novos casos de Infecções sexualmente transmissíveis (ISTs) curáveis todos os dias (incluindo a sífilis), um cenário preocupante visto que aumenta o risco de infecção por HIV.

O reaparecimento dos casos de sífilis se iniciou com aumentos significativos na década de 1980, nos anos 2000 surgiram vários surtos isolados em todo o mundo frequentemente associados à coinfecção do HIV (Ruiz-Sancho, 2007). A OMS estima que 11 milhões de novos casos de sífilis acontecem anualmente no planeta em adultos de 15 a 49 anos.

No Brasil em 2016, o governo federal por meio do Ministério da saúde (MS), decretou epidemia de sífilis no país juntamente com um plano de ação para combater a infecção. A fim de monitorar e registrar com mais precisão os casos, a notificação compulsória foi decretada para as três formas da infecção, pelas seguintes portarias: Portaria GM nº 542 de 22 de dezembro de 1986 (sífilis congênita), portaria GM nº 33 de 14 de julho de 2005 (sífilis em gestantes), portaria nº 2.472 publicada em 31 de agosto de 2010 (sífilis adquirida).

Segundo o boletim epidemiológico da sífilis (BRASIL, 2021), divulgado pelo MS em outubro de 2021, houve um crescimento do número de casos da sífilis adquirida de 2010 a 2018 com um decaimento em 2019, que foi acentuado em 2020 e 2021 (Figura 1).

Figura 1 –Números de casos de sífilis adquirida de 2010 a 2021



Fonte: Adaptado do Boletim Epidemiológico 2021.

Com a pandemia da Covid-19 foi reforçado a importância do registro e acompanhamento dos dados para apoiar e embasar as estratégias de contingência e controle da disseminação da doença, diminuindo os grandes fluxos de pessoas através de restrição do funcionamento de vários estabelecimentos, confinamento (Lockdown Haider, 2020) e os protocolos de segurança sanitária. Segundo Fawcett 2013 a Decisão Orientada a Dados (DOD), refere-se a prática de basear as decisões em análise de dados, ao invés da intuição. Segundo o autor, a partir dos dados é possível utilizar a Tecnologia da Informação (TI), para encontrar atributos descritivos da entidade de interesse e aplicar o DOD. No contexto deste trabalho a entidade de interesse são os casos de sífilis adquirida no Brasil.

Uma das ferramentas que possibilita o DOD é o *Machine Learning* ou Aprendizado de Máquina (ML). Segundo Aurlien (2017), o ML é a ciência da programação para que os computadores possam produzir conhecimento através dos dados. Segundo Goodfellow (2016) é possível criar soluções que permitem que os computadores aprendam com experiência, entendendo o mundo a partir de uma hierarquia de conceitos. O autor fala que o computador formula conhecimentos mais complexos a partir dos conceitos mais simples, chegando ao nível mais profundo denominado de Aprendizado Profundo (*Deep Learning)*, que está dentro do ML.

Devido ao contexto mundial da pandemia de Covid-19 e ao aumento no número de casos de sífilis adquirida no país na última década (Figura 1), evidenciam a importância do presente estudo. Além disso, este trabalho está inserido no contexto do Projeto Sífilis Não desenvolvido pelo Laboratório de Inovação Tecnológica em Saúde (LAIS) da Universidade Federal do Rio Grande do Norte (UFRN) que vem mudando o cenário da doença no Brasil (Governo Federal, 2021). Portanto o presente estudo busca explorar os dados da sífilis adquirida, utilizando *Deep Learning*, com o objetivo de auxiliar nas tomadas de decisões estratégicas para controle e contingência da infecção, a partir do acompanhamento e previsão das tendências no número de casos.

OBJETIVOS

1.1.1. OBJETIVO GERAL

Desenvolver um modelo de *Deep Learning* para a previsão do número de casos de sífilis adquirida no Brasil, prevendo as tendências de aumento ou diminuição dos casos.

1.1.2. OBJETIVOS ESPECÍFICOS

1. Coletar, tratar e transformar os dados de sífilis adquirida no Brasil em formato de série temporal, através de programação.
2. Estudar e aplicar algoritmos de *Deep Learning* para previsão de séries temporais, decorrentes dos casos de sífilis adquirida no Brasil.
3. Realizar uma análise exploratória dos dados para construir gráficos que ajudem a entender o comportamento da infecção sob várias perspectivas.
4. Estudar as áreas relacionadas a área da saúde em que foi aplicado o trabalho - Infectologia e Medicina.
5. Avaliar os resultados obtidos, analisando a desempenho do algoritmo.

2. ESTADO DA ARTE

Ibáñez-Cervantes (2020), utiliza uma variação do modelo Auto-Regressivo Integrado de Médias móveis (ARIMA) que permite a modelagem direta da componente sazonal da série temporal, o modelo Sazonal Auto-Regressivo Integrado de Médias móveis (SARIMA). Através desse algoritmo é feita a previsão do número de casos de sífilis adquirida nos anos de 2018 e 2019 no México. O autor utiliza o resultado do modelo para propor possíveis medidas de contenção.

Weng (2020), também utiliza o SARIMA para a previsão dos casos de outra IST, a clamídia trachomatis (TC), na cidade Shenzhen da China. O autor mostra como o modelo foi capaz de identificar bem a sazonalidade e as tendências com uma análise mensal dos resultados. Além disso, propõe possíveis medidas para o controle das infecções, como campanhas de prevenção para grupos populacionais específicos.

Com a pandemia da Covid-19 modelos Deep Learning foram amplamente utilizados para previsão de número de casos, com diversas estratégias de treinamento e arquiteturas de rede diferentes, como Gautam (2021) que utilizou o LSTM e transferência de aprendizado, para aumentar a capacidade de generalização do modelo. Para a transferência de aprendizado, utilizou um modelo previamente treinado em dados de países como Itália e Estados Unidos para previsão nos países da Alemanha, França, Brasil, Índia e Nepal. O pré-treinamento permite que a rede aprenda em uma maior variabilidade de casos, o que melhora seu desempenho.

Sanchez-Gendriz (2021), também utiliza Deep Learning e LSTM para prever a incidência de dengue na cidade brasileira de Natal (Rio Grande do Norte). Para isso o modelo se baseia na quantidade de ovos em diferentes locais das cidades, esses locais possuem focos controlados. Esses pontos com características favoráveis à deposição dos ovos do mosquito da dengue são chamados de ovitrampas. O autor se baseia nos registros anteriores e nas ovitrampas para prever a incidência da dengue. Com o isso o autor propôs um modelo para previsão antecipada da incidência da doença, correlacionando-a com dados socioeconômicos dos Bairros.

Zain (2021) propõe uma abordagem híbrida, combinando duas arquiteturas, a Rede Neural Convolucional (CNN) com o LSTM para realizar a previsão dos casos de Covid-19. O autor compara o desempenho dessa abordagem com outros 17 modelos, que servem como linha de base, além disso mostra que a combinação dessas arquiteturas de Rede Neural supera o desempenho de cada uma individualmente.

Chandra (2021), utiliza diferentes variações na arquitetura de modelos baseados em LSTM para a previsão dos casos de Covid-19 na Índia. São comparados três métodos: o LSTM padrão, o *Bi-directional* LSTM (BD-LSTM) e o *encoder-decoder* LSTM (ED-LSTMM). Os resultados mostraram que os modelos variam o desempenho para diferentes localidades geográficas.

3. METODOLOGIA

3.1. BASE DE DADOS

Os dados consistem nas notificações anonimizadas dos casos de Sífilis congênita no Brasil de 2010 a 2019, que foram retirados da base de dados epidemiológicos do Sistema de Informações de Agravos de Notificações (SINAN, 2016). Este é um sistema ligado ao MS que monitora e atualiza as notificações. A base de dados de notificações de sífilis adquirida contém informações categóricas como sexo, idade, raça ou cor e faixa etária. Além de informações geográficas como a localidade da Unidade de Saúde onde ela foi feita bem como a localidade de residência da pessoa notificada.

Entre os vários tipos de dados estavam o resultado do exame para detecção da infecção (positivo ou negativo) e a data do diagnostico, esses foram utilizados para a construção da série temporal do número de casos diário de sífilis adquirida no Brasil.

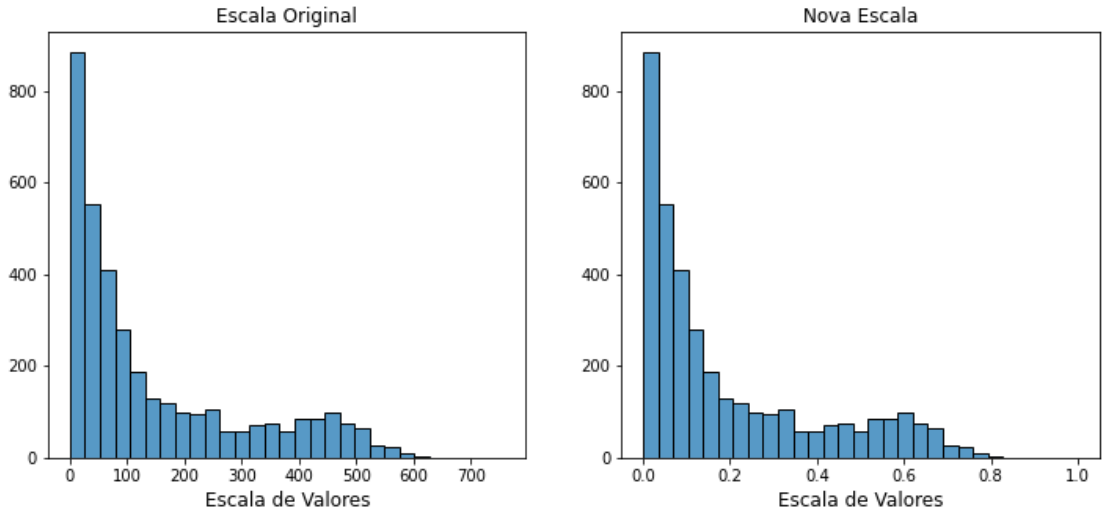
3.2. PRÉ-PROCESSAMENTO

Essa etapa consiste no conjunto de operações necessárias para transformar os dados brutos de tal forma que possa ser entendido pelo modelo de Deep Learning. Como descrito na seção 3.1, os dados estavam organizados por registro de notificação de um possível caso de sífilis adquirida. Para a formulação da série temporal de interesse foram utilizadas as colunas da data do diagnóstico (com frequência diária), e o resultado do diagnóstico. Com essas informações foi construída a série temporal do número casos da infecção no Brasil, a partir da soma do número de diagnósticos positivos agrupados por dia.

Os algoritmos de Deep Learning e Machine Learning em geral criam relações a partir de operações matemáticas entre os números (Goodfellow, 2016), portanto, conjuntos de dados com valores em escalas diferentes podem levar os modelos a atribuírem maior importância às informações trazidas pelos maiores números.

Adicionalmente, os modelos que utilizam a descida do gradiente para atualizar seus pesos (aprendizado), como as redes neurais em *Deep Learning*, convergem mais rápido em dados com a mesma escala (Aurlien, 2017). Portanto os dados foram reescalonados para a faixa de zero a um. Como observado na Figura 2, o escalonamento não modifica a distribuição.

Figura 2 – Distribuição dos valores do número de casos na escala original e escalonado de zero a um.



Fonte: Autoria própria.

Além do escalonamento, um dos modelos utiliza o filtro de média móvel (Sanchez-Gendriz, 2021) nos dados. Esse filtro vai ajudar o modelo a compreender melhor a tendência da curva, além disso, tende a diluir melhor ao agrupar o conjunto de valores dentro da janela (Eq 1).

$$y[n] = \frac{1}{N+1}\sum_{i=0}^{N} x[n-i] \tag{1}$$

3.3. ANÁLISE EXPLORATÓRIA

Segundo Knaflic (2015), existe uma diferença entre análise exploratória e explanatória. A análise exploratória consiste na compreensão dos dados por meio do esforço de analisar e testar hipóteses para todo o conjunto de dados, já a análise explanatória representa o processo de modelar e transformar os dados em informação que podem ser consumidas e comunicam sobre um tema, contando uma história.

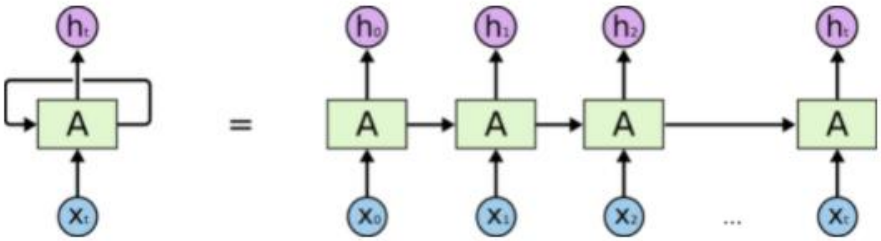
Foi realizado uma renomeação das categorias dos dados brutos de acordo com as categorias do dicionário de dados (SINAN, 2018) em conjunto com o Boletim Epidemiológico, facilitando a visualização das variáveis categóricas presentes nos dados.

Para a construção das taxas de detecção da IST foram utilizados os dados demográficos do Instituto Brasileiro de Geografia e Estatística (IBGE). Os gráficos construídos retratam a evolução do número de casos ao longo do tempo, para as diferentes categorias das pessoas infectadas como, escolaridade, sexo, entre outros.

3.4 REDE NEURAL LSTM

As redes neurais tradicionais apresentam ótimos desempenhos em várias aplicações para as mais diversas áreas, porém para essas arquiteturas a informação não é persistida (memorizada) para que possam ser utilizadas em previsões futuras. As Redes Neurais Recorrentes (RNN) são capazes de resolver esse problema aplicando *Loops* (Figura 3) aos neurônios das camadas. Permitindo que a rede "memorize" as inferências nos dados (Connor, 1994).

Figura 3 –Estrutura Rede Neural Recorrente



Fonte: Retirada de Olah (2015).

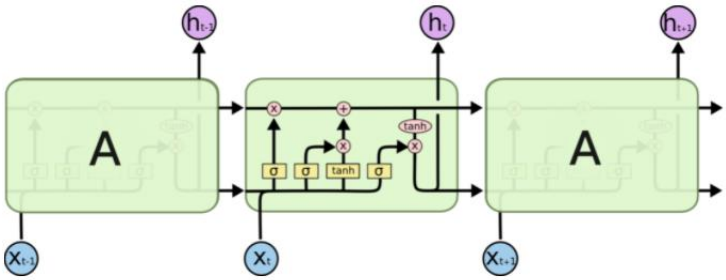
A RNN possui um problema com as dependências de longo prazo. Entende-se por dependência de longo prazo a necessidade de usar relações temporais longas. Isso aumenta o custo computacional e pode introduzir informações irrelevantes, com isso as RNNs perdem a capacidade de generalizar e aprender com os dados (Bengio, 1994).

Além das dependências de longo prazo, as RNNs sofrem com os problemas de desaparecimento e explosão do gradiente. Esses problemas estão relacionados ao aumento nas interações e atualização dos pesos da rede causados pelos loops da RNN. Isso pode fazer com que o gradiente diminua ou aumente demais. Esses problemas foram resolvidos com um tipo especial de RNN, a LSTM, adicionando um novo estado de memória, introduzido por Hochreiter em 1997.

Como observado na Figura 4, o LSTM apresenta a característica de encadeamento de uma RNN, no entanto, com quatro camadas que interagem de tal forma a selecionar as

informações mais relevantes a serem memorizadas em cada loop através das células de memória. A alteração dessas informações muda o estado da célula, essa estrutura permite que o LSTM supere o problema das dependências de longo prazo.

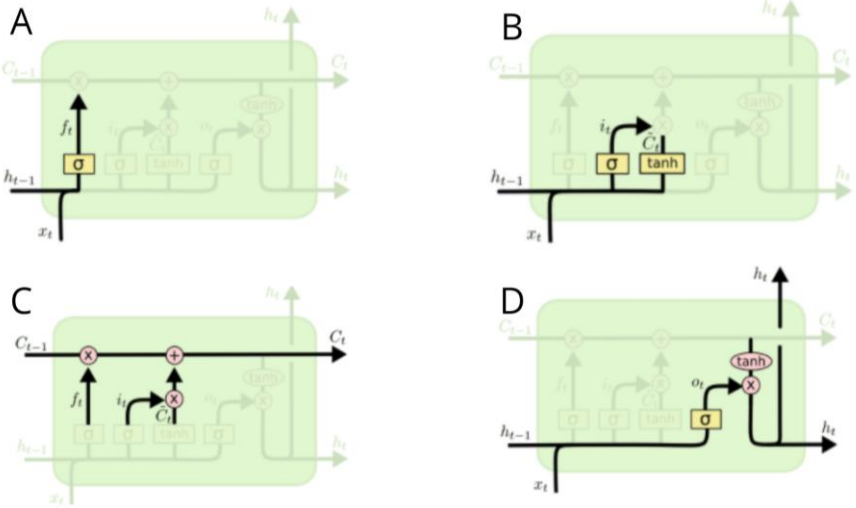
Figura 4 – Termo de memória de longo prazo



Fonte: Retirada de Olah (2015).

A Figura 5 esquematiza como o LSTM seleciona e atualiza o estado da célula com as informações mais relevantes. Os σ representam as camadas da função sigmóide, um fator multiplicativo que varia de zero a um, decidindo o quanto de informação será levada a diante. Primeiramente são selecionadas as informações que serão descartadas do estado da célula, essa decisão é tomada a partir da aplicação do estado anterior da célula ***ht-1*** e a entrada atual ***Xt.*** Esse processo é realizado pelo "forget gate" (Figura 5.A), gerando a saída ***ft*** (Eq 2, Olah 2015).

Figura 5 – Fluxo de Funcionamento do LSTM



Fonte: Retirada de Olah (2015).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right) \tag{2}$$

O segundo passo é decidir quais informações serão armazenadas ao estado da célula. Esse processo acontece em duas etapas: primeiro de forma similar ao passo anterior, uma camada sigmóide chamada de "input gate" coleta os valores que serão atualizados. Em seguida uma camada de tangente hiperbólica (***tanh***), com valores que variam de -1 a 1, decidem as novas entradas. Essas duas camadas geram os vetores de entrada ***it*** e ***Ĉt*** respectivamente Eqs. 3, 4 (Figura 5 B).

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C) \tag{4}$$

Fonte: retirada de Olah (2015).

Sendo assim, o estado da célula pode ser atualizado. Para isso é combinado o **ft** e ***Ct-1***, que compõem as informações que serão esquecidas, com as novas informações ***it*** e ***Ĉt***, criando o novo estado da célula ***Ct*** (Figura 5 C) calculado pela Eq 5 (Olah, 2015). Por fim, é produzida a saída do LSTM, que será uma versão filtrada do estado final da célula. O filtro ***Ot*** (Eq 6, Olah 2015) consiste em uma camada sigmóide que decide quais partes do ***Ct*** são mantidas. Então a saída ***ht*** (Eq 7, Olah 2015) é obtida através da multiplicação do filtro ***Ot*** pela ***tanh*** do estado atual.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right) \tag{6}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{7}$$

3.5. ARQUITETURA

3.5.1. PARÂMETROS DE DESEMPENHO

O Erro Médio Absoluto (MAE), e a Raíz do Erro Médio Absoluto (RMSE), demonstrados nas Eqs. 8 e 9 (Gautam, 2015), foram as duas métricas utilizadas para avaliar os resultados do modelo proposto.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - y_{forecast(t)}\right| \quad (8)$$

$$RMSE = \sqrt{\sum_{t=1}^{n}\frac{\left(y_t - y_{forecast(t)}\right)^2}{n}} \quad (9)$$

3.5.2. PARÂMETROS DA REDE

Foram desenvolvidos modelos com arquiteturas similares, porém com frequências temporais diferentes, sendo um semanal e outro diário. Para isso o número de casos foi agrupado de acordo com a frequência do seu respectivo modelo.

A estrutura das redes são correspondentes, sendo compostas por uma entrada com o LSTM, seguida de uma camada escondida (Hidden-Layer). A diferença entre os modelos está na saída (*Output*) e na Hidden - Layer conforme a Tabela 1. O modelo diário tem como output os próximos 7 dias, já o semanal a semana seguinte.

Tabela 1 –Estrutura das Redes Neurais

Frequência	Método	window	Input	Hidden Layer	Output
Diária	LSTM	20	(1,50)	50	7
Semanal	LSTM	3	(1,256)	256	1

Fonte: Autoria própria.

A janela do modelo diário é de tamanho 20, dessa forma o modelo utiliza o número de casos dos 20 dias anteriores para prever o atual, já o modelo semanal utiliza as 3 semanas anteriores para prever a atual. A LR (do inglês *Learning Rate*) inicia com o valor presente na Tabela 2, porém é atualizada conforme a inércia do aprendizado, ou seja, caso a rede passe por muitas iterações sem diminuir o erro, a LR é reduzida. Isso ajuda a rede a não ficar presa em mínimos locais e acelera o aprendizado. A função e custo (do inglês Loss Function) foi o RMSE (seção 3.5.1).

O RMSProp (do inglês *Root Mean Squared Propagation*) é uma extensão do algoritmo clássico de descida do gradiente. Este otimizador utiliza uma média móvel dos gradientes parciais para atualizar os pesos, isso induz o modelo a se concentrar nos gradientes mais atuais que diminuam o erro. Em conjunto com a arquiteturas da RNN, permite maior desemmpenho na seleção de informações para o aprendizado da rede.

Tabela 2 – Parâmetros do modelo

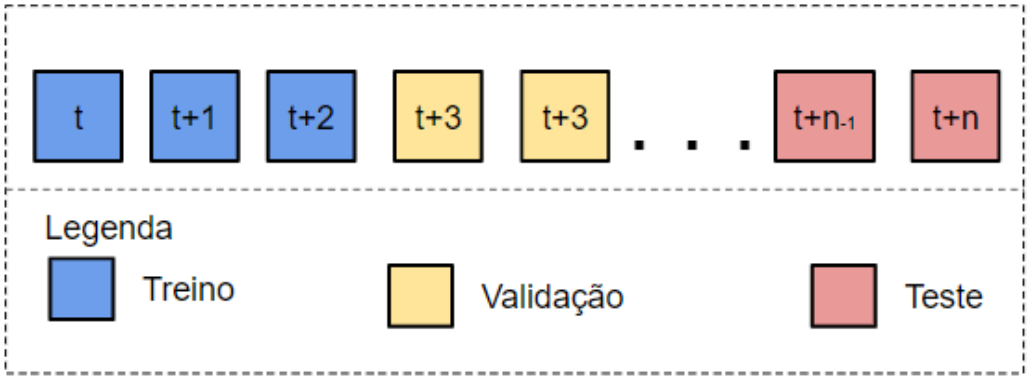
Batch Size	Optimizer	Momentum	Epochs	Função Custo	LR
16	RMSprop	0.9	150	RMSE	10^{-4}

Fonte: Autoria própria.

3.5.3. DIVISÃO DE TREINO, TESTE E VALIDAÇÃO

Os conjuntos de treino, validação e teste foram separados conforme a Figura 6. Essa forma de separação garante que não aconteça vazamento temporal, sendo assim, bons resultados no teste são indicativos confiáveis da boa capacidade de generalização do modelo.

Figura 6 – Estratégia de divisão de treino, teste e validação



Fonte: Autoria própria.

O tamanho de cada conjunto é mostrado na Tabela 3. No diário essas porcentagens representam 180 dias. Dessa forma, o conjunto de teste corresponde aos últimos 180 dias de 2019, como a janela desse modelo é de tamanho 20 (Tabela 1) ficam 130 dias como previsão de teste. O conjunto de dados a nível semanal é sete vezes menor que o diário, por isso, foram separadas menores porcentagens para teste e validação. Além disso, se estes conjuntos forem muito grandes (principalmente o semanal) o teste acaba ficando muito distante temporalmente do treino.

Tabela 3 – Tamanhos das bases de treino, teste e validação

Modelo	Treino (%)	Teste (%)	Validação (%)
Diário	90	5	5
Semanal	92	6	2

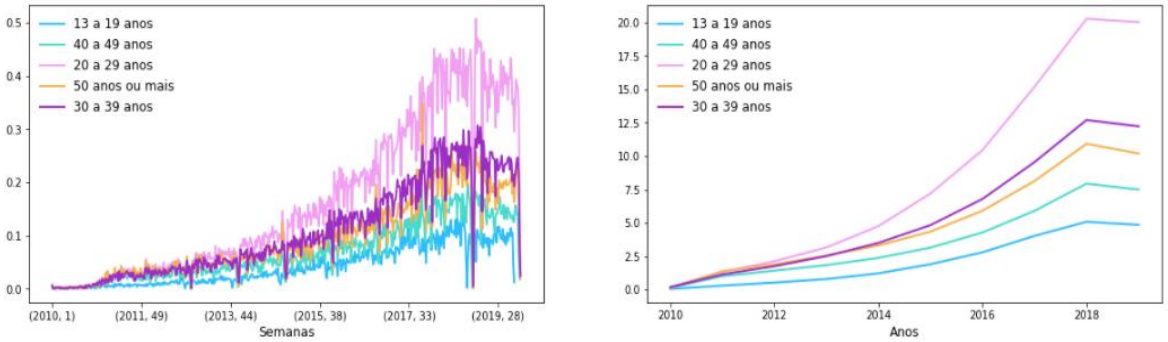
Fonte: Autoria própria.

4. RESULTADOS E DISCUSSÕES

4.1. ANÁLISE DE DADOS

Os dados consistem nas notificações de casos de sífilis congênita em todo o país ao longo da última década. Por se tratar de um evento no tempo, as análises desenvolvidas se baseiam no acompanhamento gráfico da série temporal, sob diferentes visões presentes nos dados, em diferentes frequências de tempo. Como na Figura 7, é verificado a evolução da taxa de detecção por faixa etária por semana e ano, o gráfico mostra uma tendência de crescimento para todas as faixas etárias, porém a maior quantidade se concentra nas pessoas de 20 a 29 anos.

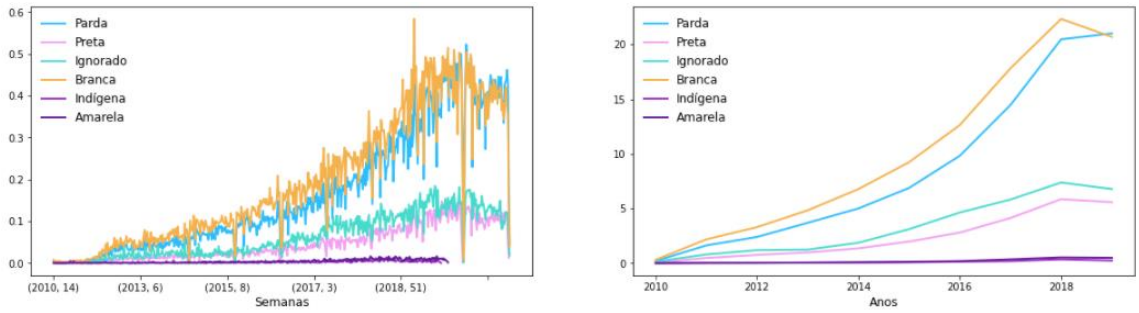
Figura 7 – Taxa de detecção de sífilis adquirida (por 100.000 habitantes) no Brasil segundo a faixa etária, de 2010 a 2019, por semana e por ano.



Fonte: Autoria própria.

Na Figura 8 observa-se o comportamento dos casos por cor ou raça. As curvas mostram alguns grupos com comportamentos similares: pessoas pardas e brancas concentram a maior quantidade de casos, seguidos das pessoas pretas, e por fim, com os menores números os indígenas e amarelos. A classe ignorada tem o comportamento próximo a curva de pessoas de cor preta.

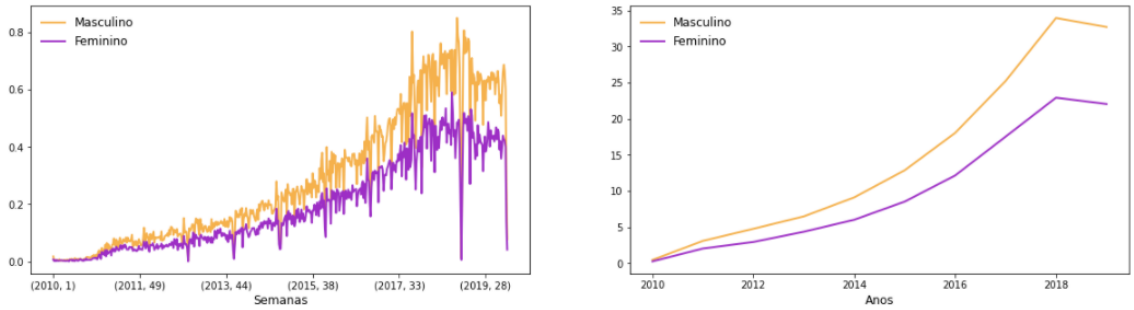
Figura 8 – Taxa de detecção de sífilis adquirida (por 100.000 habitantes) no Brasil segundo a cor, de 2010 a 2019, por semana e por ano.



Fonte: Autoria própria.

Na Figura 9 se destaca a forte correlação entre as séries temporais do sexo Masculino e Feminino. Revela também a predominante incidência entre os homens. A visão anual evidencia um aumento na magnitude da diferença entre as curvas dos sexos.

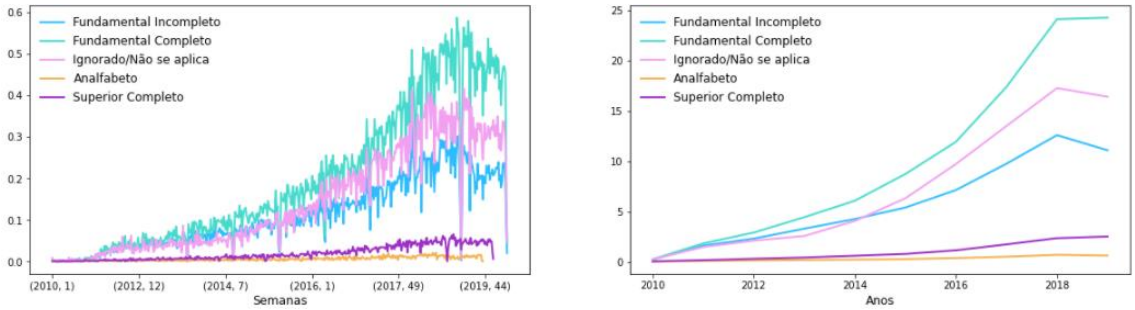
Figura 9 – Quantidade de casos de sífilis adquirida no Brasil segundo o Sexo, de 2010 a 2019, por semana e por ano.



Fonte: Autoria própria.

A Figura 10 é observado que para a escolaridade, a maioria dos casos da infecção se concentram nas pessoas com ensino fundamental completo ao longo do tempo. A classe ignorada representa boa parte dos dados com uma tendência de crescimento similar às pessoas com fundamental incompleto. Por fim, as menores taxas estão entre as pessoas analfabetas e com ensino superior completo.

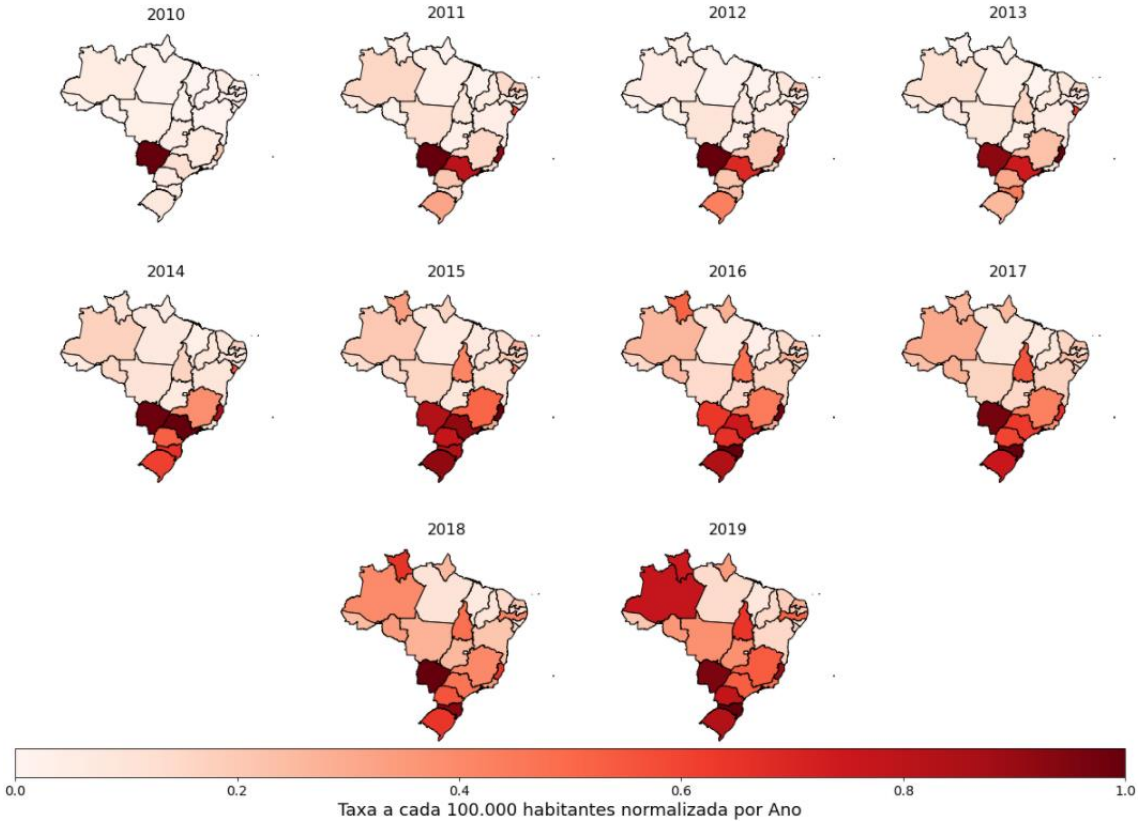
Figura 10 – Taxa de detecção de sífilis adquirida (por 100.000 habitantes) no Brasil segundo a escolaridade, de 2010 a 2019.



Fonte: Autoria própria.

A Figura 11 contém os mapas de calor do Brasil de acordo com a taxa de detecção da sífilis adquirida, normalizada para cada estado ao longo da última década. As Regiões sul e sudeste concentram grande parte dos casos. Mato Grosso do Sul (MS) possui uma forte taxa de detecção para a maioria dos anos, com exceção de 2015 e 2016. Nos últimos três anos da base de dados, Santa Catarina, Espírito Santo e MS aglutinam as maiores taxas.

Figura 11 – Mapa de calor da taxa de detecção de sífilis adquirida (por 100.000 habitantes) no Brasil normalizada por ano, de 2010 a 2019, por semana por ano.



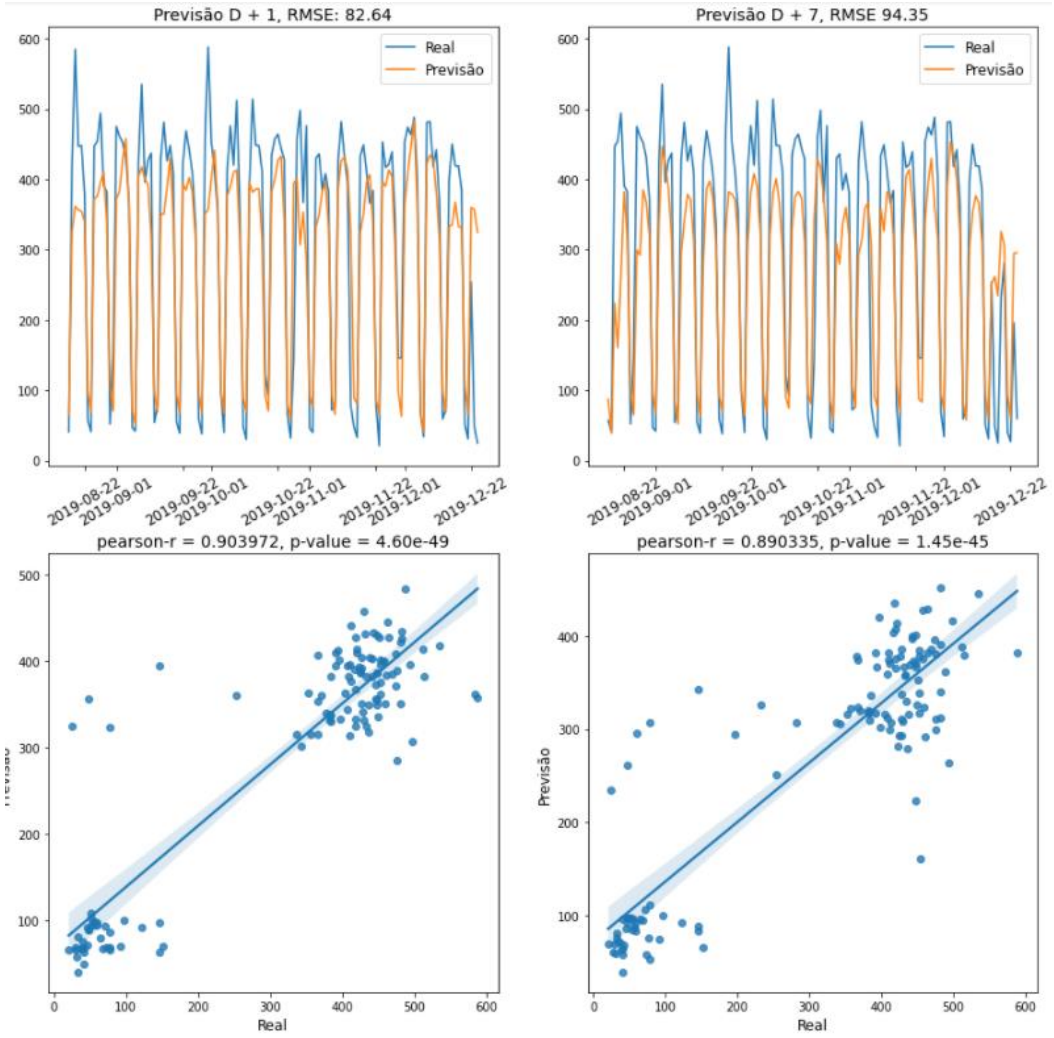
Fonte: Autoria própria.

4.2 MODELAGEM

4.2.1 MODELO DIÁRIO

A primeira linha de gráficos da Figura 12 mostra a proximidade entre a série temporal real, e as previsões do modelo para o dia seguinte (D + 1), como também para sete dias à frente (D + 7). O modelo apresenta uma boa capacidade de generalizar o comportamento da curva. Na segunda linha o gráfico de dispersão mostra alta correlação entre os valores reais e os previstos (para D + 1 e D + 7).

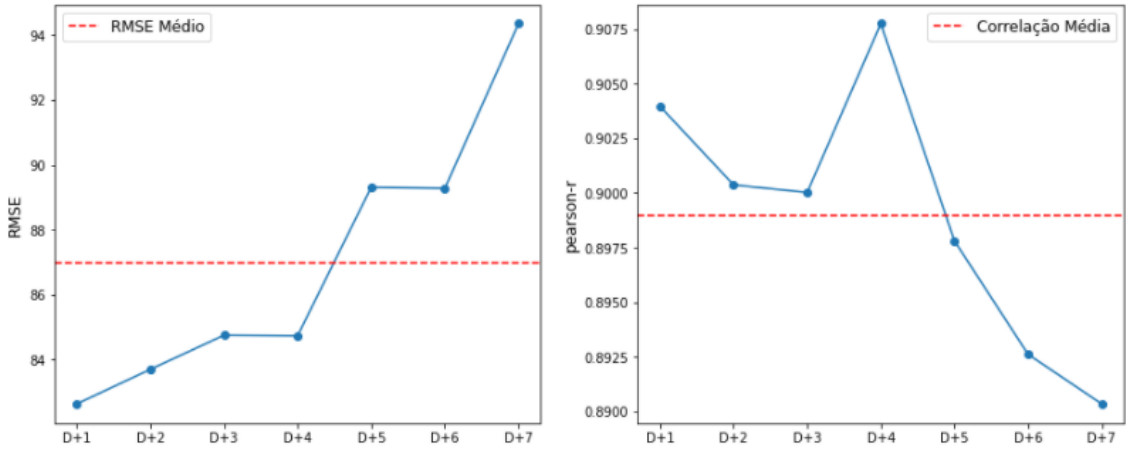
Figura 12 – Comparação do número real de casos e das previsões para o modelo diário.



Fonte: Autoria própria.

Os RMSE para cada um dos dias previstos (D_n) são plotados na Figura 13, na medida que o dia de previsão D_n fica mais distante, maior é o erro e menor a correlação. Um comportamento esperado, uma vez que quanto maior o espaço de tempo entre os dados de previsão e inferência, maior é a probabilidade de que novos eventos e tendências que o modelo ainda não viu, aconteçam.

Figura 13 – RMSE e coeficiente de correlação, para a previsão de cada um dos dias.



Fonte: Autoria própria.

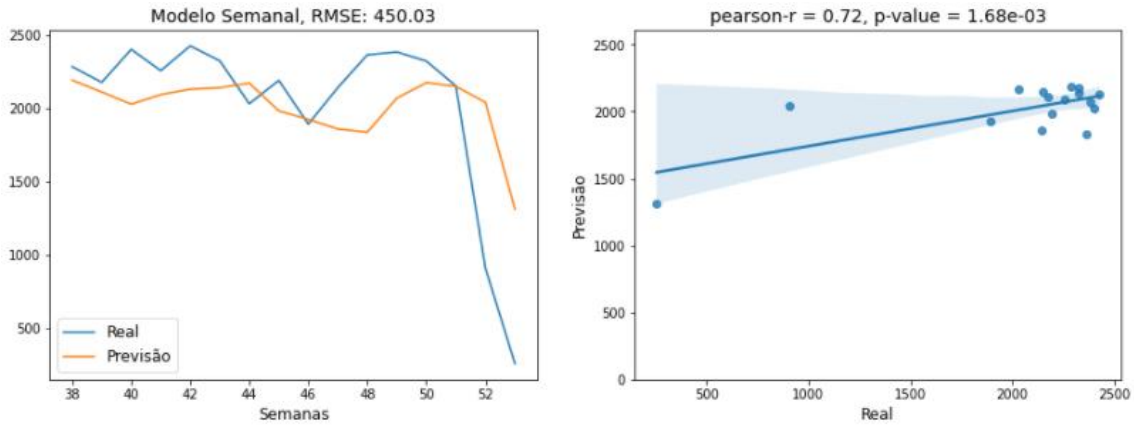
A média do número de casos no conjunto de teste do modelo diário é de 317,63 e o RMSE médio foi de 86,97. Esse RMSE representa aproximadamente 27,7% da quantidade média do número de casos no conjunto de teste o que caracteriza um baixo erro. A correlação é uma medida de similaridade, com valor variando de zero, nenhuma similaridade, até 1, que indica máxima similaridade positiva, ou -1, que indica máxima similaridade negativa. A correlação média entre serie temporal real do número de casos e a prevista pelo modelo diário foi de 0.90, indicando alta similaridade das previsões com a realidade.

4.2.2. MODELO SEMANAL

O modelo semanal possui um pequeno deslocamento (lag) de atraso em relação a curva real, como mostrado na Figura 14. Também apresentou forte correlação, porém inferior, e um RMSE maior quando comparado ao diário. Essa queda de desempenho

ocorre devido a mudança de frequência. Ao agrupar o número de casos diários por semana, os erros são acumulados e interferem no aprendizado do modelo.

Figura 14 – Comparação do número real de casos e das previsões para o modelo semanal.



Fonte: Autoria própria.

O RMSE do modelo semanal representa 22,1% do número médio de casos no conjunto de teste semanal que foi de 2031,93. Mesmo com esse erro representativo menor quando comparado ao modelo diário, o comportamento de atraso das previsões do modelo semanal impacta em uma menor correlação com valor de 0,72 que poder ser observada no gráfico de correlação da Figura 14.

A menor quantidade de dados do modelo semanal afeta diretamente a capacidade de aprendizado da rede neural. O lag desse modelo diminui a capacidade antecipativa de previsão da incidência da infecção, e consequentemente o poder do algoritmo enquanto ferramenta de DOD. No entanto, o modelo semanal foi capaz de acompanhar (mesmo com o lag) a curva original com uma boa correlação.

5. CONCLUSÕES

O uso de Deep Learning e redes neurais se mostraram eficazes para a previsão dos casos de sífilis adquirida no Brasil. As arquiteturas baseadas em LSTM obtiveram resultados satisfatórios para prever a quantidade de casos da infecção. Sendo assim, os modelos são ferramentas de ML que podem ser utilizadas como apoio às ações de contingência e controle da sífilis adquirida no Brasil.

A análise exploratória mostrou que a infecção possui maior grau de incidência em determinados locais do Brasil (Figura 11). Além de atingir em maior quantidade determinados grupos sociais, como os indivíduos com baixo nível de escolaridade, ensino fundamental incompleto (Figura 10) e pessoas na faixa etária de 20 a 29 anos (Figura 7). Esses fatores podem ser indicativos de correlação da sífilis adquirida a características socioeconômicas e aspectos geográficos.

Como perspectivas futuras para este trabalho, segregar o modelo através das diferentes categorias dos dados como escolaridade, raça e as regiões para analisar e compreender a influência destas nas previsões, e com isso utilizar modelos mais precisos e robustos a outros fatores além do próprio número de casos. Além disso aplicar os modelos da sífilis adquirida a fim de encontrar possíveis correlações com outros categorias da infecção como a congênita.

O desenvolvimento deste trabalho converge com o atual cenário que a sociedade como um todo vive. A pandemia da Covid-19 mostrou a importância dos dados como um forte aliado na condução de políticas e ações em prol da saúde pública. Nesse contexto, os modelos de Deep Learning e Machine Learning no geral são poderosos instrumentos de Decisão Orientada a Dados e quando aplicados de forma correta podem salvar vidas.

6. REFERÊNCIAS

Aurlien Gron. 2017. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (1st. ed.). O'Reilly Media, Inc.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult." *IEEE transactions on neural networks* 5.2 (1994): 157-166.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância, Prevenção e Controle das IST, do HIV/ Aids e das Hepatites Virais. Boletim Epidemiológico de Sífilis 2020. Brasília: Ministério da Saúde, 2020.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância, Prevenção e Controle das IST, do HIV/ Aids e das Hepatites Virais. Boletim Epidemiológico de Sífilis 2021. Brasília: Ministério da Saúde, 2021.

Chandra, Rohitash, Ayush Jain, and Divyanshu Singh Chauhan. "Deep learning via LSTM models for COVID-19 infection forecasting in India." *arXiv preprint arXiv:2101.11881* (2021).

Connor, Jerome T., R. Douglas Martin, and Les E. Atlas. "Recurrent neural networks and robust time series prediction." *IEEE transactions on neural networks* 5.2 (1994): 240-254.

Fawcett, Tom & Provost, Foster. (2013). Data Science for Business.

Gautam, Yogesh. "Transfer Learning for COVID-19 cases and deaths forecast using LSTM network." *ISA transactions* (2021).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Governo Federal. Projeto “Dífilis Não” muda cenário da doença no Brasil. Disponível em: https://www.gov.br/mec/pt-br/assuntos/noticias/projeto-201csifilis-nao201d-muda-cenario-da-doenca-no-brasil. Publicado em 21 de outubro de 2021.

Haider, Najmul, et al. "Lockdown measures in response to COVID-19 in nine sub-Saharan African countries." BMJ Global health 5.10 (2020): e003319.

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

Ibáñez-Cervantes, G., et al. "Epidemiological behavior and current forecast of syphilis in Mexico: increase in male population." Public Health 185 (2020): 386-393.

Knaflic, Cole Nussbaumer, *Storytelling With Data: A Data Visualization Guide for Business Professionals*. Hoboken, New Jersey: John Wiley & Sons, Inc, 2015.

Olah, Christopher. Understanding LSTM Networks, 2015. Disponível em: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Postado 15 de agosto de 2015.

Ruiz-Sancho, Andrés, et al. "Outbreak of syphilis, but not of acute hepatitis C, among HIV-infected homosexual men in Madrid." *HIV clinical trials* 8.2 (2007): 98-101.

Sanchez-Gendriz, Ignacio, et al. "Knowledge discovery and dengue forecasting applied in a four-year dataset collected at Natal/RN–Brazil." (2021).

Sanchez-Gendriz, Ignacio. "Signal processing basics applied to Ecoacoustics." *bioRxiv* (2021).

Stamm, Lola V., and Benjamin Mudrak. "Old foes, new challenges: syphilis, cholera and TB." *Future microbiology* 8.2 (2013): 177-189.

SINAN, Sistema de Informação de Agravos de Notificação, 2016. Disponível em: <http://portalsinan.saude.gov.br/o-sinan>. Publicado em 07 de Março de 2016.

SINAN, Dicionário de Dados – SINAN NET – Versão 5.0, 2016. Disponível em: < http://portalsinan.saude.gov.br/images/documentos/Agravos/Notificacao_Individual/DIC_DADOS_NET---Notificao-Individual_rev.pdf>. Revisado em Novembro de 2018.

Weng, R. X., et al. "Time series analysis and forecasting of chlamydia trachomatis incidence using surveillance data from 2008 to 2019 in Shenzhen, China." Epidemiology & Infection 148 (2020).

Zain, Zuhaira M., and Nazik M. Alturki. "COVID-19 pandemic forecasting using CNN-LSTM: a hybrid approach." Journal of Control Science and Engineering 2021 (2021).