MAC0499 - Trabalho de Formatura Supervisionado
Projeto: Continuous Training no projeto SPIRA

Orientador: Prof. Alfredo Goldman vel Lejbman
Orientador: Renato Cordeiro Ferreira
Aluno: Daniel Angelo Esteves Lawand
Número USP: 10297693

1 Introdução

Insuficiência respiratória é um sintoma médico definido como o comprometimento da troca gasosa entre o sangue do indivíduo e o meio ambiente [1] e pode resultar na redução de oxigênio e o aumento de dióxido de carbono na corrente sanguínea. Tais sintomas podem levar à tosse, ao cansaço, à falta de ar e, em alguns casos, à morte. A insuficiência respiratória pode ser causada por diferentes doenças, tais como asma, gripe, doenças cardíacas, COVID-19.

Apesar da insuficiência respiratória ser amplamente estudada, há dificuldade de realizar o diagnóstico de um dos sintomas da insuficiência respiratória: a hipóxia silenciosa [5], que causa baixa concentração de oxigênio na corrente sanguínea sem ocasionar falta de ar. Durante a pandemia do COVID-19, surgiu o projeto SPIRA [2], que visa desenvolver um sistema inteligente capaz de realizar o pré-diagnóstico da insuficiência respiratória com o sintoma de hipóxia silenciosa, usando, para tanto análise de fala baseada em modelos de *Machine Learning* (ML).

O estado atual do projeto SPIRA é resultado de diferentes trabalhos anteriores [2, 4]. Este projeto, tomará como base o trabalho descrito no artigo '*SPIRA: Building an Intelligent System for Respiratory Insufficiency Detection*' [4]. Nesse trabalho, foram desenvolvidos 3 componentes: uma interface que realiza a coleta de áudios de pacientes e os armazena em um banco de dados, uma interface para a visualização dos dados coletados, e um sistema de inferência inteligente que realiza o pré-diagnóstico do paciente a partir dos áudios enviados via um aplicativo.

Conforme novos dados são coletados ao longo do tempo, é possível realizar o retreino do modelo de ML para gerar predições com maior acurácia. Contudo, atualmente, o projeto SPIRA não possui uma *pipeline* de treinamento que permita retreino periódico automatizado.

Dessa forma, este projeto da área de *Machine Learning Operations* (MLOps) tem como objetivo habilitar o treinamento contínuo de modelos de aprendizado de máquina no projeto SPIRA, implementando uma *pipeline* que permitirá o retreino automático, sob demanda, de modelos à medida que novos dados rotulados sejam adquiridos.

2 Fundamentos

2.1 SPIRA

O SPIRA é um projeto de pesquisa iniciado durante a pandemia de COVID-19. Ele tem por objetivo desenvolver um sistema inteligente capaz de realizar o pré-diagnóstico da insuficiência respiratória com o sintoma de hipóxia silenciosa, usando, para tanto, análise de fala baseada em modelos de *Machine Learning*. O sistema inteligente auxiliará profissionais de hospitais (médico, enfermeiros, etc.) para realizar a triagem de pacientes com urgência no tratamento de insuficiência respiratória.

2.2 Engenharia de *Machine Learning*

A engenharia de *Machine Learning* envolve o plano de criação, deploy, manutenção e atualização de sistemas baseados em ML. Isso permite que os responsáveis por manter esse sistema possam colher os benefícios que uma adoção baseada em aprendizado de máquina pode fornecer [6].

2.3 Treinamento de Modelos

O treinamento de um modelo de ML envolve o uso de um algoritmo de ML com dados dos quais o modelo irá aprender. O termo "modelo de ML" refere-se ao artefato criado pelo processo de treinamento [3].

2.4 Ciência de Dados

A Ciência de Dados fornece mecanismos para a exploração de dados em busca de padrões ou associações. O resultado dessa exploração permite a tomada de decisões estratégicas.

Existem diferentes processos para o desenvolvimento de projetos de Ciência de Dados. Uma das descrições amplamente conhecidas é a *Cross Industry Standard Process for Data Mining* (CRISP-DM) [7]. Essa descrição segmenta o processo de desenvolvimento em diferentes fases, como visto a seguir:

1. **Entendimento do negócio**: Busca entender qual o objetivo de negócio a ser atingido.
2. **Entendimento dos dados**: Visa coletar e entender os dados relacionados ao problema de negócio.
3. **Preparação dos dados**: Visa selecionar e tratar um subconjunto dos dados coletados.
4. **Modelagem**: Visa treinar um modelo com base nos dados preparados.
5. **Avaliação**: Visa avaliar a acurácia do modelo e entender se os seus resultados atingem os objetivos de negócio.
6. **Implantação**: Visa organizar e apresentar o modelo de forma que o cliente possa utilizá-lo.

3 Objetivos

Com o intuito de aperfeiçoar os modelos de aprendizado de máquina do projeto SPIRA, este projeto tem como objetivo habilitar o treinamento contínuo de modelos, ou seja, implementar uma *pipeline* que permitirá o retreino automático, sob demanda, de modelos à medida que novos dados rotulados sejam adquiridos ao longo do tempo.

Para que esse objetivo geral seja alcançado, será necessário atingir os seguintes objetivos específicos:

1. implementar um modelo *baseline* que servirá como base para a *pipeline* de MLOps,
2. criar a *pipeline* de MLOps, e
3. automatizar a execução da *pipeline*.

4 Plano de Trabalho

Para a realizar este projeto, elaborou-se um cronograma de execução dividido nas seguintes tarefas:

1. **Estudo de trabalhos anteriores:** Estudar artigos, relatórios de iniciação científica e TCC's anteriores sobre os avanços no projeto SPIRA.
2. **Criação do conjunto de treinamento:** Extrair dados do SPIRA para treinar um modelo *baseline*.
3. **Implementação de um modelo *baseline*:** Treinar uma rede neural, cuja arquitetura será baseada na v1 do modelo implementado no artigo principal do SPIRA [2], que sirva como modelo *baseline* para a estruturação da *pipeline* de MLOps.

4. **Criação do sistema treinador:** Transformar *notebooks* que foram utilizados nas etapas anteriores em scripts. Essa etapa permitirá aplicar boas práticas de desenvolvimento. Além disso, permitirá a execução de cada módulo separadamente.

5. **Implantação de um agendador:** Adicionar uma ferramenta que executará a *pipeline* de treino sob demanda, por exemplo, o *Apache Airflow* ou outra ferramenta similar.

6. **Automatização da execução do sistema treinador:** Desenvolver uma *pipeline* de execução automática do sistema treinador implementado usando o sistema agendador.

7. **Escrita da monografia:** Escrever a manografia ao longo da pesquisa, descrevendo os resultados obtidos.

8. **Criação da apresentação e pôster:** Desenvolver o material de apresentação do TCC.

O tempo a ser despendido nessas tarefas estão contemplados no seguinte cronograma:

		Meses (2023)								
	Etapas	04	05	06	07	08	09	10	11	12
1	Estudo de trabalhos anteriores	X	X							
2	Criação do conjunto de treinamento		X							
3	Implementação de um modelo		X	X						
4	Criação do sistema treinador			X	X	X				
5	Implantação de um agendador					X	X			
6	Automatização da execução do sistema treinador						X	X		
7	Escrita da monografia	X	X	X	X	X	X	X	X	X
8	Criação da apresentação e pôster								X	X

Figura 1: Cronograma das atividades.

Adicionalmente, no momento, os autores deste projeto estão tentando submetê-lo para a bolsa de empreendedorismo da USP. Caso seja aceito, parte do projeto será desenvolvido na *Jheronimus Academy of Data Science* (JADS) parte da *Eindhoven University of Technology* (TUe) na Holanda.

5 Bibliografia

Referências

[1] EJM Campbell. Respiratory failure. *British medical journal*, 1(5448):1451, 1965.

[2] Edresson Casanova, Lucas Gris, Augusto Camargo, Daniel da Silva, Murilo Gazzola, Ester Sabino, Anna Levin, Arnaldo Candido Jr, Sandra Aluisio, and Marcelo Finger. Deep learning against COVID-19: Respiratory insufficiency detection in Brazilian Portuguese speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 625–633, Online, August 2021. Association for Computational Linguistics.

[3] Documentação da AWS. Treinamento de modelos, 2023.

[4] Renato Ferreira, Dayanne Gomes, Vitor Tamae, Francisco Wernke, and Alfredo Goldman. Spira: Building an intelligent system for respiratory insufficiency detection. In *Anais do II Workshop Brasileiro de Engenharia de Software Inteligente*, pages 19–22, Porto Alegre, RS, Brasil, 2022. SBC.

[5] Franco Laghi Martin J Tobin and Amal Jubran. Why covid-19 silent hypoxemia is baffling to physicians. *Am. J. Respir. Crit. Care Med.*, 202(3):356–360, 2020.

[6] Ben Wilson. *Machine Learning Engineering in Action*, chapter What is machine learning engineer? Manning Publications Co., 2022.

[7] Rüdiger. Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 01 2000.