Identifying High-Impact Artificial Intelligence and Machine Learning use-cases at CDC

Link to form: Microsoft Forms (office.com)

Thank you for your interest developing Artificial Intelligence (AI) and Machine Learning (ML) applications at CDC.

In April 2024, the Office for Public Health, Data and Surveillance, and Technology (OPHDST) is launching an "AI/ML Acceleration Program" to support the development and operationalization of AI and ML applications to address public health goals. (For more information on Artificial Intelligence and Machine Learning at CDC, please consult: https://intranet.cdc.gov/ai)

The AI Acceleration Program partners with technical teams throughout CDC to provide technical support and resources to partner CDC programs to develop high-impact use-cases. Resources provided include, but are not limited to:

- Data Science and Data Engineering support

- Design and Product Management

- Best practices to build responsible AI/ML applications at scale

- Guidance through clearance and launch of AI/ML powered applications

- Increased visibility of selected AI/ML projects from senior leadership at CDC

The AI Acceleration Program accepts submissions both for existing projects (which might need additional support to operationalize a prototype) and for projects which are still only at an ideation stage.

We expect partner organizations to have identified an internal project lead, and ideally already have a team in place, ready to receive additional technical support.

If you are interested in submitting a project for consideration, please fill in the following form to share with us a high-level overview of the AI/ML project you wish to develop. The AI Acceleration team may reach out to you for further information if necessary.

We do not expect teams to have fully worked out the implementation details for the proposed use cases. Rather, the goal is to quickly identify the most promising use cases. Once use cases are identified, the AI Acceleration Team will partner with you to further refine scope and work out an implementation plan.

Submissions will be scored using the RICE framework (assessing the Reach, Impact, Confidence and Effort required for each project) - the highest-scoring use-cases will receive resources to develop and operationalize the AI/ML applications.

The projected timeline for project selection is as follows:

March 21: submissions open.
April 2: deadline for submissions.
First week of April: prioritization and selection of AI/ML projects
Mid-April: work begins on selected projects (based on partners' availability)

Please provide a brief description of the project:

4. *What are you trying to build? If you can articulate this, how does this project employ Artificial Intelligence or Machine Learning capabilities?*

Using the CDC Data Hub (CDH) Databricks environment, we are building a reusable data pipeline to clean markup in clinical narratives that later use name entity recognition (NER) from BERT to identify symptoms, diseases, and medications. For this project we are using standard CDH data and custom Databricks GPU compute with parameterized "Notebooks" in Job Workflows with MLFlow (Databricks Experiments) Integration, and all source code is in CDCENT Github repositories. This approach can be used for featurization of multiple machine learning models. We are using electronic health records as a test case for this pipeline and expect this to be of used for a wide mix of projects that need preprocessing and featurization.

5. What is the problem that this is trying to solve? The question this is trying to answer?

Medical narratives available in electronic health records and electronic case reporting come with software specific markups that makes the extraction of meaningful information difficult, this is because it adds background noise to a machine learning model that provide biased results. We are developing an all-purpose pipeline that cleans XML, HTML, and RTF markup from clinical narratives that are later passed through a NER task to identify symptoms, medications, and diseases. At scale, these notes are cumbersome to clean because of the computational resources required to process clinical notes. To address that, we used the CDC Data Hub Databricks cloud computing assets based on Spark technology to clean this data and make it analysis-ready for featurization. We include a NER step which allows the user to characterize quickly relevant covariates that later can be used for prediction.

6. How does this project solve the problem? How does the problem go away if it is solved? Who would use this solution?

Please describe how the user(s) would use this. How many users will be utilizing this product? For example, would this include Healthcare (AKA Labs, healthcare systems, etc), STLTs (AKA Jurisdictions, non-Federal Public Health Agencies, etc), Public (AKA NGO, citizens, etc), a division/program of National Public Health (AKA CDC, Federal Partners), or Other?

This project reduces the burden of cleaning markup in medical notes and streamlines the process for analysis. Enables the creation of analysis ready clinical notes for featurization which is achieved by using pre-trained NER models. The users of this solution are researchers within the CDC or any other agency this tool can be deployed that have access to unstructured data that can be used for further analysis.

7. Does the project support "Readiness and Response" efforts?

yes

8. Who would benefit from the solution? Please describe how the user(s) would benefit from this. *For example, would this include Healthcare (AKA Labs, healthcare systems, etc), STLTs (AKA Jurisdictions, non-Federal Public Health Agencies, etc), Public (AKA NGO, citizens, etc), a division/program of National Public Health (AKA CDC, Federal Partners), or Other?*

CDC researchers and practitioners within the agency or local agencies that use compatible technology and share interest in using nonstructured data.

9. How would you describe the impact of this project? *Impact can be measured in many different ways. For example, reducing burden, increasing efficiency, direct public health impact...)*

It reduces the burden of cleaning clinical notes. Using this approach, the clinical notes are cleaned efficiently and make analysis-ready for modeling depending on user needs. The core Data Asset will be published on the CDC Datahub. This can support public health efforts when dealing with emerging diseases during outbreaks.

10. Who would be the project lead? lead technical developer?

The lead technical developer is Kate Belisle (ulc0@cdc.gov) and the project lead would be Oscar Rincon Guevara (ulc0@cdc.gov) from OPHDST

11. Who else would be involved in the project?

Directly, no one else. Indirectly supervisors and data engineers within the CDC Data Hub.

12. How would you describe the team's level of familiarity with Artificial Intelligence/Machine Learning technology?

Kate Belisle has over twenty years experience in distributed analytics in a “big data” environment, including over a decade of experience in processing unstructured data. The team has familiarity with artificial intelligence and machine learning. We are building reusable parameterized pipelines to support users with their machine learning needs. Oscar Rincon Guevara is a Prevention Effectiveness Fellow in the Public Health Analytics and Modeling Track. He has a PhD in Industrial Engineering with experience in Operation Research and Statistical Learning applied to a wide variety of knowledge domain including healthcare. With Kate’s technical support, Oscar designs and creates new approaches to use reusable AI/ML methodologies for easy user adherence and interaction.

13. How soon can you begin to work on this project?

The project is in progress.

14. Development status: Is this a new project, currently in development, or an extension of a completed pilot? If work has already started, what platform(s) & technology(ies) are currently in use?

The project is an upscale from a prototype created in October 2023. The platform is Azure Data Bricks using CPU and GPU dedicated clusters with spark as scripting language. The technology behind is the distributed architecture from Azure Databricks for data and compute, and Named Entity Recognition (NER) from pre-trained transformer models on the modeling side, and Spark/SparkML components including SparkNLP, XGBoost and SHAP.

15. What is the estimated timeline to bring the project to the next phase?
(For example: from ideation to the creation of a prototype, from an early pilot to operationalization...)

To make the full pipeline completely operational we are less than a week away from having all the components implemented on our test cohort in four separate pipelines. After ensuring that technical components are functioning to spec, the team will scale up processing, including compute, to the full cohort.

16. What support would you need for this project to be a success?

We would like to ask for support identifying users for this type of pipelines and additional data sources where this pipeline can be used to support analysis across the agency. We would be happy to collaborate with you in any capacity to advance NLP as an additional capability at the CDC.

17. What data sources will you plan to use for this project?
Will you need help getting access to data sources? For example, any data the team might not have access to, data from systems that you need permission to use, etc.

We are currently using an EHR called American Board of Family Medicine, which has clinical notes and visits. However, that is the only data sets where we have that type of unstructured data, reason why we would like to request your expertise in finding additional data sources where we can perform additional testing and users looking to analyze text data.