The proposal document should address the following points. Use these points as headers in your document.

Basic Info. The project title, your names, e-mail addresses, UIDs, a link to the project repository.
Project title: Using natural language processing to predict patient outcomes/trajectory based on radiological imaging findings

Team members/names:

- Karlyn Ledesma , U1287358 , u1287358@utah.edu
- Nate Hayward, U6031381, nate.hayward@hsc.utah.edu
- Andre Chu, U1523752, andre.chu@utah.edu
- Soumya Chava, U1538585, u1538585@utah.edu
- Matias Lee, U1447587, u1447587@utah.edu

Repository Link:
https://github.com/KLedesma-Men/Group-1-Project-

Background and Motivation. Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

Children frequently experience head injuries due to falls, accidents, or sports-related impacts. When a child presents with head injury, medical professionals must quickly assess the severity of the trauma and decide whether a CT scan is necessary. While CT scans are a valuable tool in detecting serious brain injuries, their frequent use raises concerns about radiation exposure, especially in young patients.

Our interest in this topic stems from a border curiosity about how data science can support medical decision-making. As students studying healthcare and informatics, we recognize the importance of evidence-based approaches to improve patient outcomes while minimizing risks. Through this project, we aim to analyze existing datasets to better understand which clinical indicators reliably predict serious injuries,helping doctors make more informed choices. By refining risk assessment models, we hope to contribute to safer and more efficient healthcare practices, reducing unnecessary imaging while ensuring that patients who truly need CT scans receive timely intervention.

Project Objectives. Provide the primary questions you are trying to answer with your data. What would you like to learn and accomplish? List the benefits of how the data could be useful.

1. Can combining Glasgow Coma Scale scores and neurological deficits improve the accuracy of predicting severe head injuries?
2. Given the clinical data, can we use natural language processing to classify patients into categories of high, moderate, or low risk for clinically important TBI?
    a. The output of a predictive model can be extremely useful for clinicians and allow for the identification of patients of high priority and risk
3. Can we predict the risk factors of a patient using the patient dataset's features such as injury mechanism and symptoms?
    a. This type of data allows clinicians to identify the most indicating symptoms that contribute to a diagnosis of significant TBI

Data:
We obtained the data from the Pediatric Emergency Care Applied Research Network (PECARN) website (https://pecarn.org/datasets/). We plan to use the data from the 'Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study' dataset which has an enrollment of 42412 patients. These data have already been collected and are available for download and analysis.

- **Data Processing.** Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?
    - Yes, substantial data clean up will be required due to the large data set. This will include missing values, duplicate values, converting datatypes, and standardized terminology for radiology reports.
    - Based on the research questions we could use the following quantities to derive data: descriptive statistics, NLM (word frequency), and regression models.
    - Data processing can be implemented in various way using multiple tools. For data cleaning we can use pandas and numpy from python. For modeling and statistical analysis, correlational and regression models to link the correlation. As mention we could use NLM to determine risks.

- **Design.** How will you display your data? Provide some general ideas that you have for the design. Develop **one alternative prototype design for your data**.. Describe your designs and justify your choices of visual encodings.
    - Prototype 1:
        - Summary Stats: basic distribution of key variables used in our analysis
        - Heatmap to visualize risk stratification:

            - Plots clinical symptoms vs risk levels, with a darker color indicating a stronger correlation to a certain risk level
            - Visualizes which symptoms contribute to higher risks in a organized manner
        - Scatter Plot showing Glasgow Coma Scale vs Neurological Deficits
            - By plotting GCS scores vs the presence of neurological deficits, we can visualize if combining the two improves the prediction of TBI severity
        - Plot of feature importance:
            - Visualize how each indicator contributed to the prediction model
        - Flowcharts:
            - Using model outputs, indicate how our project contributes to clinical decision support
    - Prototype 2:
        - Decision Tree visual:
            - Visualize the predictive pathway used for classification
                - Nodes represent symptoms, indicators from radiology report, GCS scores, etc
            - Clearly shows users the strongest predictors of high-risk patients that need immediate care
        - Risk heat map
            - Classifies patients into high, moderate and low risk
            - Indicates the correlation of injury mechanisms, symptoms and GCS scores to represent risk probability
        - Frequency Chart from Radiology Reports
            - Using our NLP approach, we can produce a chart indicating the most common terms used in reports for the three different categories of risk
            - Indicates terms in radiology reports most associated to higher risk cases

- **Must-Have Features.** List the features without which you would consider your project to be a failure.
    - Careful data wrangling and cleaning
    - Precise and accurate risk classification of patients
    - Analysis of strong indicators of TBI risk
    - User friendly output visuals, providing a clear explanation of actionable insights gained from the data
    - NLP integration to process radiology reports

- **Optional Features.** List the features which you consider to be nice to have, but not critical.
    - Interactive user features
    - User feedback opportunities
    - Analyze risk factors with the use of time-series modeling
    - Machine Learning Model for Risk Prediction (e.g. Naïve Bayes, Logistic Regression, or Decision Tree)

- **Project Schedule.** Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.
    - Perhaps meeting weekly?
    - Set up a date from where data cleaning should be done?

Week 1: Weekly meetings will be set up to go over project goals and divide work based on the personal strengths of each member. We will periodically check in and provide updates on our progress and make changes to our project design if necessary. The first week will focus on data cleaning in order for our analysis to be properly performed.

Week 2:

- Finish with data cleaning and processing.
- Do a summary statistics and visualization. This will help us understand the dataset. It will help us detect patterns, trends, and anomalies in our data before applying machine learning or statistical models.
- Identify risk factors for TBI. We have to determine which factors are most associated with severe TBI.

This proposal is the first part of your process book. As a ballpark number: your proposal should contain about 3-4 pages of text.

You will schedule a project review meeting with a staff member during regular lecture times of the week marked in the schedule. Make sure all of your team members are present at the meeting.

The proposal will be submitted by uploading it to your team's GitHub repository.

References:

1. Kuppermann N, Holmes JF, Dayan PS, Hoyle JD Jr, Atabaki SM, Holubkov R, Nadel FM, Monroe D, Stanley RM, Borgialli DA, Badawy MK, Schunk JE, Quayle KS, Mahajan P, Lichenstein R, Lillis KA, Tunik MG, Jacobs ES, Callahan JM, Gorelick MH, Glass TF, Lee LK, Bachman MC, Cooper A, Powell EC, Gerardi MJ, Melville KA, Muizelaar JP, Wisner DH, Zuspan SJ, Dean JM, Wootton-Gorges SL; Pediatric Emergency Care Applied Research Network (PECARN). Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. Lancet. 2009 Oct 3;374(9696):1160-70. doi: 10.1016/S0140-6736(09)61558-0. Epub 2009 Sep 14. Erratum in: Lancet. 2014 Jan 25;383(9914):308. PMID: 19758692.
2. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, Grover C, Suárez-Paniagua V, Tobin R, Whiteley W, Wu H, Alex B. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak. 2021 Jun 3;21(1):179. doi: 10.1186/s12911-021-01533-7. PMID: 34082729; PMCID: PMC8176715.