**Name:** Beth Birhanu

**Contact Info:** bbirhanu@regis.edu

**Title of the Project:** Analyzing CDOT Crash Data for Predictive Modeling and Pattern Analysis

**High-Level Description:** The project aims to analyze Colorado Department of Transportation (CDOT) crash data from 2007 to 2023 to develop predictive models for crash occurrences and conduct pattern analysis to identify key factors contributing to crashes.

**Type of Data Science Task:**

1. Classification using supervised learning for predictive modeling.

2. Data visualization and exploratory data analysis for pattern analysis.

**Data:** The data consists of historical CDOT crash data from 2007 to 2020, and 2021 to 2023, including variables such as crash location, time, severity, injury levels, road conditions, weather, vehicle information, driver characteristics, and more. The dataset is expected to be moderately large, containing thousands to tens of thousands of records per year.

**Data Analysis Methods:**

1. For predictive modeling, I plan to use supervised learning algorithms such as logistic regression, decision trees, and random forests. I will split the merged CDOT crash data from 2007 to 2023 into training and validation sets. Specifically, I will train the models on a portion of the data (e.g., data from 2007 to 2020) and validate them using the remaining data (e.g., data from 2021 to 2023). This approach ensures that the models are trained on historical data and validated on more recent data, allowing for robust performance assessment across different time periods.

2. For pattern analysis, I will do EDA to uncover trends, correlations, and patterns in the data across different years. I will use statistical methods and data visualization techniques to identify key factors contributing to crashes.

**Anticipated Difficulties:**

1. Managing and integrating data from multiple years: The project involves analyzing CDOT crash data from 2007 to 2023, which may have varying data formats, structures, and quality across different years. Integrating and harmonizing these datasets poses a challenge in terms of data consistency and preprocessing.
2. Dealing with different features and columns: The CDOT crash data includes a wide range of features and columns related to crash details, road conditions, weather, vehicle information, driver characteristics, and more. Handling and selecting relevant features for analysis while managing the diversity of columns requires careful consideration and domain knowledge.
3. Addressing imbalanced data: The severity levels of crashes (Property Damage Only, Injury, Fatal) may exhibit class imbalance, with certain severity levels being more prevalent than others. Balancing the data for accurate predictive modeling and pattern analysis is essential but can be challenging.
4. Interpretation of complex patterns: Discovering complex patterns and relationships in the data, especially across different years, may require advanced statistical analysis.

**Timeline:**

Week 1: Problem definition, data collection (2007-2023), and initial planning.

Week 2: Merge and preprocess the CDOT crash data from different years.

Week 3-4: Data preprocessing, exploratory data analysis (EDA), and feature selection for the merged dataset.

Week 5-6: Model training and evaluation for crash prediction using the merged data.

Week 7: Pattern analysis, visualization, and insights generation for the merged dataset.

Week 8: Final analysis, report writing, and presentation preparation.