Project Management Plan

Cory Chitwood, Eddy Doering, Kai Gui, Keith-Jordan Wilkinson

Concerns and Priorities:

In this capstone project, our goal is to apply the information we have learned over the past three months to create an in-depth data analysis within the medical field.

Communication within team/Splitting of workload:

Our group will be communicating with each other during the day through zoom, and after work hours in a Microsoft Teams group chat. Within this group chat, we can share code and other information with each other. In terms of splitting the workload, each of us has our strengths – we should play to our own strengths with what we end up doing. In general, the group will be working together on tasks.

Find possible datasets/ Select a dataset: We need to search for a dataset that will provide ample data to properly analyze. With a dataset that we are comfortable with, we can then begin to understand the dataset we select. In our case, we found a dataset of incoming diabetes patients from around the U.S. 1999-2008.

Expected date of completion: 10/22/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Find dataset	*2hrs*	*Complete*	*Everyone*
Review data; is it workable	*1hr*	*Complete*	*Everyone*
Decide to use the dataset	*1hr*	*Complete*	*Everyone*

Understand the dataset: In order to understand what questions we should ask; and the plan of attack with the data, we need to understand what the data is telling us. For our dataset, there was an additional paper that gave insight into the variables, as well as preliminary data analysis. This is an important piece of our project, having a solid foundation will allow us to create a higher quality project.

Expected date of completion: 10/25/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
What do the variables mean?	*2hrs*	*Complete*	*Kai, Cory*
Read supplemental material	*3hrs*	*Complete*	*Kai, Cory*
Find supporting material	*2hrs*	*Complete*	*Cory*
Go over findings as a group	*1hr*	*Complete*	*Everyone*

Create Outline:

Draft Questions: With our knowledge of the dataset that we chose to use, we are able to create a plan of attack with questions we want to answer throughout our project. Anywhere in the range of 5-10 questions will suffice for the project, giving us direction on where to go.

Napkin Drawings: With the questions we want to answer, we will be able to draft out some rough drawings of the visualizations we plan to use. This will also provide a general structure of how the dashboard may be laid out.

Expected date of completion: 10/25/2021 or 10/26/2021

Receive feedback on questions and napkin drawings: When we are done with the outline, we will need to present what we have so far to the instructors and from our peers. We will take any feedback they give us during this time and apply it to our project.
Expected date of completion: 10/26/2021

Sub Task:	**Estimated Time**	**Completion Status**	**Assigned To**
Find people to review	*1hr*	*Complete*	*Cory*
Receive and review feedback	*1hr*	*Complete*	*Cory*
Share feedback with group	*1hr*	*Complete*	*Everyone*

Clean Dataset and start ETL: Remove any unnecessary columns/ columns with too little data reported. We want to keep solely what we will use in our machine learning application and visualizations. When we are doing the cleaning, we will also need to draft an ETL report, allowing someone to download the dataset in the same state we did, and transform it to what we will be using in our visualizations and ML.

Expected date of completion: 10/26/2021; TBD for ETL

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Clean dataset for visualizations	*2hrs*	*Complete*	*Cory*
Clean dataset for ML Models	*2hrs*	*Complete*	*Kai, Eddy, Keith-Jordan*
Create ETL Document	*1hr*	*Incomplete*	*Cory*
Describe cleaning steps in ETL	*3hrs*	*Incomplete*	*Keith-Jordan*

Set up Producer/Consumer in Apache Kafka: Take our clean data and send it through a data pipeline. With this set up we can use cloud applications for data analysis easily
Expected date of completion: 10/27/2021-10/28/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Set up Data Bricks	*1hr*	*Complete*	*Eddy, Kai*
Set up mounting point	*1hr*	*Complete*	*Eddy, Kai*
Write producer	*6hrs*	*In progress*	*Eddy, Kai*
Write consumer	*3hrs*	*Complete*	*Eddy, Kai*
Set up data factory	*2hrs*	*Complete*	*Eddy, Kai*
Run data pipeline	*3 days*	*In progress*	*Eddy, Kai*
Troubleshoot Issues	-	*In progress*	*Eddy, Kai*

Create SQL Database: Create a SQL database that allows us to use SQL to grab specific tables for data analysis as well as use Power BI with our data. With the SQL database set up, we can begin to create our visualizations and begin to understand the story the data is telling us.

Write a python code to create CSV files and drop into SQL: Once the SQL database is set up, we will need to be able to write this information into the database itself. If this is not completed, we will not be able to send the data we collected into Power BI.
Expected date of completion: 10/28/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Create ERD for Database	*1-2hrs*	*Incomplete*	*Keith-Jordan*
Create Cloud Data Flowchart	*3hrs*	*In progress*	*Keith-Jordan*
Access SQL Database	*1hr*	*Complete*	*Everyone*
Create necessary tables	*2-3hrs*	*Complete*	*Eddy*
Push data into SQL Server	*2hrs*	*Incomplete*	*Eddy, Kai*
Ensure tables interact well	*1hr*	*Incomplete*	*Cory, Eddy*

Create Visualizations in Power BI: Create the visualizations from the napkin drawings that we made. Also create additional visualizations that we believe will tell the best story for the dataset. Need to narrow down which visualizations we will present due to time constraints and information overload.

Expected date of completion: 10/29/2021-11/3/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Access data from SQL Server	*1hr*	*Incomplete*	*Everyone*
Create Visualizations based on Napkin Drawings:			
Create visualization 1	*2hrs*	*Incomplete*	*Cory*
Create visualization 2	*2hrs*	*Incomplete*	*Kai*
Create visualization 3	*2hrs*	*Incomplete*	*Eddy*
Create visualization 4	*2hrs*	*Incomplete*	*Keith-Jordan*
Create visualization 5	*2hrs*	*Incomplete*	*Keith-Jordan*
Create visualization 6	*2hrs*	*Incomplete*	*Cory*
Create visualization 7	*2hrs*	*Incomplete*	*Eddy*

Create a Machine Learning Program: We will need to create a machine learning application that performs a predictive analysis that takes the numerical and categorical data from the diabetes dataset. The goal of this ML is to predict the probability that someone will be readmitted to the hospital based on some of their traits. This will be the most difficult section of the project, and thus will also take the most time to complete.

Test multiple ML models to determine the most accurate model for the data: It will be necessary to create multiple models in order to determine which one is the best at predicting the probability of someone being readmitted to the hospital. This is dependent on time; we can create more or less models depending on how much time we have left in the project.

Expected date of completion: 11/3/2021 – 11/5/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Create Ridge Classifier Model	*4-8hrs*	*In progress*	*Kai*
Create Random Forest Model	*4-8hrs*	*In progress*	*Kai*
Create Ensemble Model	*4-8hrs*	*In progress*	*Kai*

Create KNN Classification	*4-8hrs*	*In progress*	*Eddy*
Create SGD Classifier Model	*4-8hrs*	*In progress*	*Keith-Jordan*
Determine best ML Model	*2-3hrs*	*Incomplete*	*Everyone*

Create napkin drawings for dashboard/ receive feedback: With a good clue of which visualizations we want to use, and a thorough understanding of our dataset, we can sketch out the composition of our dashboard. Do this as a group so everyone is on the same page on the dashboard.

Expected date of completion: 11/6/2021-11/7/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Sketch out dashboard	*3hrs*	*Incomplete*	*Cory*
Reach out for feedback	*1-2hrs*	*Incomplete*	*Cory*
Present feedback to group	*1hr*	*Incomplete*	*Everyone*

Create a dashboard in Power BI or Plotly: With the visualizations constructed for our project, as well as the results of the ML application, we will need a dashboard that effectively tells the story of our dataset/ our analysis of the data. For a reader with little understanding of the subject, this dashboard should be comprehendible.

Expected date of completion: 11/5/2021 – 11/9/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Introduction Text	*2hrs*	*Incomplete*	*Keith-Jordan*
Data Visualization	*2hrs*	*Incomplete*	*Cory, Kai*
ML Visualization	*2hrs*	*Incomplete*	*Eddy*
Supplemental Information	*1hr*	*Incomplete*	*Cory*

Write executive project report: With the data analysis complete, we will need to write a report on the data analysis we performed. This report should introduce and answer the questions we asked in the beginning of the project. Providing visualizations and explanations of them to answer these questions effectively. It will also be necessary to discuss the ML application we built, and the results of running the ML on the dataset.

Expected date of completion: 1/10/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Introduction and Conclusion	*2hrs*	*Incomplete*	
Data Exploration	*4hrs*	*Incomplete*	*Cory*
Analyze visualization 1	*2hrs*	*Incomplete*	*Cory*
Analyze visualization 2	*2hrs*	*Incomplete*	*Kai*
Analyze visualization 3	*2hrs*	*Incomplete*	*Eddy*
Analyze visualization 4	*2hrs*	*Incomplete*	*Keith-Jordan*

Analyze visualization 5	*2hrs*	*Incomplete*	*Keith-Jordan*
Analyze visualization 6	*2hrs*	*Incomplete*	*Cory*
Analyze visualization 7	*2hrs*	*Incomplete*	*Eddy*
Analyze ML Model Results	*3hrs*	*Incomplete*	*Kai*

Create final presentation: Create a ten-minute presentation on the project that we worked very hard on. It is imperative that we restrict the amount of data we present because of the time constraints put on us for the presentation. Ensure everyone has a good understanding of the project and what they need to say in the presentation. Practice each part of the presentation together to ensure everyone is on the same page.

Expected date of completion: 1/11/2021

Sub Task	**Estimated Time**	**Completion Status**	**Assigned To**
Introduction Slides	*4hrs*	*Incomplete*	*Keith-Jordan, Cory*
Data Sources/Process Slides	*4hrs*	*Incomplete*	*Cory*
Visualizations	*3hrs*	*Incomplete*	*Everyone*
Machine Learning Slides	*4hrs*	*Incomplete*	*Kai, Eddy*
Conclusion Slides	*4hrs*	*Incomplete*	*Keith-Jordan*
Slide Review	*1hr*	*Incomplete*	*Everyone*
Practice Presentation 1	*1hr*	*Incomplete*	*Everyone*
Practice Presentation 2	*1hr*	*Incomplete*	*Everyone*

Present to the Cohort our findings!
Expected date of completion: 11/12/2021