SP-24-X-DATA MINING

DISEASE PREDICTION SYSTEM: CANCER AND DIABETES

FINAL REPORT

CS 4850 - Section 01 – Spring 2024

Professor Sharon Perry

Number of lines in source code:

Website:

Github link:

March 2024

Team Members:

Name	**Role**	**Cell Number**	**Email**
Alexus Glass	Developer	4049601998	lushgaloreinc2@gmail.com
Samuel Futral	Documentation	7067162064	samuelfutral@gmail.com
Kokou Adje	Developer	4703583315	adjelaime@gmail.com
Keegan Begley	Lead	6783667862	kbegley@students.kennesaw.edu
Sharon Perry	Project Advisor	7703293895	sperry46@kennesaw.edu

Table of Contents

1.0 Abstract

The data mining project is a web application designed with a machine learning-powered platform implemented to help with cancer and diabetics prediction. This project aims to assist individuals and healthcare professionals to predict the probability of cancer and diabetics using data based on health information. The project is a combination of machine learning models and algorithms, data mining, and web development using Python programming language.

The objective of this project is to design machine learning models capable of predicting cancer and diabetes using patient’s health indicator information provided such as age, blood pressure, breast density, glucose level, etc. Secondly, develop easy user-friendly interface web application to gather input data from the user, and display the result in a real time manner. Finally, implement a strong data mining and processing system to guarantee data and prediction accuracy.

The application begins with the login into the system and process input user data and then preprocesses this data to be analyzed for machine learning. Logistic Regression generates cancer and diabetics predictions using trained classification models and the result is displayed to the user.

This project aims to provide a good and rapid predictive tool which is a complement for traditional cancer and diabetics diagnostic methods. In addition, the scope of this project is for health care providers, healthcare professionals, and serves as educational resources for people who want to learn more about the relationship between healthcare, intelligence artificial, data mining in the prediction of cancer and diabetics.

1.2 Project Goals

First, the objective of this project is to design machine learning models capable of predicting cancer and diabetes using patient’s health indicator information provided such as age, blood pressure, breast density, glucose level, etc. Secondly, develop easy user-friendly interface web application to gather input data from the user, and display the result in a real time manner. Finally, implement a strong data mining and processing to guarantee data and prediction accuracy.

The goal of this document is to define the goal of SRS documentation and specify the non-functional and functional requirements for this project, and then present some user interface (UI), data flow diagram.

1.3 Definitions and Acronyms

Element	Description
UI	User Interface
OS	Operating systems
GPU	Graphics Processing Unit
GB	Gigabytes
DRAM	Dynamic Random-Access Memory

WCAG	Web Content Accessibility Guidelines
HIPAA	Health Insurance Portability and Accountability Act
BMI	Body Mass Index
HTTPS	Hypertext Transfer Protocol Secure
Data mining	Extraction of different types (pattern, correlation, insight) of data from large dataset
Machine learning	Part of artificial Intelligence to develop models which can learn from extracted data to make accurate prediction

2.0 Design Constraints

The implementation of this project was impacted by some design constraints such as environment, user characteristic, and system requirements.

2.1 Environment

Environmental constraints or resource constraints regroup hardware condition and limitations, software dependencies, and OS compatibility. The system needs to be compatible with laptops, desktops, current and previous OS. Another constraint is software dependencies where the system must be compatible with modern browsers as Google Chrome, Microsoft Edge, Mozilla Firefox, Opera. The software libraries (example of Python libraries used in machine learning) must also have a compatibility with the chosen IDE (PyCharm).

The table below resumes some limitations for the environment.

CPU	RAM(GB)	OS(GB)	External Disk	OS type	Browser
10	128	200	500	Windows	Google Chrome
14	1000	200	1000	Linus	Microsoft Edge

2.2 User Characteristics

User characteristic is the capability of user to interact friendly with the system. For this project, the application should provide accommodation for the users according to the degree of their technical knowledge. The navigation in the system should be easy.

The application should be accessible for the user with disabilities in compliance with accessibility standards (WCAG). For accessibility requirements, the application should consider text size, navigation with keyboard, screen readability, and color contrast.

The system must be able to ensure the input data accuracy by controlling the user input and handling errors and validating the input.

2.3 System

The implementation of this application has some system constraints such as: integration with other systems which are externs, performance, security, privacy. The application must use machine learning to predict cancer or diabetics with the possibility to extract data from public databases. This integration with the external system (e.g. connection with public database) should be possible although data format compatibility complexity. This project involves big data manipulation, and the system should be able to handle those data processing. For security purposes, the application should comply with data security and privacy regulations (HIPAA) by encrypting protocols and securing the transmission methods. The machine learning models chosen to train to predict the accurate result should endure serious testing and validation processes.

3.0 Functional Requirements

Functional requirements, also called behavioral requirements, are the requirements of the system's external observable behavior and focus on specific features, interactions, input, output. For this project, the functional requirements are its functionalities that help users to input data, process those data with the model of machine learning, and then make accurate prediction.

3.1 Authentication

To use the system, the users need to provide login information to have access to either cancer prediction page or diabetics prediction page, and then display the result.

3.2 Display cancer or diabetic prediction page

After successful login, and according to the user's choice, cancer or diabetics page will be displayed. On the cancer page, the user must fill the following information: age, tumor size, menopause, Tumor type, breast density. To finish, the user needs to click on the predict button to see the result. For diabetics prediction page, the user needs to input the following information: glucose level, blood pressure, BMI, and age, and then display the result by clicking on the predict button.

- Cancer prediction information

Input	Description
Age	Significant factor for risk of cancer. Some cancers are dangerous in a specific age category.
Tumor size	Measured in cm, it is key factor for cancer diagnostic
Menopause	Menopause status has a huge impact of risk of cancer, example breast cancer
Tumor type	No cancer, invasive cancer, and on-invasive cancer and impact the tumor behavior
Breast Density	It is a risk for cancer detection and is either low density or high density

- Diabetics prediction information

Input	Description
Glucose Level	The key factor in diabetic diagnostic is the glucose level and high level is a red flag
Blood	High blood pression means high risk of diabetic

pressure	
BMI	It is a value obtained from height and weight. Higher BMI means risk of diabetics
Age	Significant factor for risk of diabetic.

3.3 Data processing

Health data provided by users are processed in this phase. This data is cleaned using the method of data cleaning and is ready for machine learning models.

3.4 Integration of machine learning

In this project, artificial intelligence such as machine learning is used alongside data mining methods to retrieve insights and patterns and then make predictions using public healthcare databases. The machine learning algorithm used in this project is Logistic Regression (Supervised Learning for classification). The models are trained on labeled data and are validated with cross-validation. After the training and the validation, the models are used for the prediction of cancer and diabetes.

4.0 Non-Functional Requirements

4.1 Security

The security of data in this project needs to be guaranteed. The data must be encrypted because it is a healthcare dataset and during the transmission of data, HTTPS protocols will be used to keep the confidentiality. The access to the software is secured by authentication (login and password); also, access to the database requires a role-based access control (RBAC). For data privacy, the application should comply with data security and privacy regulations (HIPAA) by encrypting protocols and securing the transmission methods.

4.2 Capacity

The capacity requirement for this project consists of handling the load of information, data, and resources dedicated to the system to meet the performance desired. The volume of data, the hardware, the storage capacity, and the database capacity need to be evaluated. For example, we need to evaluate the amount of information or data which will be processed by the system without impact on its performance.

4.3 Usability

The usability requirement involves the design of user interface and how ease to interact with the system. The user interface must be user-friendly, intuitive and must be available and usable on different OS. The application should make the navigation between pages clear and easy.

4.4 Other

The system should be extensible and reliable. The system must allow the implementation of or integration of new features, new functionality without decreasing either its performance or capacity. The system must implement fault tolerance by handling potential errors during the system utilization.

5.0 Design constraints

The design conception and implementation of this project was impacted by some design constraints we will describe in the following section.

5.1 User authentication

The use of the system requires authentication where the user must provide their login information. The user authentication is implemented in Python with Flask-login extension which secures the data and use session.

5.2 Data processing

Data processing is a key aspect of software design for software projects which handle large amounts of data, in our case, the data from diverse public health databases. For our web application, data processing involves analysis of user’s input and then prepare them for the machine learning models for classification tasks. Panda, a python extension is used for data preprocessing and handle missing values from the extraction to prepare accurate and usable data for machine learning models.

5.3 Use of machine learning

In this step, scikit-learn library is used to demonstrate how the models trained for classification tasks are incorporated in the application. This library also implements the functions for the prediction tasks for cancer and diabetes.

5.4 User interfaces

The conception of the user interfaces needs to comply with the WCAG guidelines for usability and accessibility.

6.0 System architecture

This part of SDD document gives an overview of the system architecture and the description is detailed in the following session for the web application.

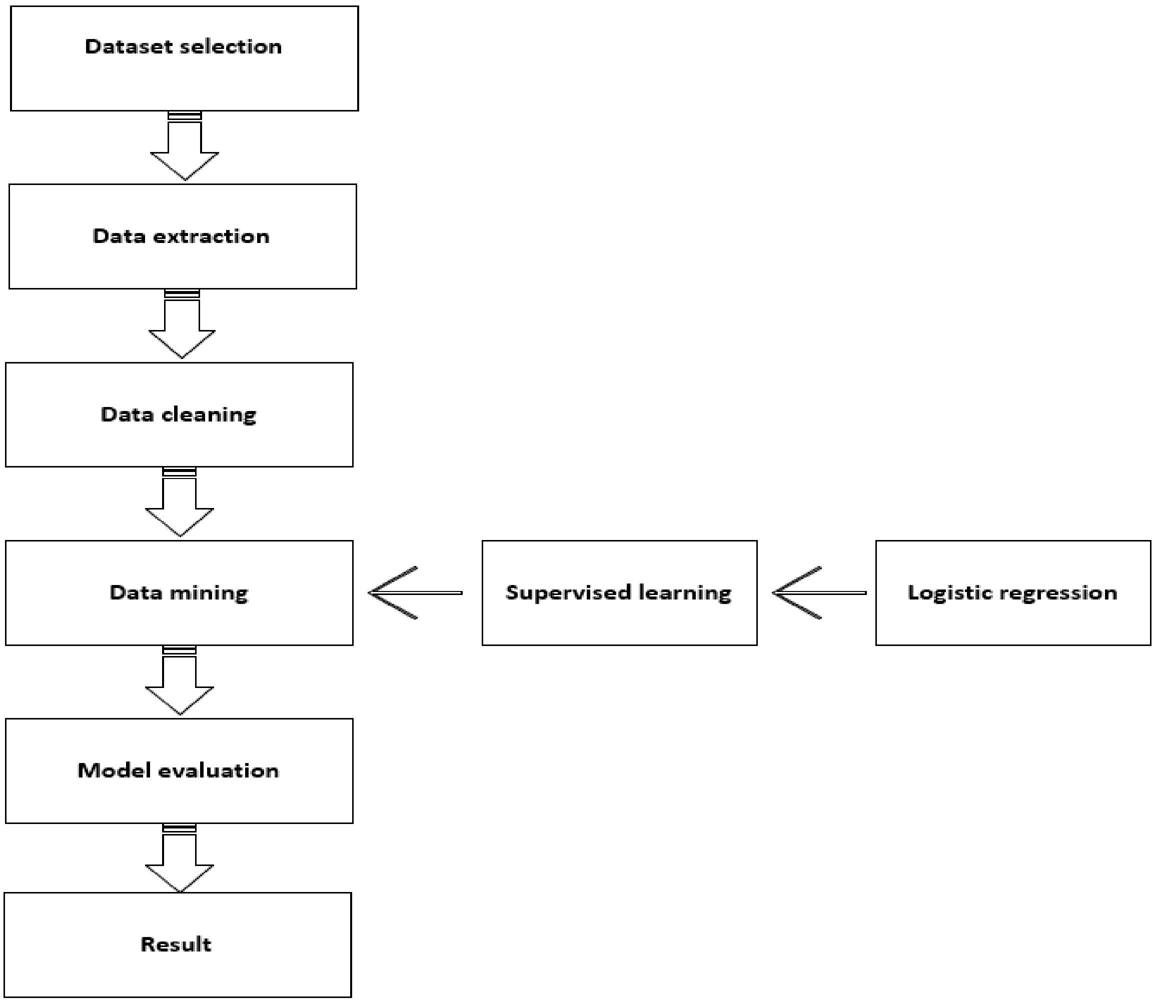
6.1 Front-end elements

Front-end elements allow interaction between the client side and the user. These elements are : HTML (HyperText Markup Language), CSS Stylesheets, JavaScript.

6.2 Back-end element

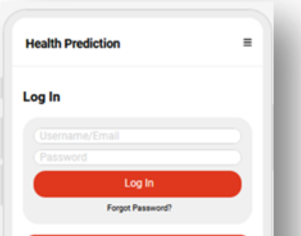
Back-end elements allow interaction between the client side and the server side. The elements used are the following: Flask Framework (provide URL mapping, request handling, routing), Machine Learning Models (Logistic Regression in our case for prediction and classification tasks, scikit-learn to train models), database connectivity (SQLAlchemy ORM connector is used to facilitate the communication between public databases and the software)

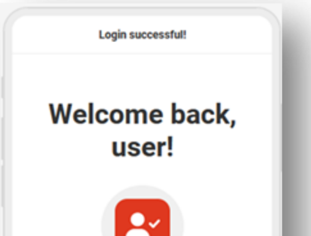
6.3 Architecture Diagram



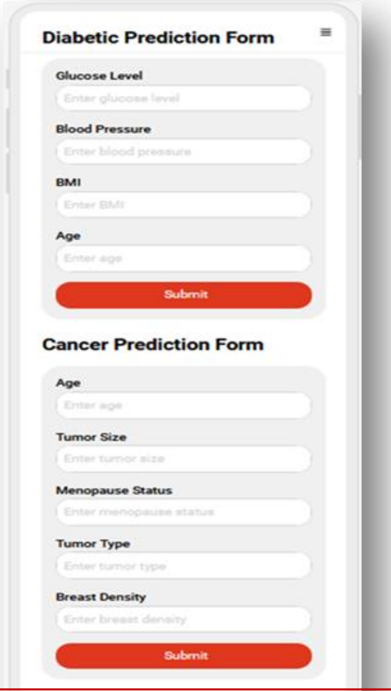
6.4 Design drawings

The following are the descriptions of how the software works.




Login successful!

Welcome back, user!



Diabetic Prediction Form

Glucose Level

Enter glucose level

Blood Pressure

Enter blood pressure

BMI

Enter BMI

Age

Enter age

Cancer Prediction Form

Age

Enter age

Tumor Size

Enter tumor size

Menopause Status

Enter menopause status

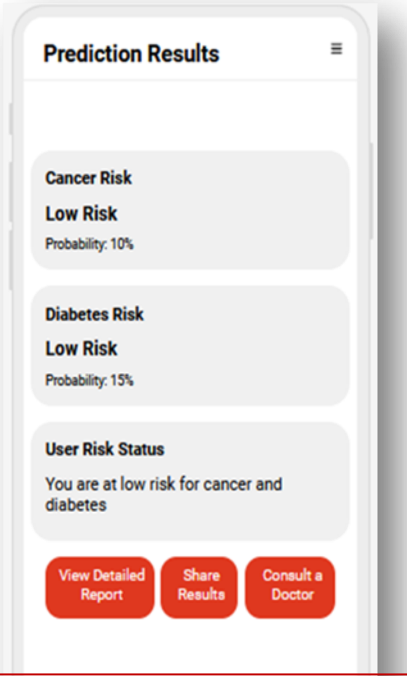
Tumor Type

Enter tumor type

Breast Density

Enter breast density

Prediction forms for the user to input their data to get their medical analysis performed. One for cancer and one for diabetes.



Prediction Results

Cancer Risk

Low Risk

Probability: 10%

Diabetes Risk

Low Risk

Probability: 15%

User Risk Status

You are at low risk for cancer and diabetes

The results page will show after the forms are submitted and compared to machine learning model that was fed research and studies.

7.0 Version control description

This section covers the setup, branching, committing, and release management using Git for a cancer and diabetic prediction web application for the version control.

8.0 Challenges, assumptions, and risk assessments

The implementation of this project involves some challenges, assumptions, and risk assessments.

8.1 Challenges

The first challenge is data availability and quality. Medical data access can be difficult because of privacy and regulatory exigences and restriction. For example, we need to adhere to HIPAA regulations. It is also difficult to adapt the data from datasets to our format. The second challenge is the machine learning models accuracy; to have a good prediction, it is crucial to have accurate data to train those models. In addition, there is user interface and experience challenge we face because the software must be accessible for all kinds of users including users with disabilities and healthcare professionals.

8.2 Assumptions

There are divers assumptions for this cancer and diabetic prediction web application. One assumption is about data access where we assume that data is available for the training of machine learning models. We also assume that healthcare providers have the technologies required for this application. For example, we assume that the users (healthcare providers and patients) have computers to use this application. We assume also that we have regulatory approval for meeting the necessary regulatory requirements to use this software in healthcare.

9.0 Test plan and Test report

10.0 Summary

Cancer and Diabetic Prediction Web Application project provides a strong tool for healthcare providers and professionals to predict and detect early cancer and diabetic from patients. The prediction is done with artificial intelligence by machine learning models. This project has many requirements and designs. A version control is integrated to control the version of the application. A test plan and report are written for the project. This project is implemented without some challenges and assumptions. We are grateful for the hard work of the team for the time dedicated to the success of this project.

10.0 Appendix

10.1 Project Plan

1. Platform:

For this project, we will use a hardware and software platform. The right hardware needs to be chosen to properly run the machine learning program and must impact the quality and the performance of the models. The hardware requirements are the following for a computer to operate this application:

- ✓ Adequate processor for the execution of instructions.
- ✓ Good storage and memory to store a large amount of data from the processor.
- ✓ Graphics Processing Units (GPUs) to handle graphical data.
- ✓ For future utilization, high-performance computing resources like cloud infrastructure can be used for the software.

For the software infrastructure, the followings are required:

- ✓ Python: is important for the implementation of this project.
- ✓ Large Language Model (LLM): a kind of AI trained to recognize and generate text.
- ✓ Web frameworks: Flask, Django which offer some features for the application.

2. Collaboration Tools:

Communication	—	Cellphones (Call/Text) / Microsoft Teams / GroupMe
Collaboration	—	Discord (Mandatory unless another tool is authorized by Perry)
Version Control	—	GitHub

3. Deliverables:

a. Team/Project Selection document (Individual Assignment)
b. Weekly Activity Reports (WARs – Individual Assignment)
c. Peer Reviews (Individual Assignment)
d. Project Plan (Group Assignment)
e. Project Requirements And Design
f. Present Prototype for Peer Review (Group Assignment)
g. Website (Group Assignment)
h. Video Demo (Group Assignment)
i. C-Day Application/Submission (Group applies to C-Day but each member

submits individual bonus points documentation in Individual Assignments)

j. Final Report Package (Group Assignment)
k. iOS and Android compatible mobile time travel apps

4. Project Schedule and Task Planning (GANTT CHART)

02/21/2024

- Research dataset
- Create graphs
- Data analysis

02/28/2024

- Data mining
- Organize Data Table
- Data analysis
- Start on System Design

03/06/2024

- System Design & architecture
- Research
- Python Coding & SQL

03/13/2024

- System Design & Architecture
- Python Coding & SQL

03/20/2024

- Documentation & Report
- Python Coding & System structure

03/27/2024

- Documentation Report

04/03/2024

- Documentation & Code implementation

04/10/2024

- Finish Final report draft

04/17/2024

- Make sure the project is precise, clean and easy to understand.

04/21/2024

- Turn in the project.

5. Meeting Schedule

The team will be meeting on Mondays and Wednesdays after class at 8pm. The length of the meetings will be an hour typically, with the option of additional time if needed. During these meetings we will discuss our current focus, allocate tasks, and update project schedules based on progress status.

6. Statement of Participation

Project ID: SP-24

STATEMENT OF PARTICIPATION:

By signing below, I Kokou Adje acknowledge that I will participate in all meetings, communications, deliverables, and other tasks necessary to complete the project. If I do not, I understand that Professor Perry will meet with me to remedy the situation.

Kokou Adje 3/7/2024

________________________________ __________

Team Member Date

By signing below, I Hailey Walker acknowledge that I will participate in all meetings, communications, deliverables, and other tasks necessary to complete the project. If I do not, I understand that Professor Perry will meet with me to remedy the situation.

Alexus Glass 3/7/2024

________________________________ __________

Team Member Date

By signing below, I William Stigall acknowledge that I will participate in all meetings, communications, deliverables and other tasks necessary to complete the project. If I do not, I understand that Professor Perry will meet with me to remedy the situation.

Samuel Futral 3/7/2024

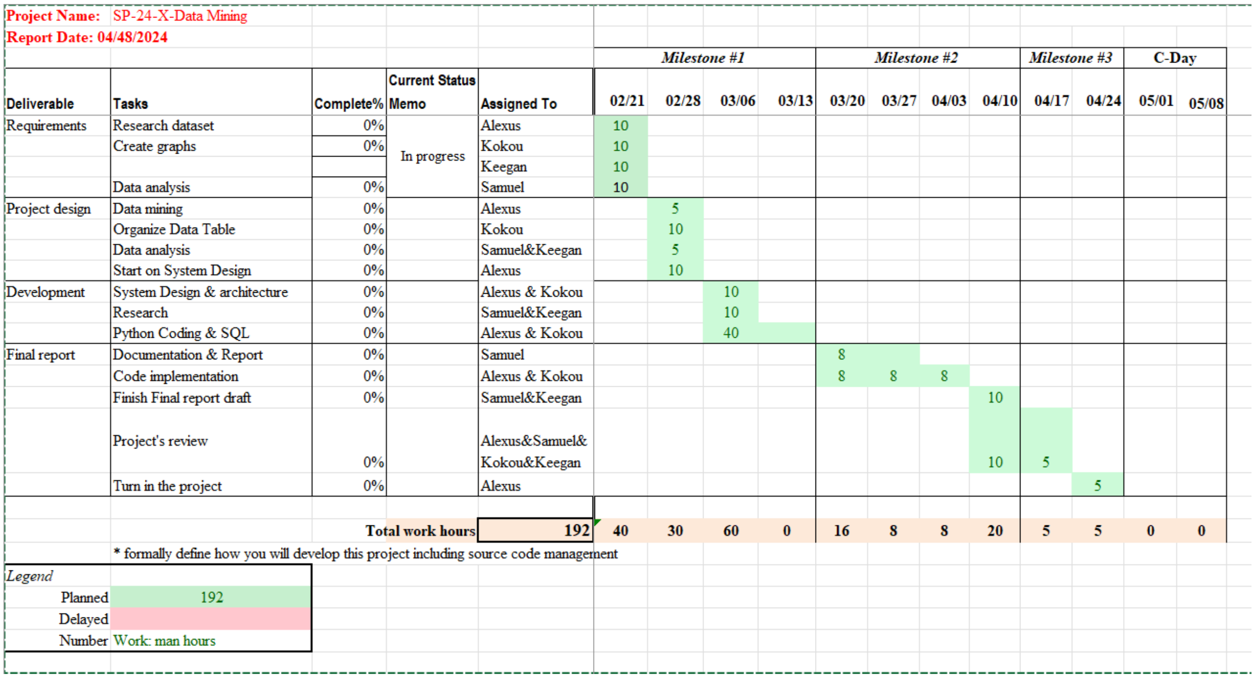
Team Member Date

By signing below, I William Stigall acknowledge that I will participate in all meetings, communications, deliverables, and other tasks necessary to complete the project. If I do not, I understand that Professor Perry will meet with me to remedy the situation.

Keegan Begley 3/7/2024

Team Member Date

10.2 Gant Chartt



Project Name: SP-24-X-Data Mining

Report Date: 04/48/2024

					Milestone #1				*Milestone #2*				*Milestone #3*		C-Day	
Deliverable	**Tasks**	**Complete%**	**Current Status Memo**	**Assigned To**	**02/21**	**02/28**	**03/06**	**03/13**	**03/20**	**03/27**	**04/03**	**04/10**	**04/17**	**04/24**	**05/01**	**05/08**
Requirements	Research dataset	0%	In progress	Alexus	10											
	Create graphs	0%		Kokou	10											
				Keegan	10											
	Data analysis	0%		Samuel	10											
Project design	Data mining	0%		Alexus		5										
	Organize Data Table	0%		Kokou		10										
	Data analysis	0%		Samuel&Keegan		5										
	Start on System Design	0%		Alexus		10										
Development	System Design & architecture	0%		Alexus & Kokou			10									
	Research	0%		Samuel&Keegan			10									
	Python Coding & SQL	0%		Alexus & Kokou			40									
Final report	Documentation & Report	0%		Samuel					8							
	Code implementation	0%		Alexus & Kokou					8	8	8					
	Finish Final report draft	0%		Samuel&Keegan								10				
	Project's review	0%		Alexus&Samuel& Kokou&Keegan								10	5			
	Turn in the project	0%		Alexus										5		
			Total work hours	**192**	**40**	**30**	**60**	**0**	**16**	**8**	**8**	**20**	**5**	**5**	**0**	**0**

* formally define how you will develop this project including source code management

Legend	
Planned	🟩 192
Delayed	🟥
Number	Work: man hours