**Computer Science Capstone Topic Approval Form**

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of these areas so that you will have a complete and realistic overview of your project. Your course instructor cannot sign off on your project topic without this information*.*

*Note: You must fill out and submit this form. Space beneath each number will expand as needed.*

*Any cost associated with developing the application will be the responsibility of the student.*

**INFORM INSTRUCTOR:**

Potential use of proprietary company information: (Y/N)

**NO**

**ANALYSIS:**

1. Project topic AND description: I Intend to use a publicly available data set that contains 768 records on type two diabetes. It will contain numerous features and the target will be whether or not a person is likely to have type 2 diabetes, and therefore is a classification project. Fortunately, no personal identify information is not attached, and the study comes from a reliable source (John Hopkins). I would like to implement a basic website where a person will be able to enter their own health data and have the machine learning model output the likelihood of them having diabetes.
2. Project purpose/goals:

My goal is for my model to be able to determine whether someone has diabetes and scoring above 70% in accuracy. I will also be tracking precision, recall, and f1 scores My hope is to actually score above 80% in all these metrics, as I would consider that a complete success and I would consider the model to be viable in a real world application.

1. Descriptive method: Data contains 768 records with the following features: number of pregnancies, glucose, blood pressure, skin fold thickness of tricep, insulin levels, body mass index, diabetes pedigree function, age, and outcome. Outcome is whether or not the patient had/has diabetes. The data set contains over 300 records with complete records, meaning no features have missing values. I will try a model with only complete records and then compare it to another model by imputing the mean value of that feature to replace that missing value and then compare the two models to see which performs better. This will be a classification model.
2. Predictive/Prescriptive method: I will be using a few different models to see which model yields the best results. At this point, from my research, it is likely that a RandomForestClassifier will more than likely give me the best results, so I will start there. If I’m not impressed with this model after it’s tuned, I will be using Naïve Bayes, Logistic Regression, and GradientBoostingClassifier from the SickitLearn library to see if I can outperform it.

**DESIGN and DEVELOPMENT:**

1. Computer science application type (select one):

* Mobile (indicate Apple or Android)
* Web
* Stand-Alone
* Web application hosting a Jupyter notebook

1. Programming/development language(s) you will use:
   1. Python
      1. Jupyter Notebook
      2. Matplotlib
      3. Numpy
      4. Conda environment
      5. Seaborn
      6. Scikit Learn
   2. Javascript
2. Operating System(s)/Platform(s) you will use:
   1. Windows 10
   2. Chrome
3. Database Management System you will use:

N/A

1. Estimated number of hours for the following:
   * 1. Planning and Design:
2. 60 hours to finish Udemy course on machine learning
3. 10 hours to review project requirements and roughly create wireframe for the dataset and consider which models, metrics, and plots/graphs will most helpful.
4. 2 hours to research basic website template to integrate with model.
5. 20 hours to review material needed to implement a website
   * 1. Development:
6. 25 hours to properly vet a model, scale, and tune the hyperameters
7. 10 hours to figure out how to use model with javascript and implement
8. 15 hours to develop website
9. 5 hours to polish website further
   * 1. Documentation:
10. 20 hours to complete documentation requirements for each main section (A,B,C,D)
    * 1. Total: 147 hours
11. Projected completion date: 10/1/2021

**IMPLEMENTATION and EVALUATION:**

1. Describe how you will approach the execution of your project:

To prepare for this project I will first finish a Udemy masterclass on machine learning. After that is complete I will use Jupyter Notebooks to organize my project and upload my dataset. I will then cleanse and normalize the data and begin modeling. I Intend to use 80 percent of the data for training, 10 Percent for validation, and 10 percent for testing. I will then scale and tune hyperameters and choose the model that gives me the best results on the following metrics: accuracy, precision, recall, f1.

After modeling is complete I will need to research how to integrate this with my website and will also need to learn some basic javascript. Once I have that complete I will then look for a basic website template. The website will only need to be 1 to 2 pages with multiple forms for the user to input various vitals, where the user will then hit submit and the model will use that information to give a prediction on whether or not it believes they have type 2 diabetes.

* **This project does not involve human subjects research and is exempt from WGU IRB review.**

**STUDENT SIGNATURE**

**­­­­­­­­­­­­­­­­­­­­­­­­­­­­­\_Kamron Zell\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**By signing and submitting this form, you acknowledge** any cost associated with development and execution of the application will be your (the student) responsibility.

**COURSE INSTRUCTOR’S NAME:**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**COURSE INSTRUCTOR APPROVAL DATE:**