# Week 2 Project Update: Coming Up With Ideas

Matthew Vu

2/1/2023

## An Initial Idea

According to the Centers for Disease Control and Prevention[1], as of September 30, 2022, ”more than 37 million people in the United States have diabetes, and 1 in 5 of them don’t know they have it” (CDC, 2022). This means diabetes remains a widespread issue in the United States, and an estimated 7.4 million of these individuals who have diabetes are unaware that they have it. As a result, a potential topic for this capstone course is to design a modern web-based questionnaire application that uses machine learning to predict if an individual has diabetes based on his or her answers to questions. Additionally, the goal of this questionnaire is that it can be taken at home, and the questions will not ask the user to input any exact measurements that require specialized tools to measure. For example, the questionnaire will not ask for an individual’s exact glucose level because that can only be exactly measured using a glucose meter, which many individuals do not have at home. In the past, researchers have investigated the potential for machine learning algorithms to predict diabetes. For instance, in an article titled ”Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms”[2] researchers Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun experimented with machine learning algorithms (naïve bayes, random forest, and decision trees) to predict diabetes (Chang et. al, 2022). However, upon inspection of their dataset, many of the numerical variables that the models were trained on required special equipment to precisely measure for an individual. For example, the glucose and blood pressure level variables require glucose meters and sphygmomanometers, respectively. Consequently, the dataset that was used to train the model are not well suited for an at-home questionnaire. Nevertheless, in the past, groups have created web-based questionnaire applications that use machine learning algorithms to predict diabetes when receiving very simple input values. For example, Jaana Lindström and Jaako Tuomilehto[3] wrote an academic journal article outlining their research and development of a web-based application questionnaire that used 1992 survey data and logistic regression to predict the risk of diabetes (Lindström & Tuomilehto, 2003). Now, although Jaana Lindström and Jaako Tuomilehto researched and designed a web-based questionnaire application that did not use variables consisting of precise measurements from special tools (like exact numerical glucose levels) or ask users to enter exact measurements (like exact glucose levels), the data that were used for training were from 1992. Also, they only experimented with one machine learning algorithm (logistic regression). Given my previous remarks on past research, the overall plan of my initial project idea is to analyze, train, and experiment with multiple machine learning algorithms (logistic regression, naïve bayes, support vector machines, and random forest) using recent (post-2015) diabetes data. Furthermore, the end product should be a web-based questionnaire application running off the machine learning algorithm that performed the best. Finally, the machine learning algorithm that is eventually chosen and questionnaire questions will be designed for very simple answers that do no require exact measurements from special equipment.

## References

[1] CDC. (2022, September 30). *Diabetes Fast Facts*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/quick-facts.html

[2] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing & Applications*, 1–17. https://doi.org/10.1007/s00521-022-07049-z

[3] Lindström, J. & Tuomilehto, J. (2003). The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk. *Diabetes Care, 26*(3), 725–731. https://doi.org/10.2337/diacare.26.3.725