Group 2 Team Members:

Darren Yu

Fyez Dean

**Project proposal Fall 23’**

1. **Define the problem/ state hypothesis:** 
   * Can we predict the stage of or implication of Diabetes given the subset risk factors within the data set?
2. **Literature Review:** 
   * End-to-End Data Science Example: Predicting Diabetes with Logistic Regression.
     1. Using Python and its respective libraries to predict whether females of age 21 and above from the Pima Indians Diabetes Database have diabetes.
   * Diabetes Prediction Using Machine Learning (Amaan Preeti Gulati)
     1. This article explains how to use machine learning to predict diabetes in individuals using the Pima Indians Diabetes Database.
   * Using Machine Learning to predict if someone has Diabetes (Edward Leoni)
     1. The article explains how to predict if someone has diabetes or not based on 8 variables using Keras, a high-level neural networks API written in Python and TensorFlow, CNTK, or Theano.
3. **Data Collection** : Diabetes Health Indicators (from Kaggle):
   * A three file containing data collected from the Behavioral Risk Factor Surveillance System (BRFSS) through an automated telephone system. The data is diverse as it includes fields such as: Income, Education, Sex, Veggies, Alcohol Consumption, and Exercise, all will be useful in providing accurate predictions.
4. **Methodology**: We will be using Classification techniques along with a simple decision tree model in order to verify if enough risk factors meet the acceptance criteria and percentage of whether or not an individual has diabetes. We could also group individuals based on similar symptoms or individuals who have similar risk factors and use the Clustering technique as a whole to verify the different classes of diabetes that people might have with respect to the grouping they fit in.
5. **Implementation**: We will be using Python to implement the data visualization and predicted outcomes. The different libraries we may use are:
   * Sklearn- a machine learning library that implements classification using algorithms to create decision trees.
   * NumPy: useful for showing spread of data and for providing statistical analysis
   * Pandas: useful for data manipulation and analysis
6. **Evaluation**: We will evaluate the effectiveness of the model based on the following
   * Qualitative analysis and Visualization: we will describe our findings and compare them to our literature review articles to verify if the parameters meet in any area and whether or not our visualizations offer some insight into the effectiveness of the classification prediction.
   * Data Splitting: We can split our data set into three subsets, a training model to initialize our mining technique, then a validation set to test which of the first two models we produce provides more accurate results that we are looking for. Then a final test model that we will report our official findings on and provide insight on the accuracy of our hypothesis.
   * Model Verification: we can test the validity of our findings based on the similarities and differences we find through the two mining techniques we will use: Clustering and Classification
7. **Conclusion**:
   * In conclusion, this proposal outlines a focused research effort to address the critical challenge of predicting diabetes and its associated implications using machine learning techniques. The primary motivation behind this project lies in the need to better understand and anticipate the onset of diabetes, a condition with significant health and societal implications.
   * Our problem statement centers on the feasibility of predicting diabetes based on a subset of risk factors within the dataset. The research draws inspiration and guidance from relevant literature, emphasizing the practical application of machine learning models in similar scenarios.
   * For data collection, we have selected the "Diabetes Health Indicators" dataset from Kaggle, a valuable source of information essential for our analysis.
   * In terms of methodology, we propose to utilize classification techniques, particularly a simple decision tree model, to assess the dataset's risk factors against acceptance criteria for predicting diabetes. Additionally, clustering techniques will be employed to group individuals based on similar symptoms or risk factors, facilitating the identification of distinct diabetes classes.
   * Python will serve as our primary programming language for data visualization and modeling, with essential libraries like Scikit-Learn, NumPy, and Pandas enhancing our analytical capabilities.
   * The evaluation plan is multifaceted. We will engage in qualitative analysis and visualization, validating our findings against existing literature to ensure the effectiveness of our model. Data splitting will enable the creation of training, validation, and test subsets, allowing for the assessment of model accuracy. Model verification will serve to confirm the validity of our results, comparing outcomes from clustering and classification techniques.
   * In summary, our research is driven by the ambition to leverage the potential of machine learning to contribute to the understanding and prediction of diabetes. We believe our comprehensive approach, including data collection, methodology, and evaluation, will lead to valuable insights that can positively impact healthcare and disease prediction practices.
8. **Timeline and Resources**:

Training set preparation: 11/1

* TA or Prof Check 10/25
* Write outline & check documentation 11/2
* Progress Report 11/2

validation set prepared: 11/8

* TA or Prof Check 11/15
* Write outline & check documentation 11/ 13
* Progress Report: 11/15

final test set 11/22

* TA or Prof Check 11/19
* Write outline & check documentation 11/20
* Progress Report: 11/19

Compile official findings: [EVERYTHING BEFORE FINALS WEEK - - Writing Week]

* Compile all the progress reports and notes
* Clarify and organize findings
* Create final abstract
* Write about the significance drawn from data
* Identify sources of discrepancies, flaws, and analyze overall reliability

**References**

Dhandhania, Keshav. “End-to-End Data Science Example: Predicting Diabetes with Logistic Regression.” *Medium*, Towards Data Science, 24 May 2018, towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16.

Gulati, Aman Preet. “Diabetes Prediction Using Machine Learning.” *Analytics Vidhya*, 4 Jan. 2022, [www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/](http://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/)

Leoni, Edward. “Using Machine Learning to Predict If Someone Has Diabetes.” *Medium*, Learn stuff with Ed, 12 Apr. 2020, medium.com/edward-leoni/using-machine-learning-to-predict-if-someone-has-diabetes-21dc52118d8f.

Teboul, Alex. “Diabetes Health Indicators Dataset.” *Kaggle*, 8 Nov. 2021, [www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset](http://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset).