**Project II: Machine Learning**

Pick a readily available dataset of your choice to build a regression or classification model. Write a report with the following details:

describe the problem with proper literature review (10 points)

provide a descriptive analysis of the data set including visualization (5 points)

build at least three machine learning models and compare the performances of these models (10 points)

Write a detailed description of the evaluation metrics used and the data splitting techniques (10 points)

provide a thorough error analysis and propose recommendations for improving the models (5 points)

using the principles of responsible artificial intelligence discussed during the tutorials, discuss how your project meets responsible AI guidelines.

**Predicting diabetics using related health conditions and lifestyles Factors**

# Introduction

Diabetes mellitus, often referred to simply as diabetes, encompasses a set of metabolic disorders marked by elevated blood sugar levels. This condition can precipitate severe long-term complications such as cardiovascular issues, stroke, kidney failure, heart attacks, peripheral arterial disease, and nerve damage (Maniruzzaman et al., 2020). In 1980, approximately 122 million individuals globally were affected by diabetes, a number that surged to about 422 million by 2014. Projections suggest that by 2040 this figure will rise to about 642 million (Zimmet et al., 2016). Additionally, there were roughly 1.6 million diabetes-related deaths directly attributed to the condition.

Data science solutions have catalysed a revolution in the healthcare sector, facilitating advancements in areas like drug discovery and understanding genetic diseases. Consequently, there remains substantial untapped potential in this domain that warrants further exploration for the betterment of society. Machine learning offers an array of algorithms capable of delivering heightened accuracy in predicting outcomes based on input data, bolstered by statistical analysis (Kumar et al., 2019).

Examining diabetes data poses a formidable challenge due to the nonlinear, nonnormal, correlation-structured, and complex nature of most medical data sets (Maniruzzaman et al., 2020). Furthermore, it aids individuals in accurately diagnosing diabetes, with identifying the most effective classifier being crucial for precise diabetes risk assessment.

In the medical realm, there is a significant focus on preventing diabetes. Leveraging data mining expedites the analysis of data, allowing analysts to scrutinize existing datasets to uncover patterns and trends associated with diabetes (Llaha & Rista, 2021). Through employing data mining techniques, physicians can predict illnesses with increased effectiveness, enabling them to better handle individuals deemed at high risk (Bisandu, 2019). The abundant data on diseases and the intricate interrelationships among them render medicine a suitable arena for the application of data mining techniques. Data mining enables the scrutiny of numerous extensive datasets encompassing a vast array of variables, surpassing the capacity of a single analyst, doctor, or even an analytical team.

Literature Review

# 3. Materials and Methods

**3.1. Data collection:**

Data was collected from

**3.2. Data Splitting Techniques:**

# Results

## 4.1 Descriptive analysis of the data set including visualization

# References

Kumar, Y. J. N., Shalini, N. K., Abhilash, P. K., Sandeep, K., & Indira, D. (2019). *Prediction of Diabetes using Machine Learning*. *8*(7).

Llaha, O., & Rista, A. (2021). *Prediction and Detection of Diabetes using Machine Learning*.

Maniruzzaman, Md., Rahman, Md. J., Ahammed, B., & Abedin, Md. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, *8*(1), 7. https://doi.org/10.1007/s13755-019-0095-z

NCD Risk Factor Collaboration (NCD-RisC). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 192 million participants. Lancet. 2016;387(10026):1377–96.

Zimmet, Paul, et al. "Diabetes mellitus statistics on prevalence and mortality: facts and fallacies." Nature Reviews Endocrinology 12.10 (2016): 616-622.

Bisandu, D. B., Datiri, D. D., Onokpasa, E., Thomas, G., Haruna, M. M., Aliyu, A., & Yakubu, J. Z. (2019). Diabetes prediction using data mining techniques. International Journal of Research and Innovation in Applied Science (IJRIAS), 4(6), 103-111.