


Faculty of Engineering
Cairo University

# Machine Intelligence

#CMP4040

## Project Proposal

## Team 18

*Submitted to:*

*Eng. Mohamed Shawky*

*Submitted By:*


| NAME | SEC | BN | ID |
|---|---|---|---|
| Sarah Elzayat | 1 | 29 | 9202618 |
| Abdelrahman Fathy | 2 | 2 | 9202846 |
| Yasmine Ashraf Ghanem | 2 | 37 | 9203707 |
| Yasmin Abdullah Nasser | 2 | 38 | 9203717 |

## Selected Problem 1: Diabetes Health Indicators

- *Definition:*

  The dataset consists of health indicators collected from the Behavioral Risk Factor Surveillance System, focusing on diabetes. It includes a wide range of variables, such as health behaviors, health outcomes, and the use of preventive services. This dataset offers a comprehensive view of factors that may influence diabetes conditions in individuals.

- *Motivation:*

  Diabetes is a growing health concern worldwide, leading to various complications and affecting millions of people's quality of life. Understanding the factors that contribute to diabetes can help in early detection, prevention, and management strategies. By analyzing this dataset, we aim to uncover patterns and relationships between different health indicators and diabetes status. Insights derived from this analysis could inform public health policies, individual lifestyle choices, and medical interventions aimed at reducing the incidence of diabetes and improving the lives of those living with the condition. Specifically, we could identify high-risk groups based on demographics or behaviors, determine key factors contributing to diabetes, and suggest targeted interventions to mitigate these risks.

## References

*Dataset: Diabetes Health Indicators Dataset (kaggle.com)*
CDC Diabetes Health Indicators - UCI Machine Learning Repository

- *More information:*

  Features: *21*
  Instances: *253,680*

## Selected Problem 2: Students Dropout Prediction

- *Definition:*

  The dataset compiles data aimed at understanding factors that predict students' academic outcomes, specifically focusing on dropout rates and success metrics. This dataset includes variables like students' demographics, academic performance, engagement metrics, and other relevant indicators that could influence their academic journeys.

- *Motivation:*

  Education is a cornerstone for individual development and societal progress. However, student dropout rates pose a significant challenge, impacting individuals' future opportunities and broader societal well-being. By analyzing the dataset, the goal is to identify predictors of academic success and risk factors for dropping out. Such insights can empower educators, policymakers, and institutions to implement targeted interventions designed to support students at risk of dropping out, enhance academic achievement, and ultimately, reduce dropout rates.

## References

*Dataset: Predict Students' Dropout and Academic Success - UCI Machine Learning Repository*

- *More information:*

  Features: *36*
  Instances: *4,424*

## Selected Problem 3: PhiUSIIL Phishing URL (Website)

- *Definition:*

  The dataset is designed for the detection and analysis of phishing URLs. Phishing attacks involve malicious websites pretending to be legitimate with the aim of deceiving individuals into proclaiming personal and sensitive information. It contains various features of URLs, including the lexical characteristics of the web addresses, website content features, and host-based features. This can give us a comprehensive framework for distinguishing between phishing and legitimate websites.

- *Motivation:*

  In the digital current age, cybersecurity is of great importance. Phishing attacks represent a significant threat, leading to severe financial losses and breaches of personal privacy. The motivation lies in developing effective, intelligent systems capable of identifying and blocking phishing attempts before they reach end-users. By using machine learning techniques to understand and predict phishing URLs, we can significantly enhance online security measures, protect individuals' sensitive information, and mitigate the risks associated with these deceptive practices.

  Then develop more robust cybersecurity tools, educating users about the characteristics of phishing attempts, and ultimately, creating a safer online environment for everyone.

## References

*Dataset: PhiUSIIL Phishing URL (Website) - UCI Machine Learning Repository*

- *More information:*

  Features: *54*
  Instances: *235,795*

## Evaluation Metrics for all ideas

Since all selected ideas are classification problems, the following metrics can be applied to all:

1. *Accuracy:*
   The fraction of predictions our model got right. It's a good starting point but doesn't work well with imbalanced datasets.

2. *Precision and Recall:*
   Precision measures the accuracy of positive predictions, while recall measures the fraction of positives that were correctly identified.

3. *F1 Score:*
   The harmonic mean of precision and recall. It combines both metrics into one, balancing their contributions.

4. *ROC-AUC:*
   The area under the Receiver Operating Characteristic curve. It measures the model's ability to distinguish between classes.

*Those metrics are subject to modification and change depending on the final analysis.*