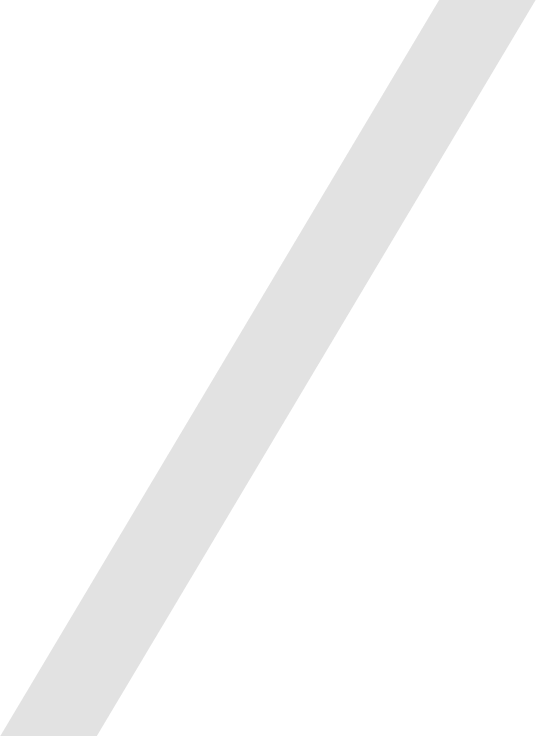
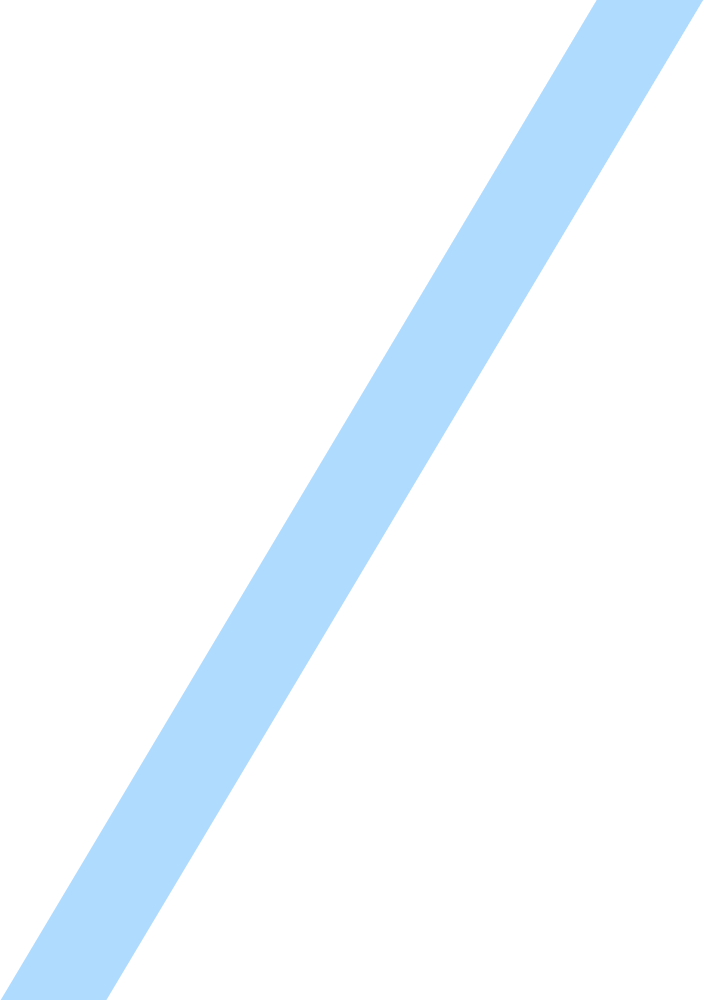
|  |
| --- |
| TECHNICAL REPORT |

|  |
| --- |
| Data Science |

|  |  |
| --- | --- |
| SPRING 22 |  |



Contents

Project Name…………………………………………..2

[Executive Summary 3](#_Toc96341550)

[Technical Report 4](#_Toc96341551)

[Highlights of Project 5](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 5](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 6](#_Toc96341554)

[Methodology 9](#_Toc96341555)

[Results Section 9](#_Toc96341556)

[Discussion 12](#_Toc96341557)

[Conclusion 13](#_Toc96341558)

[Contributions/References 13](#_Toc96341559)

Diabetes Prediction

Using Machine Learning

|  |  |  |
| --- | --- | --- |
| Executive Summary Medical Science is one of the sectors which uses Machine Learning very frequently at its core. From diagnosis and prognosis to drug development, ML is infiltrating this sector to a great extent. Our Business Problem belongs to Medical Science domain, that is; predicting if a patient is diabetic or not. About 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.5 million deaths are directly attributed to diabetes each year. This fact is more than sufficient to build a production level system which can help actual patients who might be having Diabetes and not knowing it. So, let’s clearly state our objective. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Name 1: Siva Chandan Chakka**  **Name 2: Sun Gajiwala**  **Name 3: Shriya Raj Mahan** | **Questions?**  Contact:  schak5@unh.newhaven.edu  [sgaji1@unh.newhaven.edu](mailto:sgaji1@unh.newhaven.edu)  srajm1@unh.newhaven.edu |  |

|  |
| --- |
|  |

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
| Highlights of Project In this project we use four different machine learning models. The model was then deployed using Flask.    We divided the project into three phases:  Phase 1: Research and Information gathering.  Phase 2: Implementation using python  Phase 3: Documentation    Till Date, all the phases of the project have been achieved, the project is 100% completed and the desired results have been obtained.   Submitted on: 05/03/2022 |  |

## 

## Abstract

Diabetes is a disease that has no permanent cure; hence early detection is required. [Data mining](https://www.sciencedirect.com/topics/computer-science/data-mining), machine learning (ML) algorithms methods are used in diabetes prediction in our research. We used the Pima Indian Diabetes (PID) dataset for our research, collected from Kaggle. The dataset contains information about 768 patients and their corresponding nine unique attributes. We used seven ML algorithms on the dataset to predict diabetes. We found that the model with [Logistic Regression](https://www.sciencedirect.com/topics/computer-science/logistic-regression) (LR), K-Nearest Neighbors(KNN), Support Vector Machines(SVM) and Random Forest Classifier (RFC) worked well for the considered dataset.

Cover Page

PROJECT TITLE:  Diabetes prediction application using machine learning.

**AUTHORS:**

Siva Chandan Chakka ([schak5@unh.newhaven.edu](mailto:schak5@unh.newhaven.edu))

Sun Gajiwala ([sgaji1@unh.newhaven.edu](mailto:sgaji1@unh.newhaven.edu))

Shriya Raj Mahan ([srajm1@unh.newhaven.edu](mailto:srajm1@unh.newhaven.edu))

Introductory Section

GITHUB LINK: <https://github.com/sivachandan1996/DistributedSystemProject>

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose, which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. The most common is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin. In the past three decades the prevalence of type 2 diabetes has risen dramatically in countries of all income levels.

Predicting the probability of an individual's risk and susceptibility to a chronic illness like diabetes is an important task. Diagnosing chronic illness at an early stage saves on medical costs and reduces the risk of more complicated health problems. Even in emergencies where a patient may be unconscious or unintelligible, it is pertinent that deductions can be made accurately from immediately measurable medical indicators to help clinicians make better decisions for patient treatment in high-risk situations.

Data mining and Machine learning have been developing, reliable, and supporting tools in the medical domain in recent years. The [data mining method](https://www.sciencedirect.com/topics/computer-science/data-mining-techniques) is used to preprocess and select the relevant features from the healthcare data, and the [machine learning method](https://www.sciencedirect.com/topics/engineering/machine-learning-method) helps automate diabetes prediction. Data mining and [machine learning algorithms](https://www.sciencedirect.com/topics/engineering/machine-learning-algorithm) can help identify the hidden pattern of data using the cutting-edge method; hence, a reliable accuracy decision is possible.

Review of available research

Nicos Maglaveras and Ioannis [1]: The dataset was evaluated in two phases, 1) Biomarker identification and Prediction of Diabetes mellitus, 2) disease prediction takes place, although this task is always performed to evaluate the predictive accuracy of the identified biomarkers. Hence, in the framework of this study, efforts were made to review the current literature on machine learning and data mining approaches in diabetes research.

Varun Jaiswal and Anjeli [2]: This paper is an effort to summarize the majority of the literature concerned with machine learning and data mining techniques applied for the prediction of diabetes and associated challenges. Reinforcement learning was used to determine the accuracy of prediction of the model. The Model

## 

Arianna Dagliati and Simone Marini [3]: Data mining strategies can be also used to provide new predictive models that, starting from already available risk prediction calculators, may be fused with the data available at a single clinical site to effectively support disease management and patient care. The analysis pipeline has been made up of four sequential steps: Center Profiting, Predictive model targeting, Predictive model construction and Predictive model validation.

Sidong Wei and Xuejiao Zhao [4]: In this paper, a comprehensive exploration to the most popular techniques were made (e.g., DNN (Deep Neural Network), SVM (Support Vector Machine), etc.) used to identify diabetes and data preprocessing methods. Basically, we examine these techniques by the accuracy of cross-validation on the Pima Indian data set. The accuracy of each classifier over several ways of data preprocessors and we modify the parameters to improve their accuracy. The best technique found has 77.86% accuracy using 10-fold cross-validation. Also, they analyze the relevance between each feature with the classification result.

## Methodology

## The problem we have at hand is binary classification. So, we decided to use four different supervised learning algorithms to create our ML model.

* K-Nearest Neighbors: The KNN algorithm is a non-parametric technique used for classification.
* Logistic Regression: It is a process of modeling the probability of a discrete outcome given an input variable.
* Support Vector Machines: SVMs are a set of supervised learning methods used for classification, regression and outlier detection.
* Random Forest Classifier: It is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time.

## Results Section

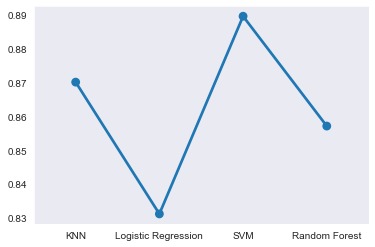
Metrics:

As we are using supervised learning classification models, accuracy is the best metric for evaluating our model.

Table Comparing Accuracies:

|  |  |  |
| --- | --- | --- |
| **Model Number** | **Model Name** | **Accuracy Score** |
| 1 | K-Nearest Neighbor | 87.01 % |
| 2 | Logistic Regression | 83.11 % |
| 3 | Support Vector Machines | 88.96 % |
| 4 | Random Forest Classifier | 85.71 % |

Point plot describing the accuracy of all the models used:



## Discussion

Comparing the considered four machine learning we can observe that SVM is the best fit model for the considered dataset. On deploying the web application, we can predict the chances of a person being diabetic based on the parameters that lead to the diagnosis.

## 

## Conclusion

## The conclusion drawn from this project is that, SVM is the best fit model for the considered problem.

## Contributions/References

[1] <https://www.sciencedirect.com/science/article/pii/S2001037016300733>

[2] <https://www.sciencedirect.com/science/article/abs/pii/S175199182100019X>

[3] <https://journals.sagepub.com/doi/abs/10.1177/1932296817706375>

[4] <https://ieeexplore.ieee.org/abstract/document/8355130>

Link for dataset:

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>