Prediction of diabetes and analysis of its symptoms among patients of different ages using machine learning approaches

Submitted By

**Name**

**Id:**

**Name 2**

**Id:**

This Report in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

**Supervisor Name**

Lecturer

Department of CSE

University Name

**University LOGO**

**UNIVERSITY**

**Dhaka, Bangladesh**

**2021**

**Declaration**

We hereby declare that this thesis has been done by us, under the supervision of**Teacher Name, Lecturer,**and Department of CSE**i University**, affiliated to the partial fulfillment of the requirement for the award of the degree of Bachelor of Computer Science and Engineering, from 2017 through 2021 at **i University**. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

**Submitted by:**

Signature: ………….

Date: ………………..

**Name 1**

**ID:**,

Batch:

Signature: ……………

Date: ………………….

**Name 2**

**ID:**

Batch:

**Supervised by:**

Signature: ………….

Data: ……………….

**Name**

**Lecturer**

Department of CSE, City University

**ACKNOWLEDGEMENT**

We are deeply indebted to my supervisor **Name, Lecturer**, name University, Dhaka, Bangladesh. Without her help guidance, sympathetic co-operation, stimulating suggestions, and encouragement, the planning, and development of this thesis would have been very difficult for us. We would like to express our heartiest gratitude to **Name, Associate Professor, and Head** of, Department of CSE, for his kind help to finish our thesis. We are bound to the Honorable **Prof. Dr. Engr. NAme, Dean, Faculty of Science and Engineering & Adviser**, Name University, Dhaka, Bangladesh, for his endless support Not to mention, and also to other faculty members and the staff of the CSE department of Name University. We are also very grateful to all those people who have helped us to complete our thesis including our parents & family**.**

Signature:………………

Name 1

ID:

Signature:……………….

Name 2

ID:

**Abstract**

This thesis is on “**Prediction of diabetes and analysis of its symptoms among patients of different ages using machine learning approaches**”.

Diabetes is a common problem in different ages people and those people who not only suffering diabetes but also, they have suffered from other symptoms which affect their health. However, to get a good recognition process we need a large amount of data which may be sufficient to generate good results. It is difficult to collect that amount of data from any health care service. On the other hand, we can use an online existing dataset to establish a proper system to get a better process for Diabetes problems. There are numerous ML algorithms and used for several reasons for examination and prediction. This work shows, how this disease gives better accuracy from the exit classification algorithms. Classification algorithms like SVM, KNN, Random Forest and Naïve Bayes algorithms execute their better performance based on the dataset. Until now there are three types of diabetes which are Type 1, Type 2, and Type 3. We generate accuracy and predict those types by distributing our whole dataset according to the age of the patients. Some algorithms are implemented on the types to find out accuracy. In diabetes prediction, the Random Forest algorithm gives better accuracy than another algorithm that is 80%.

**Keywords:** Diabetes, ML, Classification Algorithm, Diabetes Type, Age group Symptom.

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| KNN | K-nearest neighbor |
| RF | Random Forest Classifier |
| NB | Gaussian Naive Bayes |
| SVM | Support Vector Machine |

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| CONTENS | PAGE |
| Declaration | ii |
| Acknowledgements | iii |
| Abstract | iv |
| List of Abbreviations | v |
| Chapter | vi-vii |
| List of Figures | viii-ix |
| List of Tables | ix |
|  |  |
| CHAPTER |  |
| CHAPTER-1: Introduction | 1-3 |
| **CHAPTER**-**2: Related Work** | 4-7 |
| **CHAPTER**-**3: Methodology** | 8-12 |
| 3.1 Support Vector Machine (SVM) Algorithm | 10 |
| 3.2 Random Forest Algorithm | 10 |
| 3.3 KNN Algorithm | 11 |
| 3.4 Naïve Bayes Algorithm | 11 |
| 3.5 Proposed Methodology | 11-12 |
| **CHAPTER**-**4: PERFORMANCE EVALUATION** | 13-35 |
| 4.1 Diabetes Prediction | 14 |
| 4.2 Dataset | 15 |
| 4.3 Result Analysis | 15-30 |
| 4.3.1 Mean and Standard Deviation of Dataset | 15 |
| 4.3.2 Balanced Dataset | 16 |
| 4.3.3 Patients Percentage | 17 |
| 4.3.4 Simple Age Distribution on Diabetes | 18 |
| 4.3.5 Heat Map | 19 |
| 4.3.6 Histogram Plot | 20-28 |
| 4.3.6.1 Pregnancies | 20-21 |
| 4.3.6.2 Glucose | 21-22 |
| 4.3.6.3 Blood Pressure | 22-23 |
| 4.3.6.4 Skin Thickness | 23-24 |
| 4.3.6.5 Insulin | 24-25 |
| 4.3.6.6 BMI | 25-26 |
| 4.3.6.7 Diabetes Pedigree Function | 26-27 |
| 4.3.6.8 Age | 27-28 |
| 4.3.7 Classification Algorithm Accuracy | 28-30 |
| 4.3.7.1 SVM Algorithm | 28 |
| 4.3.7.2 Random Forest Algorithm | 29 |
| 4.3.7.3 KNN Algorithm | 29 |
| 4.3.7.4 Naïve Bayes | 30 |
| 4.4 Types | 30-31 |
| 4.4.1 Type 1 | 30 |
| 4.4.2 Type 2 | 31 |
| 4.4.3 Type 3 | 31 |
| 4.5 Age Group Symptom Analysis | 32-35 |
| 4.5.1 Age Group 1 | 32-33 |
| 4.5.2 Age Group 2 | 33-34 |
| 4.5.3 Age Group 3 | 34-35 |
| **CHAPTER** -**5: DISCUSSION** | 36-37 |
| **CHAPTER**-**6: CONCLUSION** | 38-39 |
| REFERENCES | 40-44 |

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **FIGURES** | **PAGE NO** |
| Figure 3.1: Proposed Methodology | 12 |
| Figure 4.1: Dataset Information | 15 |
| Figure 4.2: Mean and Standard Deviation of Dataset | 15 |
| Figure 4.3: Check Balanced Data | 16 |
| Figure 4.4: Diabetes Patients Percentage | 17 |
| Figure 4.5: Simple age Distribution on Diabetes | 18 |
| Figure 4.6: Heat Map Plot | 19 |
| Figure 4.7: Pregnancies Distribution on Dataset | 20 |
| Figure 4.8: Pregnancies Distribution on positive and negative Diabetes | 21 |
| Figure 4.9: Glucose Distribution on Dataset | 21 |
| Figure 4.10: Glucose Distribution on positive and negative Diabetes | 22 |
| Figure 4.11: Blood Pressure Distribution on Dataset | 22 |
| Figure 4.12: Blood Pressure Distribution on positive and negative Diabetes | 23 |
| Figure 4.13: Skin Thickness Distribution on Dataset | 23 |
| Figure 4.14: Skin Thickness Distribution on positive and negative Diabetes | 24 |
| Figure 4.15: Insulin Distribution on Dataset | 24 |
| Figure 4.16: Insulin Distribution on positive and negative Diabetes | 25 |
| Figure 4.17: BMI Distribution on Dataset | 25 |
| Figure 4.18: BMI Distribution on positive and negative Diabetes | 26 |
| Figure 4.19: Diabetes Pedigree Function Distribution on Dataset | 26 |
| Figure 4.20: Diabetes Pedigree Function Distribution on positive and negative Diabetes. | 27 |
| Figure 4.21: Age Distribution on Dataset | 27 |
| Figure 4.22: Age Distribution on positive and negative Diabetes | 28 |
| Figure 4.23: Confusion Matrix for SVM Classification | 28 |
| Figure 4.24: Confusion Matrix for Random Forest Classification | 29 |
| Figure 4.25: Confusion Matrix for KNN Classification | 29 |
| Figure 4.26: Confusion Matrix for Naïve Bayes Classification | 30 |
| Figure 4.27: Age Group 1 all Symptoms | 32-33 |
| Figure 4.28: Age Group 2 all Symptoms | 33-34 |
| Figure 4.29: Age Group 3 all Symptoms | 34-35 |
|  |  |

**LIST OF TABLES**

|  |  |
| --- | --- |
| **TABLES** | **PAGE** |
| Table 4.1: Type 1 Accuracy | 30 |
| Table 4.2: Type 2 Accuracy | 31 |
| Table 4.3: Type 3 Accuracy | 31 |