# 

# Abstract

Patients with diabetes must constantly check their blood glucose levels. Deviations from the normal blood glucose range might cause substantial short- and long-term consequences. People would be able to take preventive action, if an automated prediction model alerted them to impending changes in their blood glucose levels. In this paper we describe the prediction of diabetic by using various machine learning classification models like logistic regression, support vector machine, decision tree classifier, random forest and Naive Bayes on a specific patient’s data. For this the data considered is the history of patients like Pregnancies, Blood Pressure, Insulin, Glucose level and Diabetes Pedigree Function. Hence such a system has a significant place in healthcare system of any hospitals. The main objective of this paper is to analyze previous year’s patient’s historical data and analysis of performance of machine learning algorithms on diabetes disease for better decision making to increase the accuracy of the model in the percentage of the hospitals.

[Electronic Health Records](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/electronic-health-record) (EHR) generates a high volume of clinical information, which is very complicated to make management decision by using business intelligence alone. In order to end this, machine learning and data mining methods has been introduced where machine is responsible for functioning human intelligence.

In this study 769 patients’ data is taken and various algorithm has been used to predict the accuracy of the machine learning model on diabetes. It has been found that 65.1% of females has been diagnosed with diabetes and 34.9% with no diabetes. Glucose is one of the main factors which is responsible for diabetes with the positive correlation of +0.47. Naive Bayes are undoubtedly superior to the other classifiers. The best result obtained for diabetic dataset is 0.7727. It suggests that machine learning may be used to predict diabetes, but it's crucial to identify the right features, classifier, and data mining technique. Based on the diabetic dataset, it was unable to forecast the type of the diabetes which is considered to be the future scope of the work.

# **Table of Contents**

**Contents**

[Abstract 1](#_Toc116644169)

[Table of Contents 2](#_Toc116644170)

[List of Tables 3](#_Toc116644171)

[List of Figures 4](#_Toc116644172)

[Abbreviation and Acronyms 5](#_Toc116644173)

[Chapter 01: Introduction and Motivation 6](#_Toc116644174)

[1.1 Business Intelligence 6](#_Toc116644175)

[1.2 Artificial Intelligence 7](#_Toc116644176)

[1.3 Machine Learning 7](#_Toc116644177)

[1.4 Difference Business Intelligence and Artificial Intelligence 7](#_Toc116644178)

[1.5 Transformation of Business Intelligence into Artificial Intelligence 8](#_Toc116644179)

[1.6 Motivation for the Study 8](#_Toc116644180)

[Chapter 02: Literature Review and Problem Formulation 10](#_Toc116644181)

[2.1 Critical Literature Review 10](#_Toc116644182)

[2.2 Review of Literature Summary 13](#_Toc116644183)

[Chapter 03: Problem Statement 20](#_Toc116644184)

[3.1 Title 20](#_Toc116644185)

[3.2 Aim 20](#_Toc116644186)

[3.3 Objectives 20](#_Toc116644187)

[3.4 Scope of Present Investigation 20](#_Toc116644188)

[3.5 Methods and Methodology/Approach to attain each objective 21](#_Toc116644189)

[Chapter 04: Methodology 22](#_Toc116644190)

[4.1 Materials and Methods 22](#_Toc116644191)

[4.1.1 Details of dataset 22](#_Toc116644192)

[4.2 Data Collection 23](#_Toc116644193)

[4.2.1 Diabetic Dataset 23](#_Toc116644194)

[4.3 Diabetes Classification Techniques 23](#_Toc116644195)

[4.3.1 Logistic Regression 23](#_Toc116644196)

[4.3.2 Random Forest Classifier 24](#_Toc116644197)

[4.3.3 Support Vector Machine 24](#_Toc116644198)

[4.3.4 Decision Tree Classifier 25](#_Toc116644199)

[4.3.5 Naïve Bayes 25](#_Toc116644200)

[4.4 Explanatory Data Analysis 25](#_Toc116644201)

[4.4.1 Feature information 25](#_Toc116644202)

[4.4.1 Descriptive statistics 26](#_Toc116644203)

[4.5 Data Visualization 26](#_Toc116644204)

[4.5.1 Histogram and density graph 27](#_Toc116644205)

[4.5.2 Correlation Matrix 28](#_Toc116644206)

[4.5.3 Pie and Bar chart for distribution of outcomes 31](#_Toc116644207)

[4.5.4 Pairplot – relationship between variables 32](#_Toc116644208)

[Chapter 05: Results and Discussions 33](#_Toc116644209)

[5.1 Results and Discussion 33](#_Toc116644210)

[5.1.1 Methods to improve the accuracy of the machine learning model 35](#_Toc116644211)

[Chapter 06: Conclusions and Future Directions 36](#_Toc116644212)

[6.1 Conclusion 36](#_Toc116644213)

[6.2 Limitations 36](#_Toc116644214)

[6.3 Future Directions 37](#_Toc116644215)

[Appendix 38](#_Toc116644216)

[References 43](#_Toc116644217)

# 

# List of Tables

[Table 1 difference between business intelligence and artificial intelligence 10](#_Toc109917367)

[Table 2 Summary of Literature Review 15](#_Toc109917368)

[Table 3 Objective - Methodology 23](#_Toc109917369)

[Table 4 Correlation Matrix 31](#_Toc109917370)

[Table 5 Accuracy Table 36](#_Toc109917371)

[Table 6 Confusion Matrix 36](#_Toc109917372)

# List of Figures

[Figure 1 diabetic dataset 25](#_Toc110983310)

[Figure 2 feature information of dataset 28](#_Toc110983311)

[Figure 3 Descriptive statistics 28](#_Toc110983312)

[Figure 4 Histogram and density graph 29](#_Toc110983313)

[Figure 5 Correlation Matrix 32](#_Toc110983314)

[Figure 6 Distribution of Outcomes 33](#_Toc110983315)

[Figure 7 Pairplot 34](#_Toc110983316)

[Figure 8 Confusion Matrix 36](#_Toc110983317)

[Figure 9 Logistic Regression Model 40](#_Toc110983318)

[Figure 10 Logistic Regression Confusion Matrix 40](#_Toc110983319)

[Figure 11 Support Vector Machine Model 41](#_Toc110983320)

[Figure 12 Support Vector Machine Confusion Matrix 41](#_Toc110983321)

[Figure 13 Naive Bayes Model 42](#_Toc110983322)

[Figure 14 Naive Bayes Confusion Matrix 42](#_Toc110983323)

[Figure 15 Decision Tree Classifier Model 43](#_Toc110983324)

[Figure 16 Decision Tree Classifier Confusion Matrix 43](#_Toc110983325)

[Figure 17 Random Forest Classifier Model 44](#_Toc110983326)

[Figure 18 Random Forest Classifier Confusion Matrix 44](#_Toc110983327)

# Abbreviation and Acronyms

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Abbreviations** | **Full Form** |
| 1 | AI | Artificial Intelligence |
| 2 | BI | Business Intelligence |
| 3 | ML | Machine Learning |
| 4 | BMI | Body Mass Index |
| 5 | SVM | Support Vector Machine |
| 6 | LG | Logistic Regression |

# Chapter 01: Introduction and Motivation

This chapter contains a brief introduction on business intelligence and artificial intelligence, along with the impact of artificial intelligence on business decision making.

## Business Intelligence

Managers make decisions that have long-term effects every day. Executives must comprehend their company's dynamics, assess risks, and predict market behaviour in order to make these decisions. Business Intelligence (BI) is the process of transforming data into actionable insights that help a company make strategic and tactical choices. To offer users with detailed insight about the condition of the business, BI technologies access and analyze data sets and show analytical findings in reports, summaries, dashboards, graphs, charts, and maps. The overall goal of business intelligence is to help a company make better decisions. Data will be accurate, thorough, and structured in a corporation with a strong BI strategy. Business intelligence may be used to display historical patterns in order to help stakeholders assess the health of their firm, alerting them to both issues and opportunities.

## Artificial Intelligence

One of the fundamental aims of artificial intelligence is to model human intellect. By simulating human actions and thinking processes. AI approaches have recently made huge impacts in the healthcare industry. The recent effective uses of AI in healthcare have been made feasible by the rising availability of healthcare data and the quick development of big data analysis methodologies. Powerful AI algorithms may reveal clinically important information hidden in vast amounts of data, which can aid clinical decision making when guided by relevant clinical queries.

[Zohuri .V.A et al., 2020]

## Machine Learning

Machine learning is a branch of science that studies how machines learn from their experiences. For many scientists, the terms "machine learning" and "artificial intelligence" are interchangeable, because the ability to learn is the most important attribute of an object classified as intelligent in the largest context. Machine learning aims to create computer systems that can adapt and learn from their past experiences. [[Ioannis Kavakiotis](https://www.sciencedirect.com/science/article/pii/S2001037016300733" \l "!) et al., 2017]

## Difference Business Intelligence and Artificial Intelligence

**Table 1 difference between business intelligence and artificial intelligence**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Artificial Intelligence** | **Business Intelligence** |
| Theme conception | Human-like computer intelligence is involved in artificial intelligence. | Business intelligence includes making informed decisions. |
| Area of focus | It covers statistical analysis principles. | It is concerned with deep learning and machine learning algorithms. |
| Scope | Its scope is related with future occurrences. | Its extent is determined by what has occurred previously. |
| Disadvantages | It has disadvantages, including a threat to privacy and safety. | It contains flaws such as outdated technology and data abuse. |
| Goal | Artificial intelligence's major goal is to create machines that can function similarly to the human brain. | The goal of the business intelligence is to analyze data and forecast the future using historical data. |
| Tools | Uses Algorithms | Uses data mining tools |

## Transformation of Business Intelligence into Artificial Intelligence

In the subject of data science, Business Intelligence (BI) and artificial intelligence (AI) are both monitored. The combination is seen to be a perfect fit. Data science is generally employed by research analysts to produce results for artificial intelligence and business intelligence research. According to the study, Artificial Intelligence (AI) is capable of coping with a greater number of algorithms than Business Intelligence (BI). Big data management makes extensive use of algorithms along with artificial intelligence, business intelligence, and data science. For statistical analysis, business intelligence and data science principles are applied. For supervised machine learning, artificial intelligence and data science are generally employed. The two can be combined to produce cutting-edge dashboards for emerging technology. [Zohuri .V.A et al., 2020].

## Motivation for the Study

Artificial intelligence is mostly employed in the creation of expert systems that employ human intellect. Business intelligence, on the other hand, aims to create effective and efficient business solutions. AI can 'learn' features from a big volume of healthcare data using complex algorithms, and then utilize the results to aid clinical practice. It might also include learning and self-correcting capabilities to enhance accuracy depending on input. Physicians can benefit from AI systems that provide up-to-date medical knowledge from journals, textbooks, and clinical practices to help them provide effective patient care. Furthermore, an AI system can assist in the reduction of diagnostic and treatment mistakes that are unavoidable in human clinical practice. Furthermore, an AI system collects usable data from a huge patient population to aid in developing real-time conclusions for health risk warning and prediction. [Jiang, Fei et al., 2017]

# Chapter 02: Literature Review and Problem Formulation

This chapter examines the impact of AI in decision making and machine learning techniques to predict diabetic prediction. Several research papers are critically reviewed in this chapter.

## 2.1 Critical Literature Review

1. **Serge-Lopez Wamba-Taguimdje, et al (2020)**, in their paper **“Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects”** stated that the purpose of their study is to analyze the influence of Artificial Intelligence (AI) on firm performance, notably by building on the business value of AI-based transformation projects. This study was conducted using a four-step sequential approach: (1) analysis of AI and AI concepts/technologies; (2) in-depth exploration of case studies from a great number of industrial sectors; (3) data collection from the databases (websites) of AI-based solution providers. Detailed analysis of the case studies shows that AI and its technologies offer a wide range of options, benefits and services, with the aim to improve Influence of artificial intelligence on firms organizational performances.
2. **Mohammad Hossein Ghahramani (2020)**, in their paper “**AI-Based Modeling and Data-Driven Evaluation for Smart Manufacturing Processes**” proposed a dynamic algorithm for gaining useful insights about semiconductor manufacturing processes and to address various challenges an emerging area of ML, AI, and IoT, it was a promising solution for efficient, cost-effective manufacturing production The goal of manufacturing enterprises is to develop cost-effective and competitive products. Manufacturing intelligence can significantly improve effectiveness by bridging business and manufacturing models with the help of low-cost sensor data
3. **Manjeet (2019),** in their paper **“The product manager in the artificial intelligence world”** stated that Cloud applications, artificial intelligence, machine learning, data insights, rapid prototyping, design thinking, and faster decision making are becoming more and more significant in the daily life of the product manager. With the advancement in machine learning, products can now significantly differ from the traditional style of product designs. A collaboration between humans and machine, This is what is going on in the business world and product management.
4. **Hany H Ammar, et al (2012), in their article “Software Engineering Using Artificial Intelligence Techniques: Current State and Open Problems”** stated the application of artificial intelligence approaches to the software engineering processes. These approaches can have a major impact on reducing the time to market and improving the quality of software systems in general. In this paper, surveyed promising research work on applying AI techniques to solve some of the most important problems facing the software engineer and surveyed research in the development activities of requirements engineering, software architecture design, and coding and testing processes.
5. [**David Higgins**](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Higgins%2C+David)**, et al (2020),** in their paper **“A Practical Framework for Artificial Intelligence Product Development in Healthcare”** stated that Artificial intelligence (AI) in healthcare holds great potential to expand access to high-quality medical care, while reducing systemic costs. The development process proposed for AI in healthcare software strongly diverges from modern consumer software development processes. This paper developed a framework which outlines, considered to be, best practices in AI in healthcare development.
6. **Long, et al (2020)**, in their paper “**Developing an Artificial Intelligence (AI) Management System to Improve Product Quality and Production Efficiency in Furniture Manufacture**” stated there are some problems in furniture production industry, such as low production efficiency, low accuracy, and lack of innovation for products. To resolve those problems, an AI management system is developed to improve the product quality and production efficiency in furniture enterprises. Although AI cannot really achieve intelligence, and cannot fully replace human beings, its ability to receive and process information is far beyond people's ability.
7. **Crispin Coombs, et al (2020)**, in their paper “**The strategic impacts of Intelligent Automation for knowledge and service work**” stated a significant recent technological development concerns the automation of knowledge and service work as a result of advances in Artificial Intelligence (AI) and its sub-fields. his development presents organisations with a new strategic opportunity to increase business value. Paper conceptualize Intelligent Automation and its associated technologies. It provides a business value-based model of Intelligent Automation for knowledge and service work and identify research gaps that hinder a complete understanding of the business value realisation process
8. **Neha Sonia, et al (2020),** in their paper “**Artificial Intelligence in Business: From Research and Innovation to Market Deployment**” investigates the overall impact of AI - from research and innovation to deployment. The paper addresses the influential academic achievements and innovations in the field of AI; their impact on the entrepreneurial activities and thus on the global market. the top 200 AI start-ups are analyzed, the investors and entrepreneurial actions are investigated in launching AI-based services in existing and new industries. The analysis of the top 200 AI start-ups explicitly shows the influence of advance research and innovation in AI on the global market. The study shows that the AI wave is on and an appetite for AI growth is exponential.
9. **Zou, et al (2018)**, in their paper “**Predicting Diabetes Mellitus with Machine Learning Techniques**” implemented decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. Study reveals the root node is glucose, which can show the glucose has the maximum information gain, so it confirms the common sense and the clinical diagnosis basis. But there are diabetic patients whose fasting blood glucose is less than 6.8 in Luzhou dataset, they considered the reason that they injected insulin before the physical examination to control blood sugar levels.
10. **Alghamdi M, et al (2017),** in their paper “**Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project**” investigate the relative performance of various machine learning methods such as Decision Tree, Naïve Bayes, Logistic Regression, Logistic Model Tree and Random Forests for predicting incident diabetes using medical records of cardiorespiratory fitness. This study shows the potential of machine learning methods for predicting incident diabetes using cardiorespiratory fitness data. The Random Forest and NB Tree models showed greater results in all model evaluation metrics.

## 2.2 Review of Literature Summary

**Table 2 Summary of Literature Review**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SI.**  **No.** | **Authors** | **Year of Publication** | **Research Focus** | **Methodology** | **Research Findings** | **Conclusions drawn by Author** | **Limitations of the Study** | **Critical Appraisal** |
| **s** | Serge-Lopez Wamba-Taguimdje, Cameroon Samuel Fosso Wamba, Jean Robert Kala Kamdjoug and Chris Emmanuel Tchatchouang Wanko | 2020 | The main purpose of our study is to analyze the influence of Artificial Intelligence (AI) on firm performance, notably by building on the business value of AI-based transformation projects | The research process was based on a review of 500 case studies. Studying the influence of AI on the performance of organizations, and more specifically, of the business value of such organizations’ AI-enabled transformation projects, required us to make an archival data analysis following the three steps, namely the conceptual phase, the refinement and development phase, and the assessment phase. | The results of our study have highlighted the AI benefits in organizations, and more specifically, its ability to improve on performance at both the organizational (financial, marketing and administrative) and process levels | In this study, which focused on analyzing the influence of AI on organizational performance, we adopted a qualitative approach based on the analysis of 500 case studies. The choice of this approach was justified by the benefits it offers and by the fact that this type of research is based on secondary data | For the purpose of our study, we adopted a research framework geared toward a more inclusive and comprehensive approach so as to better account for the intangible benefits of AI within organizations. | This study was conducted using a four-step sequential approach:   1. analysis of AI and AI concepts/technologies; 2. in-depth exploration of case studies from a great number of industrial sectors; 3. data collection from the databases of AI-based solution providers; 4. a review of AI literature to identify their impact on the performance of organizations while highlighting the business value of AI-enabled projects transformation within organizations |
| **2.** | Mohammadhossein Ghahramani | 2020 | AI-Based Modeling and Data-Driven Evaluation for Smart Manufacturing Processes | Outlier Detection, Handling an Imbalanced Data Set, Feature Selection, Feature Extraction Methods | The experimental results show that our proposed model is superior over those conventional ones. The corresponding accuracy rate of the proposed model is over 90%. | We have proposed a dynamic feature selection method based on GA and ANN. We have compared the results achieved in this work with traditional approaches to prove the effectiveness of our proposed solution. | We plan to consider other MOEAs for solving our optimization problem in a way that both feature selections objective functions are optimized simultaneously. | It aims to achieve a high level of intelligence with the latest appropriate technology-based computing, advanced analytics, and new levels of Internet connectivity. |
| **3.** | Hany H Ammar, Walid Abdelmoez , and Mohamed Salah Hamdi | 2015 | Software Engineering Using Artificial Intelligence Techniques: Current State and Open Problems | Requirement Engineering, Processing Natural Language Requirements NLR, Software architecture design, Software coding and testing | Open problems that Artificial Intelligence can help:  Disambiguating natural language requirements  Developing knowledge based systems and ontologies to manage the requirements and model problem domains  The use of computational intelligence to solve the problems of incompleteness and prioritization of requirements | In this paper, we surveyed promising research work on applying AI techniques to solve some of the most important problems facing the software engineer. We surveyed research in the development activities of requirements engineering, software architecture design, and coding and testing processes. We summarized the most important open problems in these active research areas. | Dealing with large lines of code for analysis seems to be difficult. | Use of AI to help automate the programming process. This is done by having human specialists write a complete and concise specification of the desired software; so that, a system can generate "functions, data structures, or entire programs" directly from the specifications |
| **4.** | Guang Jin Long, Bai Hua Lin, Hong Xing Cai, Guang Zai Nong | 2020 | Developing an Artificial Intelligence (AI) Management System to Improve Product Quality and Production Efficiency in Furniture Manufacture | Establishing an AI Data Management System, Establishing an Expert System on Furniture Manufacture, Production Monitoring and Auto Controlling by the AI Management System, Management of the Internal Space by the AI Management System | The investment for an AI management system is totally 825,000 Yuan RMB. Where, the total software consumed 450,000 Yuan, accounting for 54.5% of the total investment. And the manpower pay is total 300,000 Yuan, accounting for 36.9% of the total investment. The investment of an AI management system is not too much. | The first fully automated mattress production line in Asia is the “MuSi” Bedding Project in Dongguan city, Guangdong province, China. At least 200 million Yuan has been invested to build a fully-automatic factory by the project. In addition, more furniture enterprises in Dongguan city are trying to improve their intelligence level by intelligent information management and automatic production equipment. | This paper doesn’t specify the methodologies in detail for analysis. | Using artificial intelligence technology to improve product quality and production efficiency in furniture enterprises is an effective method. There is witness that some furniture enterprises have begun to explore the footsteps of intellectualization. Intellectualization will become a developing trend of furniture enterprise |
| **5.** | Manjeet Singh | 2019 | This study will deal with the question: how can the product manager thrive in an artificial intelligence machine learning world?. With the advancement in machine learning, products can now significantly differ from the traditional style of product designs. | Research was conducted on books, journals and articles from various databases, like PROQuest, EBSCO, and other online databases, including websites that can provide important information and contexts regarding the so-called competition between human perception and artificial intelligence. Blockchain, Chatbots, Internet of Things (IoT) chatbots, The printer revolution. | An Oxford University study says that by 2025, millions of U.S. jobs will be lost due to automation. At least 173 million jobs are at risk within eight years, due to automation. Frank et al. argue that only about 12% of the jobs might be automated, which is equivalent to about 19 million jobs in the U.S. Automation will also change the G7 labor force: 12% of existing jobs will be at risk, a great portion of the jobs (75%) is altered or enhanced, but about 13% will be created. | If we do not rely on machines, then we can switch back to the traditional way. But is that not the primitive way – going backward? Why not ask the computer to find ways instead of discarding it and going back to the olden times? Those are suppositions – and beliefs that are not true. Turning back the hands of time might be catastrophic. | This paper focuses more on the traditional ways rather than AI | Humans still dominate the thinking and business operations and outperform technology on highly complex tasks. But the way smart technology has performed these past two decades will tell that there is an ongoing competition. |
| **6.** | Neha Sonia, Enakshi Khular Sharmaa, Narotam Singhb, Amita Kapoorc | 2019 | This paper investigates the overall impact of AI - from research and innovation to deployment. The paper addresses the influential academic achievements and innovations in the field of AI; their impact on the entrepreneurial activities and thus on the global market. The paper also contributes in investigating factors responsible for the advancement of AI. | Convolutional Neural Networks (CNNs), Policy Gradient Reinforcement Learning, Deep Autoencoder, Deep QNetworks | Recognition of the factors which are resulting in today’s AI exponential growth. Identification of academic achievements in AI which are advancing the commercially available intelligent products. Determination of the top AI industries and investment trends in AI. Exploration of geographically strong AI locations.  The data analysis done proves that AI is not hype | The analysis of the top 200 AI start-ups explicitly shows the influence of advance research and innovation in AI on the global market. The study shows that the AI wave is on and an appetite for AI growth is exponential | the study shows that the AI technology is confined only in a few regions in the world | Research has also suggested that information about generic drugs from generic manufacturers is inadequate |
| **7.** | Crispin Coomb, Donald Hislop, Stanimira K. TanevaSarah BarnardGrac | 2017 | A significant recent technological development concerns the automation of knowledge and ser-  vice work as a result of advances in Artificial Intelligence (AI) and its sub-fields. | Decision Support System  Expert Systems (ES)  Recommendation Agent (RA) | contextual factors influenced Intelligent Automation enabled business process performance | provide a business value-based model of Intelligent Automation for knowledge and service work and identify  twelve research gaps that hinder a complete understanding of the business value realisation process | there are significant limitations in understanding of how organisations decide the level of Intelligent Automation that  is implemented | How do organisations determine which tasks are  appropriate for Intelligent Automation, and what  factors shape this decision-making process |
| **8.** | [David Higgins](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Higgins%2C+David) | 2020 | A Practical Framework for Artificial Intelligence Product Development in Healthcare | artificial neural networks (ANNs)  support vector machines  Gaussian processes  Etc | framework for the development of AI-driven biomedical products from conception to market launch is presented. The framework highlights the risks, objectives, and key results which are typically required to navigate a three-phase process to market-launch of a validated medical AI product | aim to improve the development situation of AI-driven medical products by enhancing accountability, transparency, and planning, hopefully increasing success rates for such products, without facing some of the safety and ethical trade-offs mentioned previously | Current software product development practices are rarely a fit for both the medical certification as well as the clinical validation process. This—in turn—leads to 1) much longer development cycles and 2) quite different proof points inherent in developing a product which must be medically certified. | The development process proposed for AI in healthcare software strongly diverges from modern consumer software development processes. Key time points to guide founders, investors, and key stakeholders throughout the process are highlighted. |
| **9.** | Quan Zou1, Kaiyang Qu ,Yamei Luo , Dehui Yin , Ying Ju and Hua Tang | 2018 | Predicting Diabetes Mellitus With Machine Learning Techniques | In this study, used decision tree, random forest and neural network to predict diabetes mellitus. | The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used | ccording to the all above experiments, we found the accuracy of using PCA is not good, and the results of using the all features and using mRMR have better results. The result, which only used fasting glucose, has a better performance especially in Luzhou dataset. | Due to the data, we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and exploring the proportion of each indicator, which may improve the accuracy of predicting diabetes | The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross validation was used to examine the models. In order to verity the universal applicability of the methods, we chose some methods that have the better performance to conduct independent test experiments |
| **10.** | Manal Alghamdi, Mouaz Al-Mallah, Steven Keteyian , Clinton Brawner , Jonathan Ehrman , Sherif Sakr | 2017 | Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project | Logistic Regression  Logistic Model Tree  Random Forest | The results show that the Logistic Regression (LR) classifier achieves the highest performance | The Random Forest and NB Tree models showed greater results in all model evaluation metrics. The study shows the potential of ensembling and SMOTE approaches for predicting incident diabetes using cardiorespiratory fitness data | The variance between the two classes is considerably large and could lead to lower accuracy on the prediction of the classifiers. | The study shows the potential of ensembling and SMOTE approaches for predicting incident diabetes using cardiorespiratory fitness data. |

# Chapter 03: Problem Statement

The chapter on problem statement deals with the title, aim, objectives and scope of investigation and methodology used for each objective.

## 3.1 Title

A Study and Analysis of Performance of Machine Learning Algorithms on Diabetes Disease

## 3.2 Aim

To improve the data analysis of diabetic dataset to find the diabetic pattern in the patients by developing various machine learning models like logistic regression, KNN and random forest to predict the results and estimating the accuracy of the models and implement the best model for rational decision making.

## 3.3 Objectives

* To collect the diabetic dataset to find the diabetic pattern in the patients.
* To analyze the diabetic data & visualize them to get the insight about factors affecting diabetic and their relationship.
* To develop the Artificial Intelligence (AI) models for analyzing the collected diabetic data for better rational decision making.
* To implement & evaluate the accuracy of the machine learning model and suggest the methods to improve the accuracy of the machine learning model.

## 3.4 Scope of Present Investigation

The National Institute of Diabetes and Digestive and Kidney Diseases provided diabetic data and taken from an open-source platform Kaggle. The dataset's goal is to diagnose whether a patient has diabetes using diagnostic metrics provided in the collection. Eight variables were selected Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI) and Diabetes Pedigree Function as well as one objective variable, Outcome. The selection of these cases from a wider database was subjected to certain limitations. All of the patients at this diabetic dataset are Pima Indian women who are at least 21 years old. All the data was collected and analyzed to find the diabetic pattern.

## 3.5 Methods and Methodology/Approach to attain each objective

**Table 3 Objective - Methodology**

|  |  |  |  |
| --- | --- | --- | --- |
| **Objective No.** | **Statement of the Objective** | **Method/ Methodology** | **Resources Utilized** |
| **1** | To collect the diagnostic data | The National Institute of Diabetes and Digestive and Kidney Diseases provided diabetic data | Kaggle dataset |
| **2** | To analyze the collected data & visualize the data | Explanatory data analysis & visualization of diabetic pattern using histogram, pairplot, pie chart etc. | Python programming language |
| **3** | To develop the AI models for analyzing the collected the data for better decision making | Developed various machine learning models like logistic regression, random forest, naïve bayes, decision tree and SVM to predict the results | Python programming language |
| **4** | To implement & evaluate the accuracy of the model, suggest to improve the accuracy | Estimated the accuracy of the models and implemented the best model for rational decision making | Python programming language |

# **Chapter 04: Methodology**

This chapter consists of the methods used to carry the study specifying the following: - data collection, details of dataset and the methods used for analysis. It basically covers how the study has been designed for solving the problem definition formulated in a systematic way.

## 4.1 Materials and Methods

The study was conducted on diabetic dataset and was descriptive in nature. The data collection was mainly from The National Institute of Diabetes and Digestive and Kidney Diseases available on Kaggle.

### 4.1.1 Details of dataset

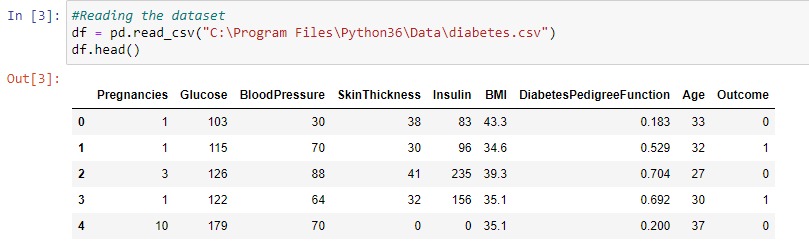
There are diverse clinical predictor factors in the dataset, as well as one objective variable, Outcome. The number of pregnancies the patient has had, their BMI, insulin level, and age are all predictor factors.

* **Pregnancies**: Number of times pregnant
* **Glucose**: Plasma glucose concentration 2 hours in an oral glucose tolerance test
* **Blood Pressure**: Diastolic blood pressure (mm Hg)
* **Skin Thickness**: Triceps skin fold thickness (mm)
* **Insulin**: 2-Hour serum insulin (mu U/ml)
* **BMI**: Body mass index (weight in kg/(height in m)^2)
* **Diabetes Pedigree Function**: Diabetes pedigree function
* **Age**: Age (years)
* **Outcome**: Class variable (0 or 1)

## 4.2 Data Collection

### 4.2.1 Diabetic Dataset

The National Institute of Diabetes and Digestive and Kidney Diseases provided this data and taken from an open-source platform Kaggle. Figure 1 shows a brief view of diabetic dataset with various variables.



**Figure 1 diabetic dataset**

## 4.3 Diabetes Classification Techniques

We suggested a system to predict diabetes using machine learning techniques and algorithms such as Logistic Regression, KNN, SVM, Random Forest, Decision Tree, and Nave Bayes.

### 4.3.1 Logistic Regression

Logistic regression is another statistical tool that machine learning has taken. It's the approach of choice for binary classification issues (problems with two class values). The logistic function, which is at the heart of the procedure, is called logistic regression. The logistic function, also known as the sigmoid function, was created by statisticians to characterize the features of population expansion in ecology, which include rapid increase and reaching the environment's carrying capacity. It's an S-shaped curve that can transfer any real-valued integer to a value between 0 and 1, but never exactly between those two points.

### 4.3.2 Random Forest Classifier

A supervised learning algorithm is random forest. It has classification and regression capabilities. It's also the most adaptable and straightforward algorithm. The trees make up a forest. A forest is thought to be stronger the more trees it has. Random forests generate decision trees from randomly chosen data samples, obtain predictions from each tree, then vote on the best option. It also serves as a strong indicator of the value of the feature.

It is technically a set of dynamic of decision trees created on a randomly divided dataset (based on the divide-and-conquer methodology). The forest refers to a collection of decision tree classifiers. Individual decision trees are created for each characteristic using an attribute selection indicator such as information gain, gain ratio, or Gini index. Each tree is based on a separate random sample. Each tree votes on a classification challenge, and the most popular class is picked as the final result. In the case of regression, the ultimate result is the average of all the tree outputs. In comparison to other non-linear classification methods, it is both simpler and more powerful.

### 4.3.3 Support Vector Machine

SVM stands for Support Vector Machine and is one of the most widely used Supervised Learning algorithms for Classification and Regression issues. However, it is mostly utilized in Machine Learning for Classification difficulties.

The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points may be readily placed in the proper category in the future. A hyperplane is the name for the optimal choice boundary. The extreme points/vectors that assist create the hyperplane are chosen via SVM. Support vectors are the extreme instances, and the method is called a Support Vector Machine.

### 4.3.4 Decision Tree Classifier

A decision tree is a tree structure that resembles a flowchart where each leaf node symbolizes the result and each inside node indicates a characteristic (or attribute). The root node in a decision tree is the first node from the top. It gains the ability to divide data according to attribute values. Recursive partitioning is the process of repeatedly dividing a tree. This framework, which resembles a flowchart, aids in decision-making. It is a flowchart-like representation that perfectly replicates how people think. Decision trees are simple to grasp and interpret.

### 4.3.5 Naïve Bayes

A generative model for classification is the naive Bayes classifier. The Naive Bayes classifier was one of the most often used classifiers for machine learning applications prior to the development of deep learning and its user-friendly libraries. The naive Bayes classifier works admirably in many situations despite its simplicity.

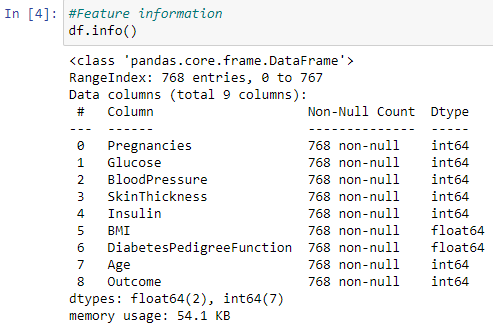
A probabilistic machine learning model called a Naive Bayes classifier is utilized for classification tasks. The Bayes theorem serves as the foundation of the classifier.

## 4.4 Explanatory Data Analysis

All the data obtained from The National Institute of Diabetes and Digestive and Kidney Diseases was collected, analyzed and interpreted with the help of python programming language. The data visualization is done with the help of built-in libraries in python programming.

### 4.4.1 Feature information

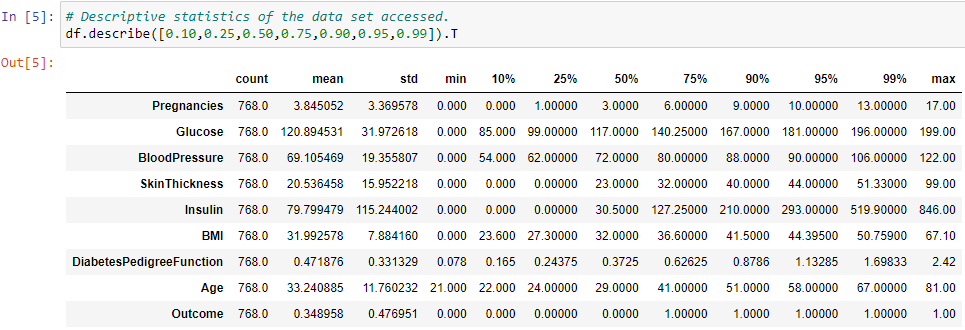
The diabetic dataset contains 8 attributes namely number of pregnancies the patient has had, their BMI, insulin level, and age with 767 entries. Figure 2 summarizes the feature information with data types.



**Figure 2 feature information of dataset**

### 4.4.1 Descriptive statistics

The descriptive statistics summarizes the attributes of diabetic dataset. Figure 3 summarizes the central tendency, mean, median and the measure of variability.



**Figure 3 Descriptive statistics**

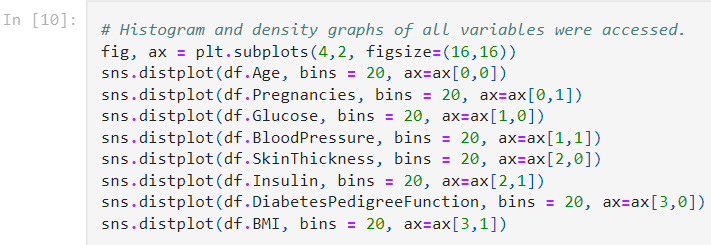
## 4.5 Data Visualization

Advanced analytics relies heavily on data visualization. When a data scientist is developing complex predictive analytics or machine learning (ML) algorithms, it's critical to display the outputs in order to track outcomes and confirm that models are working as expected. This is because sophisticated algorithm visuals are often simpler to understand than numerical results. Big data visualization typically goes beyond traditional visualization approaches like pie charts, histograms, and business graphs. Various data visualization of diabetic dataset are discussed below.

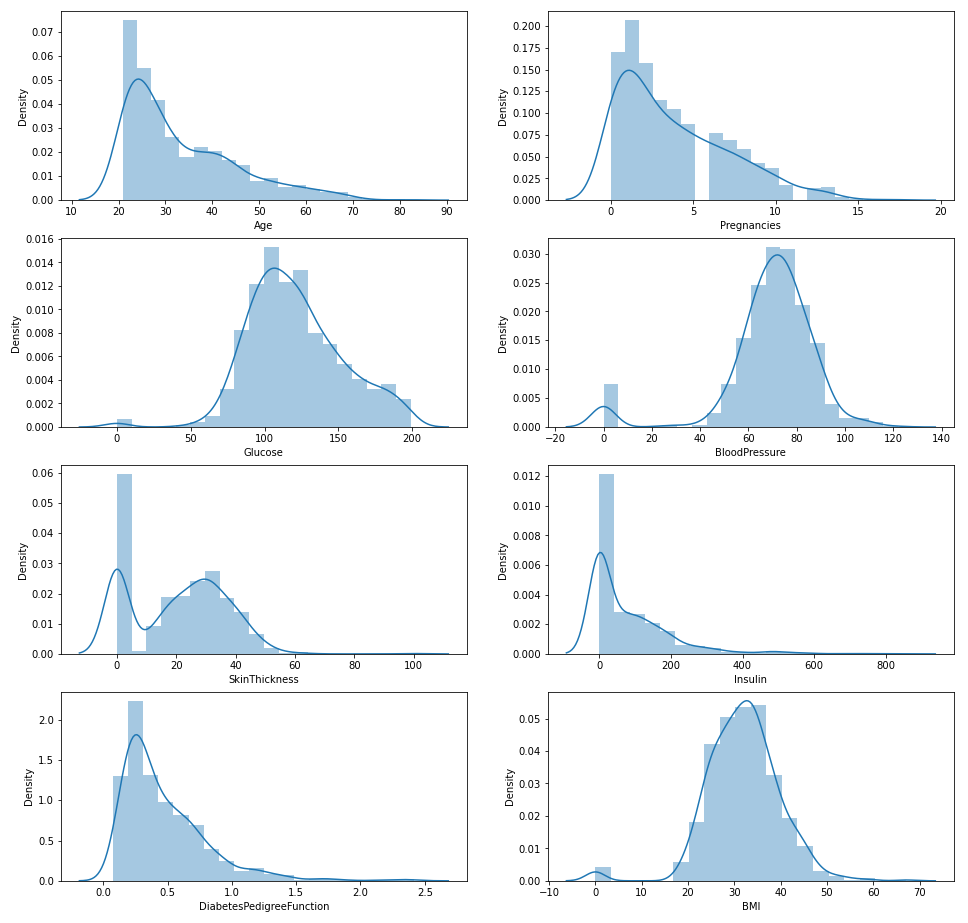
### 4.5.1 Histogram and density graph

The most popular graph for displaying frequency distributions is a histogram. Here the frequencies are absolute frequencies and the attributes are variable. The distribution of data across a continuous interval or time period is shown using a density plot. Figure 4 shows a typical histogram

graph.



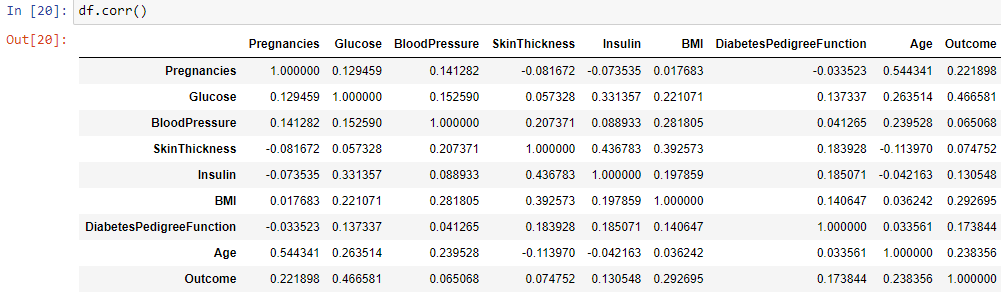
**Figure 4 Histogram and density graph**



### 4.5.2 Correlation Matrix

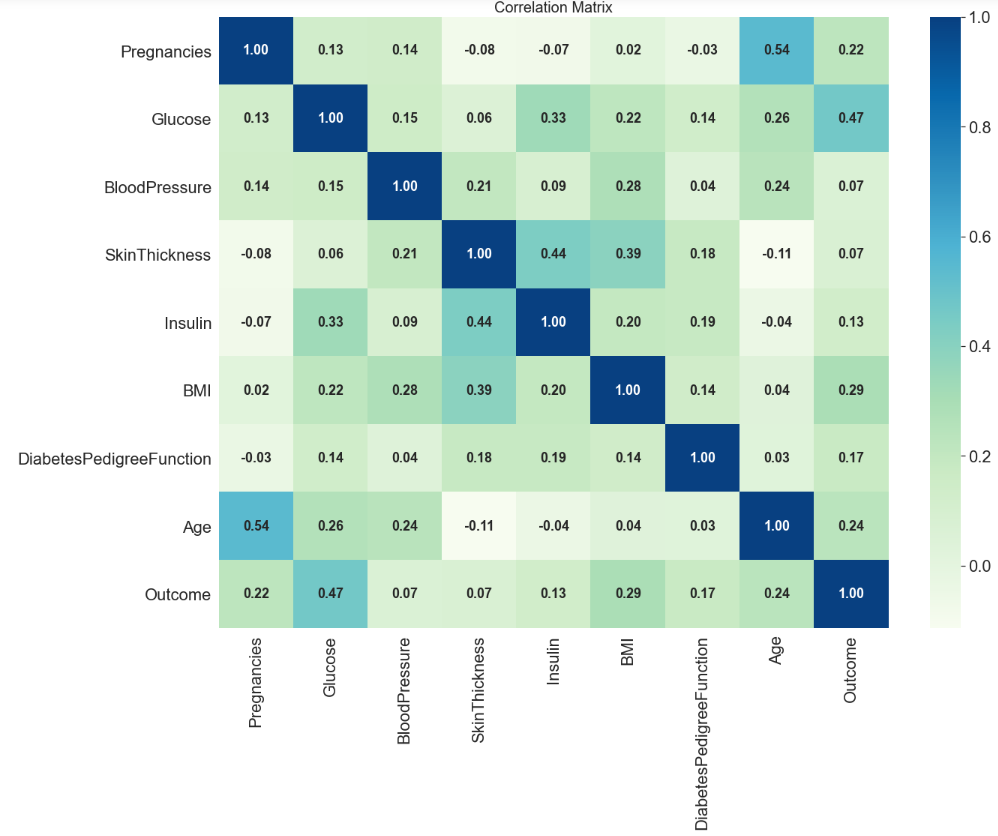
The coefficients for different variables of diabetes data set is explained by correlation matrix. A correlation is said to be positive if the correlation value is greater than zero. Hence the value of variable increases with the other variable. If the value of the correlation is zero it has been concluded that there is no relationship between the variable. Table 4 shows a correlation matrix of diabetes dataset.

**Table 4 Correlation Matrix**



When the correlations are examined, there are 2 variables that act as a positive correlation to the dependent variable.

Figure 5 shows the correlation matrix with various variables and Glucose has the positive correlation of +0.47. As the glucose level increases, outcome variable increases.



**Figure 5 Correlation Matrix**

### 4.5.3 Pie and Bar chart for distribution of outcomes

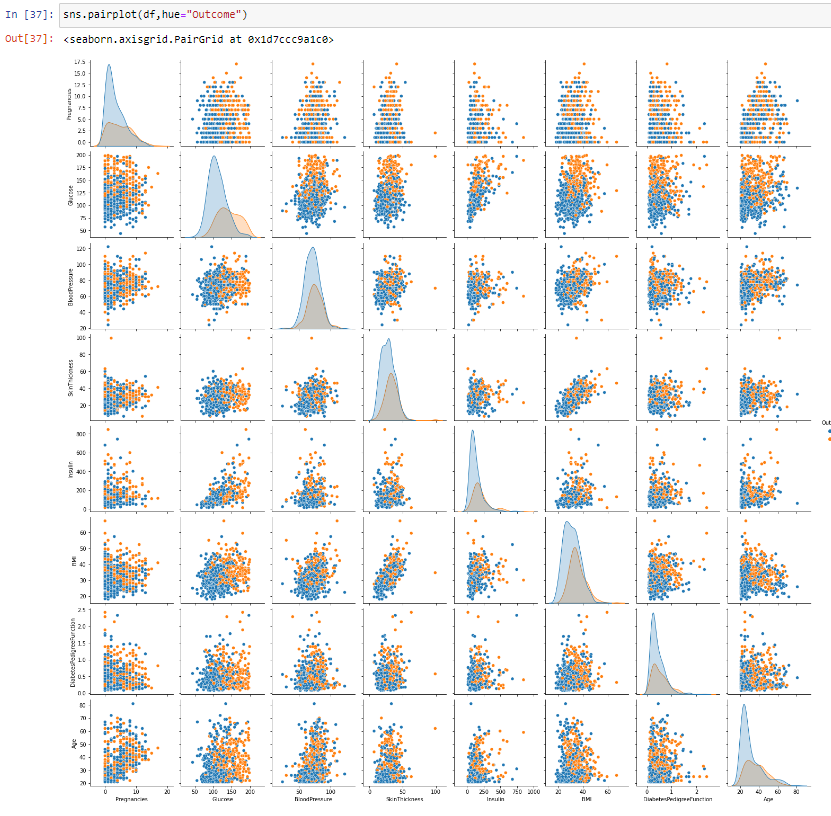
A statistical tool called logistic regression has been used to classify the outcomes. The distribution of outcomes is shown in the figure 6, It has been seen that 65.1% of females has been diagnosed with no diabetes and 34.9% of females has been diagnosed with diabetes as per the data obtained.



**Figure 6 Distribution of Outcomes**

### 4.5.4 Pairplot – relationship between variables

The pair plot is used to figure out which attributes are best for explaining a connection between two variables or for forming the most separated groups. Drawing some basic lines or making linear separation in our data-set also aids in the formation of some simple classification models. From the picture below, we can observe the variations in each plot. The plots are in matrix format where the row name represents X axis and column name represents the Y axis. The main-diagonal subplots are the univariate histograms (distributions) for each attribute.



**Figure 7 Pairplot**

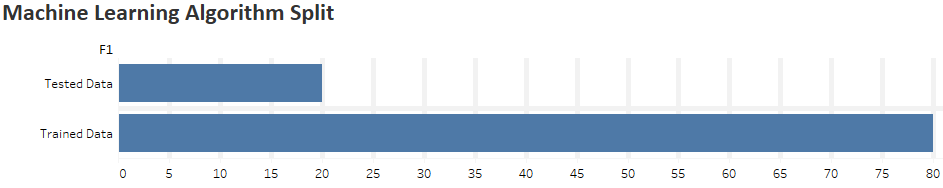
# **Chapter 05: Results and Discussions**

This chapter deals with the results obtained from the diabetic data collected in the study. The results and analysis are represented by using data visualization charts.

## 5.1 Results and Discussion

Diabetes is a condition that has a lot of potential consequences. It might be worthwhile to research how to precisely anticipate and diagnose this condition using machine learning. Particularly in the diabetic dataset, the outcome with fasting glucose shows a higher performance. It follows that while fasting glucose is the most crucial index for predicting. Using just fasting glucose alone won't get the best results, hence additional indices are required to make accurate predictions. Additionally, while comparing the outcomes of the various classifications, it has been observed that Logistic Regression, Decision Tree, Random Forests, and SVM (Support Vector Machine) don't really differ much from one another, however Naive Bayes are undoubtedly superior to the other classifiers. The best result for diabetic dataset is 0.7727. It suggests that machine learning may be used to predict diabetes, but it's crucial to identify the right features, classifier, and data mining technique. It was unable to predict the type of diabetes based on the diabetic data, therefore in the future we want to do so while also examining the proportion of each parameter, which may increase the accuracy of forecasting diabetes.

The trained and tested data split in the ratio of 0.8 and 0.2.

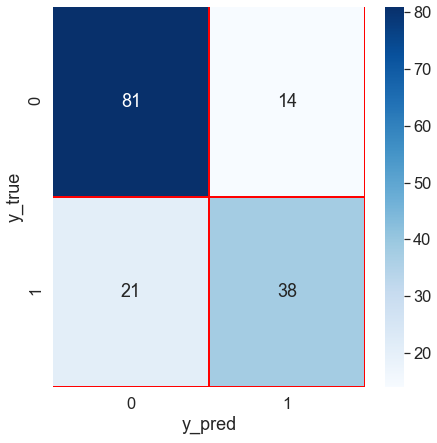


Following the use of several machine learning algorithms on the diabetic dataset, the accuracy results are as follows.

**Table 5 Accuracy Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl Num** | **Algorithm** | **Trained Data Model Accuracy** | **Tested Data Model Accuracy** |
| 01 | Logistic Regression | 0.7597 | 0.5681 |
| 02 | Support Vector Machine (SVM) | 0.7662 | 0.5878 |
| 03 | Naive Bayes | 0.7727 | 0.5985 |
| 04 | Decision Tree | 0.7597 | 0.5680 |
| 05 | Random Forest Classification | 0.7532 | 0.5585 |

Figure 8 shows the confusion matrix of Naive Bayes



**Figure 8 Confusion Matrix**

**Table 6 Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
| **Actual** |  | Non-Diabetic | Diabetic |
| Non-Diabetic | 81 | 14 |
| Diabetic | 21 | 38 |

(0, 0) = True Negative (TN)

(1, 1) = True Positive (TP)

(0, 1) = False Positive (FP) -> Type 1 Error

(1, 0) = False Negative (FN) -> Type 2 Error

Table 6 shows the confusion matrix of Naive Bayes. It has been observed that the Machine Learning algorithm has predicted most of the data with precision.

### 5.1.1 Methods to improve the accuracy of the machine learning model

* **Add More Data**

More information is certainly beneficial. Instead, than depending on hypothesis and flimsy correlations, it lets the "evidence tell for itself." Better and more accurate models are produced when there is more data.

* **Treat missing and Outlier values**

The unintentional inclusion of missing and outlier values in training data frequently decreases model accuracy or results in biased models. It results in wrong forecasts. This is a result of improper behaviour and relationship analysis with respect to other factors. Therefore, it's crucial to handle missing and outlier values with care.

* **Feature Engineering**

By using this process, current data may be expanded upon with new information. In terms of new features, fresh information is extracted. These characteristics could be better able to explain the variation in the training set of data. resulting in increased model accuracy.

* **Feature Selection**

Finding the optimal collection of features that best illustrates how independent variables relate to the target variable is a process called feature selection.

* **Multiple algorithms**

The optimal strategy to increase accuracy is to target the appropriate machine learning algorithm. It is, however, simpler said than done.

With practice and experience, one develops intuition. Certain types of data sets are more suited to certain algorithms than others. Therefore, we should use all pertinent models and evaluate the results.

# Chapter 06: Conclusions and Future Directions

This chapter on conclusion deals with the key findings observed from the study. It also contains recommendations, limitation of the study and suggestion for future direction.

## 6.1 Conclusion

Although there has been a significant amount of study into developing ways to predict incident diabetes, the bulk of these strategies rely on traditional statistical methods. Machine learning techniques are getting popular and attracting the interest of the healthcare sector. The promise of machine learning approaches for predicting incident diabetes using diagnostic data is demonstrated in this work. The System model is primarily concerned with detecting diabetes using a few factors. Physicians may utilize the system to anticipate diabetes in its early stages. So that patients can receive traditional treatments and remedies. To generate more exact findings, the system applied techniques such as Machine Learning (ML) for prediction. There has been a lot of research into the diabetic imprint. For hospitals and physicians, developing a diabetic disease prediction system is beneficial. The system predicts disease at an early stage, allowing doctors to better treat patients. The proposed model is a real-time application that is designed for several hospitals that predicts disease in a shorter amount of time. Hence receive more accurate and efficient findings when employ Machine Learning (ML) algorithms for illness prediction.

## 6.2 Limitations

* Difficulties in obtaining data from the respondents as they were very conscious.
* The in-depth study could not be done due to shortage of time and sensitivity of the topic.
* The sample size was restricted.
* The study was limited to the female who had pregnancies.
* The data collected is based on the inputs given by the National Institute of Diabetes and Digestive and Kidney Diseases and it could be biased.

## 6.3 Future Directions

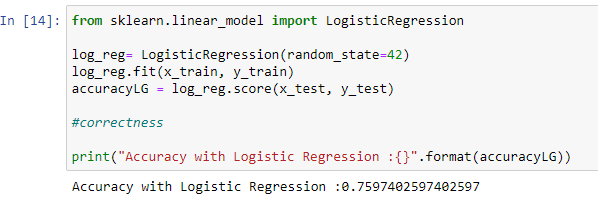
* This study can be extended to inpatients at hospitals and also hospital diagnostic centers.
* A similar kind of study could be carried out in different geographical location so that the results can be generalized.
* A study could be conducted for a longer duration of time.

# Appendix

**Appendix A**

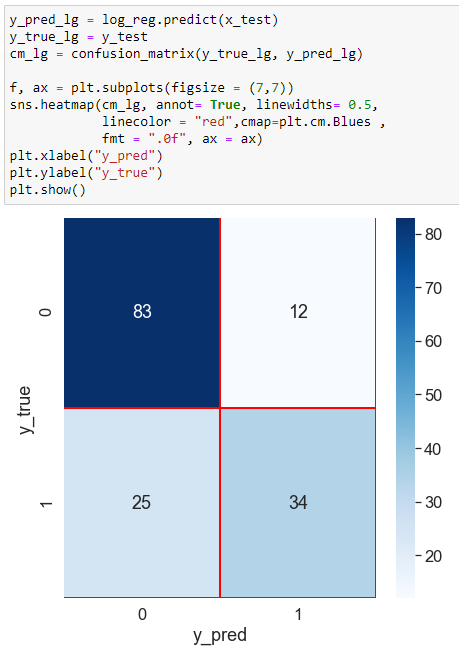
**Logistic Regression Algorithm**

**Predicting the accuracy of the Logistic Regression Model**



**Figure 9 Logistic Regression Model**

**Confusion Matrix of the Logistic Regression Model**

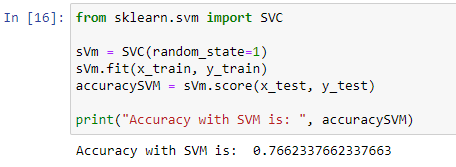
****

**Figure 10 Logistic Regression Confusion Matrix**

**Appendix B**

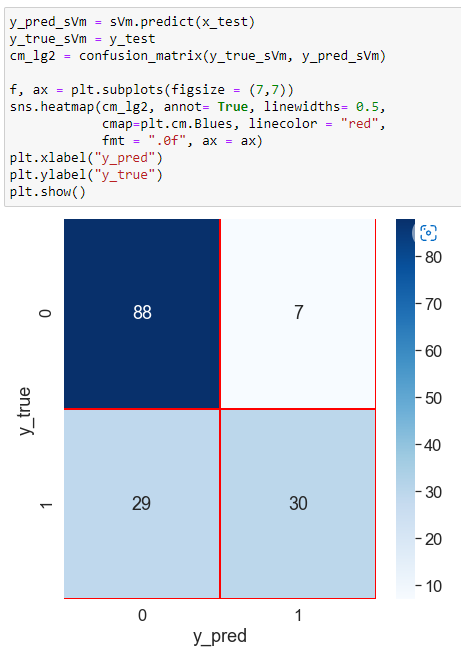
**Support Vector Machine Algorithm**

**Predicting the accuracy of the Support Vector Machine (SVM)**

****

**Figure 11 Support Vector Machine Model**

**Confusion Matrix of the Support Vector Machine (SVM) Model**

****

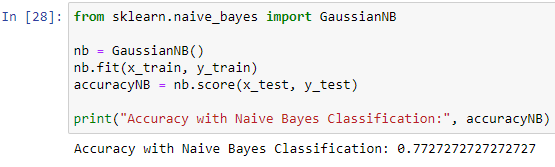
**Figure 12 Support Vector Machine Confusion Matrix**

**Appendix C**

**Naive Bayes**

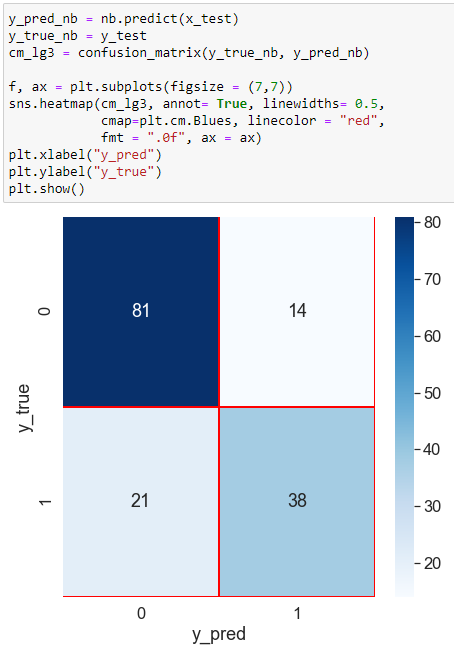
**Algorithm**

**Predicting the accuracy of the Naïve Bayes Model**

****

**Figure 13 Naive Bayes Model**

**Confusion Matrix of the Naive Bayes Model**

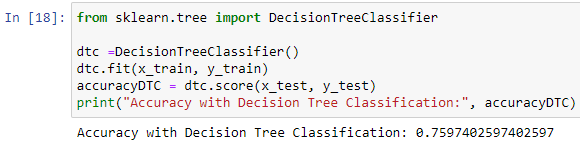
****

**Figure 14 Naive Bayes Confusion Matrix**

**Appendix D**

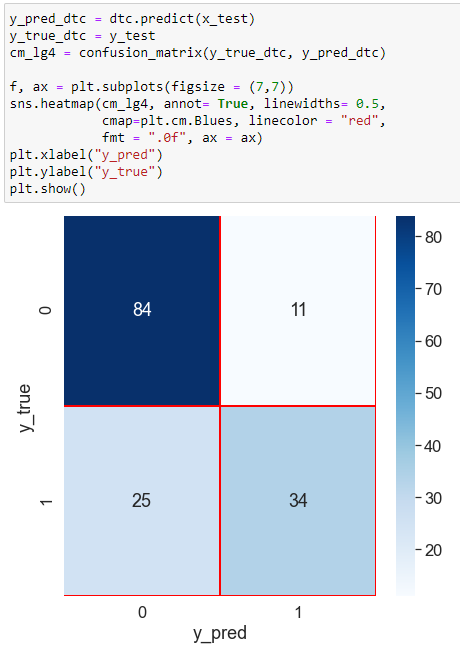
**Decision Tree Classifier Algorithm**

**Predicting the accuracy of the Decision Tree Classifier Model**

****

**Figure 15 Decision Tree Classifier Model**

**Confusion Matrix of the Decision Tree Classifier Model**

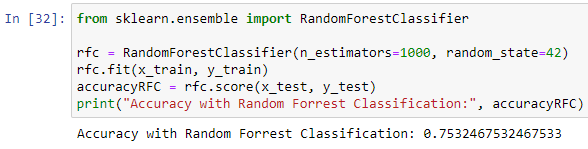
****

**Figure 16 Decision Tree Classifier Confusion Matrix**

**Appendix E**

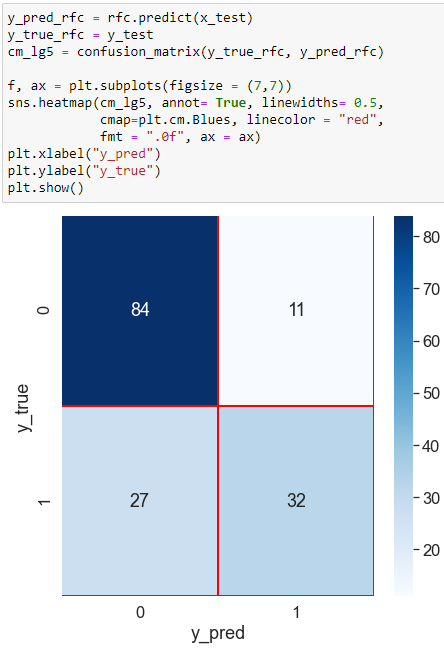
**Random Forest Classifier Algorithm**

**Predicting the accuracy of the Random Forest Classifier Model**

****

**Figure 17 Random Forest Classifier Model**

**Confusion Matrix of the Random Forest Classifier Model**

****

**Figure 18 Random Forest Classifier Confusion Matrix**

References

1. Zohuri, B. and Moghaddam, M. (n.d.). From Business Intelligence to Artificial Intelligence. doi:10.32474/MAMS.2020.02.000137.
2. Jiang, Fei; Jiang, Yong; Zhi, Hui; Dong, Yi; Li, Hao; Ma, Sufeng; Wang, Yilong; Dong, Qiang; Shen, Haipeng; Wang, Yongjun (2017). Artificial intelligence in healthcare: past, present and future. BMJ, (), svn-2017-000101–. doi:10.1136/svn-2017-000101
3. Zeithaml, V.A., Rust, R.T. and Lemon, K.N., 2001. The customer pyramid: creating and serving profitable customers. California management review, 43(4), pp.118-142.
4. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics, 9. doi:10.3389/fgene.2018.00515
5. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S (2017) Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS ONE 12(7): e0179805. [https://doi.org/10.1371/journal. pone.0179805](https://doi.org/10.1371/journal.%20pone.0179805)
6. Wamba-Taguimdje, Serge-Lopez; Fosso Wamba, Samuel; Kala Kamdjoug, Jean Robert; Tchatchouang Wanko, Chris Emmanuel (2020). Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. Business Process Management Journal, ahead-of-print(ahead-of-print), –. doi:10.1108/bpmj-10-2019-0411
7. Ghahramani, Mohammadhossein; Qiao, Yan; Zhou, MengChu; O Hagan, Adrian; Sweeney, James (2020). AI-based modeling and data-driven evaluation for smart manufacturing processes. IEEE/CAA Journal of Automatica Sinica, (), 1–12. doi:10.1109/JAS.2020.1003114
8. Higgins, David; Madai, Vince I. (2020). From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. Advanced Intelligent Systems, (), 2000052–. doi:10.1002/aisy.202000052
9. Long, Guang Jin; Lin, Bai Hua; Cai, Hong Xing; Nong, Guang Zai (2020). Developing an Artificial Intelligence (AI) Management System to Improve Product Quality and Production Efficiency in Furniture Manufacture. Procedia Computer Science
10. Coombs, Crispin; Hislop, Donald; Taneva, Stanimira K.; Barnard, Sarah (2020). The strategic impacts of Intelligent Automation for knowledge and service work: An interdisciplinary review. The Journal of Strategic Information Systems, (), 101600–. doi:10.1016/j.jsis.2020.101600