Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 8s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

AI and Data Analytics for Proactive Healthcare Risk Management

G. Sabeena Gnanaselvi[1], D. Sandhya[2], Dr R Murugesan[3], Dr. M. Geetha[4], Boyina Kavya[5], Dr. Parashuram S. Vadar[6]

[1]Designation: Assistant Professor, Department: computer science and engineering Institute: sathyabama Institute of science and technology, District: kancheepuram, City: Chennai, State: Tamil Nadu

Emai ID: sabisamuel33@gmail.com

[2]Designation: Assistant Professor, Department: Faculty of Allied Health Science, Institute: Dr MGR Educational and Research Institute University, District: City: Chennai, State: Tamil Nadu,

Emai ID: sandyy.jas@gmail.com

[3]Professor, Department of AI&DS, VSB Engineering College, Karur 639111

Emai ID: rmurugesan61@gmail.com

[4]Designation: Associate Professor, Department: Computer Science and Engineering, Institute: Koneru Lakshmaiah Education Foundation, District: Guntur, City:Guntur, State:Andhra Pradesh

Emai ID: geethasaravanan@kluniversity.in

[5]Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,

Vaddeswaram, Guntur District, Andhra Pradesh, 522302, India.

Email ID: kavyaboyina@gmail.com

[6]Designation: Assistant Professor, Department:Yashwantrao Chavan School of Rural Development Shivaji University Kolhapur, Institute: Yashwantrao Chavan School of Rural Development Shivaji University Kolhapur, District: Kolhapur, City: Kolhapur, State: Maharashtra

Cite this paper as: G. Sabeena Gnanaselvi, D. Sandhya, Dr R Murugesan, Dr. M. Geetha, Boyina Kavya, Dr. Parashuram S. Vadar, (2025) AI and Data Analytics for Proactive Healthcare Risk Management. *Journal of Neonatal Surgery*, 14 (8s), 797-813.

ABSTRACT

This work aims at examining the use of AI in anticipation of diabetes and the use of data analysis. To achieve this, we trained different classifiers using a broad Diabetes Detection Dataset which includes the following; Logistic Regression, K Nearest Neighbor (KNN), Random Forest Classifier, and Decision Tree Classifier. The observation made from the analysis showed that the Random Forest Classifier yielded the highest overal accuracy of 96. It was found that 82% of the population was affected, with a precision of 0.94 and a recall of 0.69 for positive cases. The overall performance of KNN model was also impressive it had a precision of 0.91 and a recall of 0.62, while LR and DTC gave useful information about the data but did not perform so well in some of the evaluation measurements. Some of these items included correlation heatmaps and ROC curves that helped in capturing the relationship between diabetes and other health aspects such as blood glucose level, HbA1c and model performance. It confirms that the concept enshrined in the AI-technologies actually has a huge potential in the early diagnosis and intervention programs that will eventually lead to efficiency enhancement and harmonization of health care services delivery. Future research should hence focus on issues to do with the generalizations of the models as well as ways of combining data in order to enhance the level of predictive and health care improvement. This study predicts diabetes with the aid of machine learning models. Exploratory data analysis shows that the major predictors of diabetes are age, blood glucose, and hypertension. The implementation of the models Logistic Regression, KNN, Random Forest, and Decision Tree was done. Random Forest turned out to be the best model being very accurate and precise in predicting negative cases and quite reasonable in performance for positive case detection. This research contributes to the early detection of diabetes and potentially better treatment for patients.

Keywords: *Machine Learning, Diabetes Prediction, Data Analytics, Predictive Models, Healthcare Management*

1. INTRODUCTION

Background

In the setting of the current healthcare environment, both AI and data analysis play crucial role in developing a new approach to patients' treatment and improving the techniques of risk assessment and prediction. One of the most commonly observed chronic diseases is diabetes, which shall serve as an example of the importance of the early diagnosis and treatment thereof. As its occurrence is increasingly noted around the world, diabetes poses major difficulties in the field of health care, which means that new strategies and techniques have to be developed in order to avoid and better treat this disease [3]. Earlier methods used in identifying diabetes are mostly receptive methods, normally, this results to late diagnoses, and complications. Through the help of AI and data analytics, it is now possible to move closer to a preemptive security style where security threats are prevented from reaching optimum and extreme levels. AI and data analytics act as helpful tools for improving the probability of presenting disease diagnosis and treatment [4]. It is the case that machine learning systems are capable of processing large amounts of data in orders of magnitude more than a human doctor can do and finding correlations which could be unnoticed otherwise. They can be used in early risk assessment of disease occurrence in a patient and planning of effective preventative and control measures. That is why this approach not only enhances the patients' quality of life but also may help to manage the healthcare costs, for instance, required for treating terminal diseases.

Aim and Objective

Aim:

The main aim of this research is to identify how AI and data analytical tools can be utilized to prevent diabetes risk. With the help of the tested and used machine learning models and analytical approaches, this work aims at improving the existing approaches towards detection and providing a better quality in the treatment.

Objectives:

1. To conduct evaluation of different machine learning models as to their accuracy in estimating the probability of developing diabetes based on medical and other characteristics.
2. To check the presence of diabetes and its relationship with the main clinical characteristics including blood and HbA1c levels and other existing pathologies, including hypertension and cardiovascular disease.
3. To analyze accuracy, precision, recall and F1-score of different algorithms that are logistic regression, KNN classifier, Random Forest classifier, Decisions tree classifier for the purpose of prediction of diabetic patients.
4. To offer insight into the implementation of AI-DMT predictive models into healthcare systems for effective prevention and management of diabetes.

2. LITERATURE REVIEW

2.1 Advancements in Machine Learning for Disease Prediction

Modern developments in ML have greatly improved the methods of disease prediction, more specifically, chronic diseases such as diabetes. These advances have revolutionised the conventional methods by utilizing computational structures to evaluate and diagnose intricate healthcare information for early diagnosis and effective prevention. A significant development is employing deep learning, which is based on neural networks with more than one layer. These models are also efficient at details and complex relationships in the given mass of data [5]. Some of the popular NN models are CNNs which have been used in analysis of Medical Imaging Data, RNNs and LSTM for Sequential Data like patient history and time series health record data. The update also includes the added feature of ensemble models, which uses the prediction generated by a range of models, to increase accuracy and performance. Methodologies such as Random Forests and Gradient Boosting Machines are based on the integration of several algorithms and they help to minimize overfitting as well. These methods have been found to have better prediction parameters bit over single-model schemes. Feature engineering and selection have also come a long way with auto feature selection now being available to pick the best predictors from large datasets. Methods like PCA & t-SNE, help in reducing feature dimensions and helps in focusing only on the most influencing variables of data preparation phase [6]. In addition, transfer learning has made it possible to take advantage of existing practical models to apply them to a new, though highly related task, thereby increasing the rate of creating predictive models where data is scarce. This approach is especially effective in healthcare since it may be rather difficult to get large datasets which are labeled. In general, all these techniques of machine learning have provided a grand platform to the prediction of diseases and helped in the early understanding of the risk factors and thus enhancing the proactive healthcare management approach.

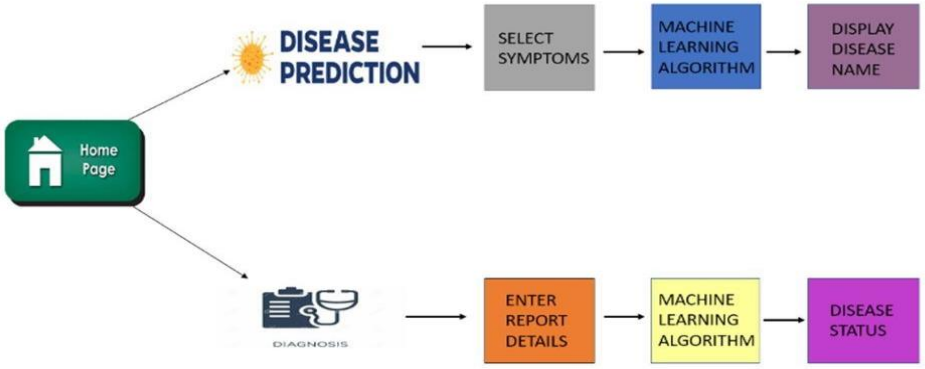

Figure 2.1.1: Multi-Disease Prediction Using MachineLearning Algorithm

2.2 Integration of Data Analytics in Healthcare

The incorporation of data analytics into healthcare has significantly transformed how physicians and other healthcare providers tackle diseases. Together with cautious usage of patients' data, it becomes possible to obtain valuable information to improve the patients' condition and the overall functioning of healthcare systems with the help of data analysis of vast and diverse data sets. Possibly the most common use of data analytics is in predictive analytics, which involves using algorithms to predict future patient events. The approach helps in early screening of those who are likely to develop complications, and thus, receive early treatment. For instance, it can be used to find out those patients who are most likely to develop corresponding chronic diseases including diabetes, and take appropriate action [7]. Data analytics also helps in the improvement of individualized medicines since patient profiles are unique, and therefore should be treated as such. Looking at EHRs data, genomic information and other lifestyle details healthcare providers will be in a position to develop individual treatment plans that will reflect the patient's individual characteristics. This precision medicine enhancing the effectiveness of the treatment and minimizing the side effects of the certain treatment. Moreover, data analytics contributes to service enhancement in healthcare organizations, technically, and operationally. Through the identification of such patterns as the rate and timing of patient visits, treatment, and other service processes, the prospects for health care centres' effciency, decreased waiting time and optimal staffing can be revealed [8]. This results in reduction of cost as well as improving patients' satisfaction. Furthermore, data analytics plays a crucial role in detecting superficial relationships that might exist and help ministry of health and policymakers in making appropriate policy decisions. For instance, the patterns of incidence of an illness can be determined from epidemiological data, or the likelihood factors of occurrence of an ailment can be identified, or the success of public health measures can be assessed and determined from the data collected. In summary, the use of data in health care enables the providers to utilize the information and data gathered for the purpose of making right decisions and improve patient care and overall health system operations. As is already seen, one of the rapidly growing fields, this strategy of data analytics is set to become even more important in the development of the upcoming health systems.

Figure 2.2.1: Data Analytics in Healthcare Industry

2.3 Correlation between Diabetes and Key Health Indicators

The relationships between diabetes and other health indicators should be well understood in order to manage the disease and identify risk factors at an early stage. Diabetes has been associated with several physiological measures that explain the development and the course of the disease especially the Type 2 diabetes. Among all diagnostic determinants, blood glucose level is one of the most crucial ones in diabetes. One of the cardinal manifestations of diabetes is hyperglycemia, and glycemic control is at the heart of diabetes management. All the research points towards high blood glucose levels of increasing the risk of getting diabetes especially type 2 diabetes [21]. These levels should be also keenly observed so that they can be diagnosed and treated early. There is HbA1c, which is a specific form of hemoglobin, involved in determining average blood glucose concentration in the body during the past 2-3 months. HbA1c > 7% is an index of suboptimal long-term glycemic regulation, and the presence of chronic diabetes end-organ complications is highly linked [9]. Over the years, there has been evidence that links HbA1c levels to the risk of having diabetes and for this reason, the tool has been proving relevant and vital in diagnosis as well as in management of the condition. Another is Body Mass Index (BMI) that is also influential in the changes in human behaviours. We know how obesity raises the chances of getting Type 2 diabetes. An increased BMI means a higher risk of getting diabetes due to the fact that extra body fat, usually in the abdominal area, cause insulin resistance. Hypertension is another characteristic that is related to the disease in question. Metabolic interactions of the components of MS increase the susceptibility of those with high blood pressure to diabetes [10]. The co-relationship between hypertension and diabetes risk shows that both conditions should not be treated in isolation from each other. Last but not the least; the lipid factors such as cholesterol and triglycerides show a relationship with the risk factor of diabetes. Dyslipidemia is another related condition whereby patients who have diabetes also have suspect cholesterol levels that lead to cardiovascular issues. High density lipids as a marker have been found to have a strong relationship with diabetes; their control is crucial. Accordingly, the HA-perceived Blood glucose level, HbA1c, BMI, hypertension, and lipid profiles all precisely bear significant relationship to Diabetes risk and control. Knowledge of these relationships improves the early identification, control measures, and general handling of the illness.

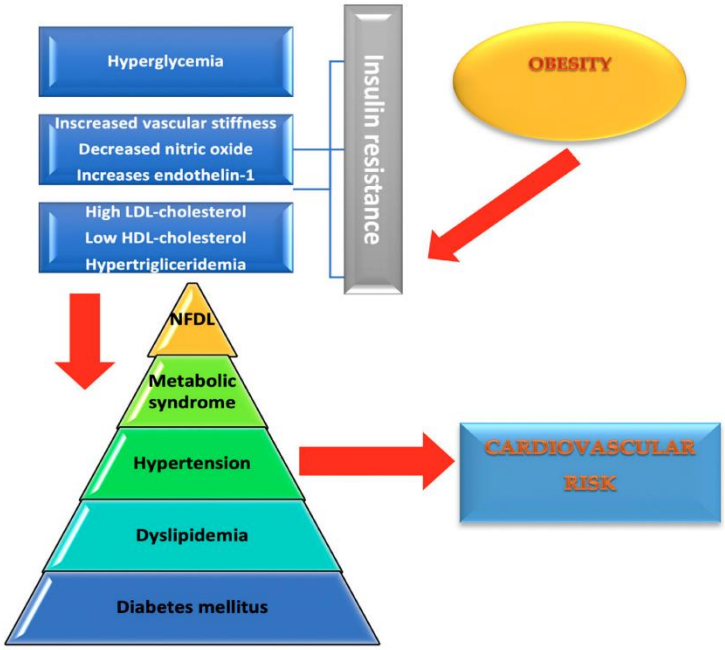

Figure 2.3.1: Current Data Regarding the Relationship between Type 2 Diabetes Mellitus and Cardiovascular Risk Factors

2.4 Comparative Analysis of Predictive Models

In the field of diabetes risk prediction using machine learning methods, several models show performance variations in their efficiencies and limitations. The comparison of these models: Logistic Regression, K Nearest Neighbors, Random Forest, Decision Tree, help in understanding the working and applicability of each model.

Logistic Regression is one of the statistical techniques that are basic in use for binary classification. It is quite good at giving probabilities and fairly easy to explain. Although it is quite basic, it works well in situations, where the dependencies between the features and the outcome are directly proportional. In the field of diabetes prediction, the performance of Logistic Regression has been commendable in terms of accuracy and precision particularly with the negative class [11]. However, it can has some drawbacks when applied to the positive class instances, where it cannot properly model interactions of the

features.

K-Nearest Neighbors (KNN) was based on the concept of similarity, where new instances were classified using the nearest labelled data. The technique is useful for identifying nonlinear relationships and has minimal preconditions about the distribution of the data. The results of using KNN are better for positive cases in terms of precision and recall, but the choice of 'k' and the distance measure affects results [22]. It also demands substantial amount of computation particularly when the data set is large.

Random Forest Classifier is an ensemble model which means it uses the output of multiple decision trees to make the final decision to increase reliability and decrease the chances of over training. It performs well in case of non-linearity and provides good forecast with low error rates. From the tables, Random Forest performs very well for negative instances as well as it has high precision for positive test cases. However, it is not easy to interpret and it normally needs some of its parameters to be tuned for better results.

Decision Tree Classifier is based on the use of a tree, the branches of which are decisions and their potential outcomes. It is generally easy to understand and apply, but its application can sometimes be sensitive to the complexity of the data set, and therefore may easily be over-trained. Nevertheless, using Decision Trees provides good accuracy of both positive and negative instances only if pruned and optimized.

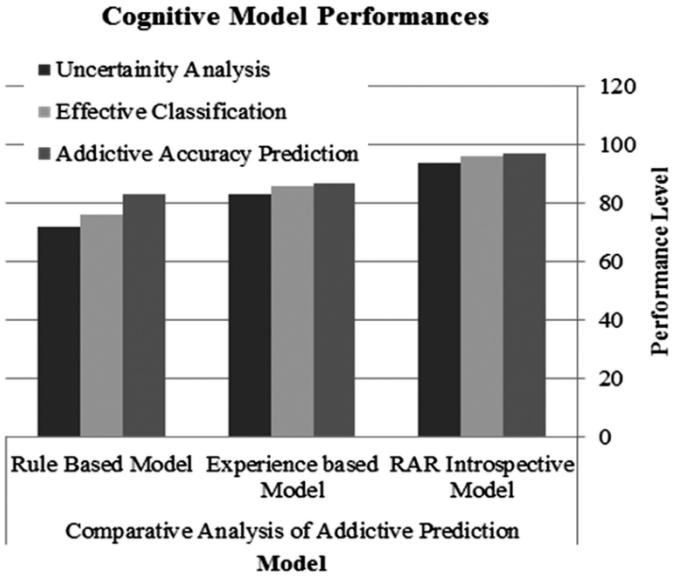

Figure 2.4.1: Comparative analysis of addictive predictive model

2.5 Impact of AI on Proactive Healthcare Management

The incorporation of AI in proactive healthcare management has made another shift in identifying, scrutinizing, and supervising health risks. Machine learning and deep learning are AI technologies that improve the performance of healthcare systems because they facilitate early diagnosis, and individualized management of diseases, thus enhancing patient satisfactions, health and reach health facility productivity. Notably, AI's perhaps the most significant effect is the analysis of large volumes of data from various sources including EHRs, wearable devices, and imaging [12]. Democratization of data means that the AI algorithms can discern patterns and correlations that can go unnoticed by analysts. For instance, probabilities of the onset of such diseases as diabetes or cardiological diseases can be predicted using the parameters of the past and present health conditions. This early detection makes it easy to treat these complications before they progress in order to enhance long term health. AI helps in implementation of personalized medicine by creating individual treatment plans based on the patient's details. AI, based on DNA data, daily habits, and disease history, can introduce individualised treatment and protection measures. It not only increases the efficacy of the treatments but also reduces side effects; thus, the patients get care that meets their needs. In addition, AI improves the flow of doing business in the overall health sector of facilities. Technologic advances and integrations of artificial intelligent technologies result in the improvement of advanced administration, patient flow, and resource allocation. This results in short waiting times cutting operating expenses while at the same time increasing patient satisfaction. In the case of public health AI supports the supervision of disease epidemic, monitoring of vaccine administration, and evaluation of health trends. Given the role of AI for these tasks, health authorities can use public health interventions and policies that are more efficient.

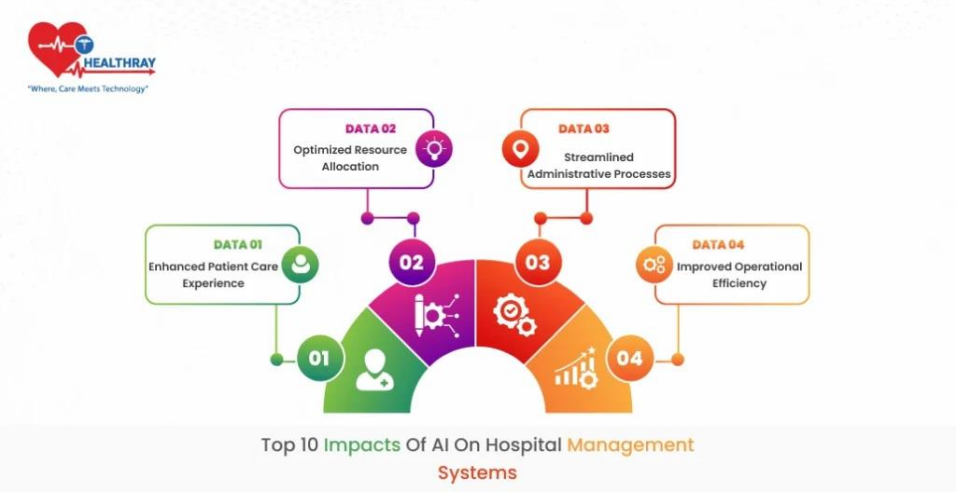

Figure 2.5.1: The Impact of AI on Hospital Management Systems

2.6 Literature Gap

Although contemporary approaches to AI and data analytics continue to improve, there are limitations in the utilization of unstructured data sources and the transferability of predictive models among different populations [13]. Nonetheless, more extensive reviews of the effects of the AI-assisted interventions in the long run on patients' health and costs have not been conducted. It is important for the two to be filled in order to get the best results in AI and in the health sector.

3. METHODOLOGY

3.1 Data Collection and Preprocessing

The collection and preprocessing of data form an important part of the research methodology in predictive modeling in healthcare. This section explains how information was gathered and processed with regards to proactive healthcare risk management.

Data Collection: The primary data for this study were obtained from the Diabetes Detection Dataset, which provides extensive information about patients' clinical characteristics. This dataset contains a number of parameters like age, gender, blood glucose level, blood pressure, BMI, etc [14]. These attributes are important in the assessment of risk and modelling of diabetes. During data acquisition, records were selected from the dataset and the goal was to give the best shot at selecting a diverse patient set to make the model more resilient.

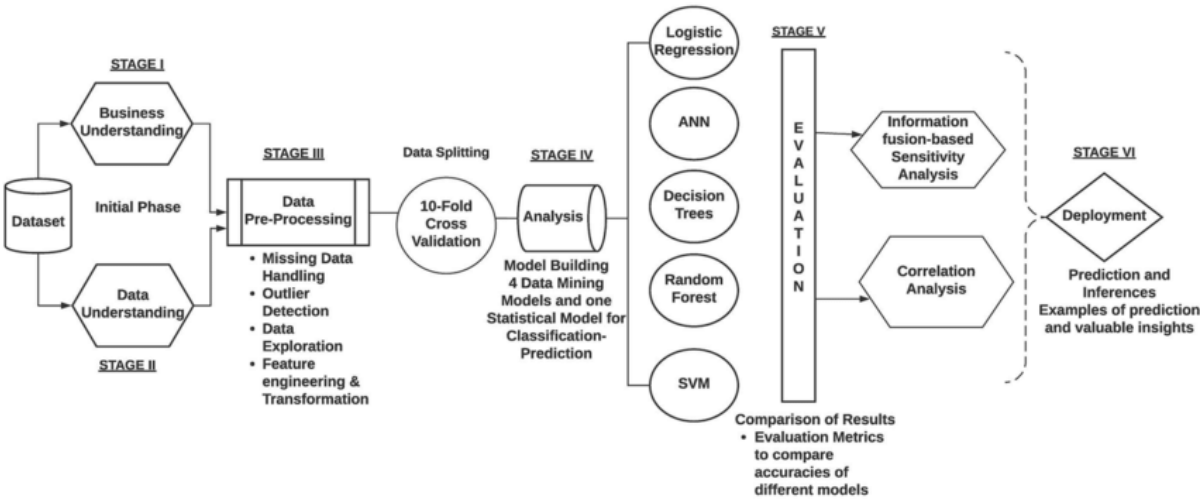

Figure 3.1.1: The flowchart for the predictive analysis of patients' no-shows

Data Preprocessing:

1. **Data Cleaning:** The preprocessing took the following steps where the first step involved handling missing values and duplicates. To prevent distorting of the analysis, any missing values observed in the dataset were dealt with,

through the use of appropriate imputation method such as mean or median imputation. To eliminate any duplication in the dataset, data were checked and cleaned so as to have a neat set of data. After these steps, the data that was cleansed contained approximately 96446 rows of data.

2. **Normalization:** To make the data comparable, normalization was done to the data set so that the variance was brought down towards a moderate level. This process included normalization of the numerical variables including the values of blood glucose and BMI to the range of 0 to 1 [15]. Scalar quantization is relevant to the kind of machine learning models that work based on the magnitude of the data such as k-Nearest Neighbors (KNN) and neural networks.
3. **Feature Selection:** These features were selected bearing in mind the importance of each of them in the diagnosis of diabetes. This step involved assessing the degree of relationship between features and the target variable that was diabetes and the process of feature selection where certain features irrelevant or those that are redundant were disposed to enhance the models performance as well as to do away with computational expenses.
4. **Data Splitting:** The dataset was split into training and testing sets. Most of the time, the data was split into 70-80% training data set used for training the models while 20-30% of the data set was used to evaluate or test the models.

Logistic Regression

$p=1+e-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn)/1$

> ***"Initialize coefficients (β) to zero***
>
> ***Repeat until convergence:***
>
> ***Compute predictions using logistic function***
>
> ***Calculate error (difference between predicted and actual values)***
>
> ***Update coefficients using gradient descent***
>
> ***Return coefficients"***

k-Nearest Neighbors (KNN)

$d(x,x')=\sum i=1n(xi-xi')2$

> ***"For each test instance:***
>
> ***Calculate distances to all training instances***
>
> ***Sort distances and select the k-nearest neighbors***
>
> ***Assign the class based on majority vote of the neighbors***
>
> ***Return predicted class for each test instance"***

Random Forest Classifier

$Prediction=T1\sum t=1TDecisionTreet(x)$

> ***"Initialize number of trees (T) and tree depth***
>
> ***For i = 1 to T:***
>
> ***Sample with replacement from the training dataset (bootstrap sampling)***
>
> ***Build a decision tree on the sampled data with random feature selection***
>
> ***Return aggregated predictions from all trees"***

Decision Tree Classifier

$Gini(p)=1-\sum i=1kpi2$

"BuildTree(Node):

If all samples at Node belong to the same class:

Return a leaf node with that class

If no features left to split:

Return a leaf node with the most common class

For each feature:

Compute the best split (e.g., using Gini impurity)

Choose the best feature and split

Recursively apply BuildTree to each child node

Return the root node"

3.2 Model Development

Model development is an important process of building effective models for early healthcare intervention. In this process, one involves using data of patients to train machine learning models to predict the chances of developing diabetes [16]. In the following section, the various stages in development of the model are described below.

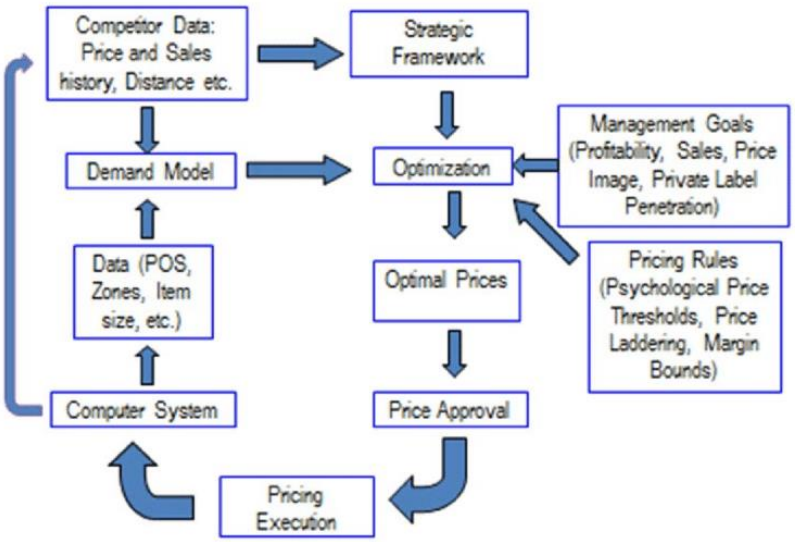

Figure 3.2.1: Flowchart for a Customer-Centric Predictive Analytics and Optimization

1. Model Selection: Several machine learning algorithms were chosen for the use of building models. Some of these are Logistic Regression, k-Nearest Neighbors also known as KNN, Random Forest Classifier as well as the Decision Tree Classifier. Every model presents different advantages.

- **Logistic Regression** is preferred due to its interpretability and ability to work well with binary classification problems.
- **K Nearest Neighbors (KNN)** is a good approach to identify non-linear relationships based on nearest neighbor classification.
- **Random Forest Classifier** averages the outputs of different decision trees to improve the model, as well as to accommodate variant and multiple data correlation.

- **Decision Tree Classifier** can be easy to understand featuring a decision tree which detailed decision-making process while highlighting the feature importance.

2. Model Training: Training phase includes feeding preprocessed dataset into every model in order to fit them. This process included:

- **Logistic Regression:** Training was done with the optimization algorithms to get the right parameters that enhanced classification with minimal errors.
- **k-Nearest Neighbors (KNN):** The model was trained by calculating the correct value for k, and the choice of the distance function to be employed in the classification process.
- **Random Forest Classifier:** Several decision trees were built from different random sub-samples of the data, and the results from each of them were then combined to make the predictions.
- **Decision Tree Classifier:** The model was trained in order to divide data by the feature values to gain the maximum classification accuracy of nodes.

3. Model Validation: Validation was done in a way where each of the developed models was tested on the test subset of the data set. Some of the measured performance indicators were accuracy, precision, recall and F1 score was used. These measures ensured it was possible to check on the ability of every model to estimate diabetes risk as well as its ability to perform on positive and negative instances.

4. Model Tuning: The decision was made to tune them, so if it is necessary the parameter value is adjusted. Techniques like grid search and cross-validation were also applied for getting the best parameters of each model to enhance their prediction and to have more generalized results.

3.3 Visualization

Visualization is a very important aspect of any analysis that involves data and models whereby patterns and merits of the model as well as data relationships can be easily identified [17]. To increase readability and facilitate analysis, interpretation of the results yielded by the prospective models of diabetes risk management, a number of visualization methods were used in this study.

1. Data Distribution and Cleaning: First of all, histograms and bar plots were employed to investigate the distribution of the main variables in the samples. For example, statistical figures including histograms helped to determine the current status of different range of blood glucose level and BMI. Furthermore, the plots for showing the data preprocessing, also focused on the deletion of repeat records and treatment of 'missing' values within the records.

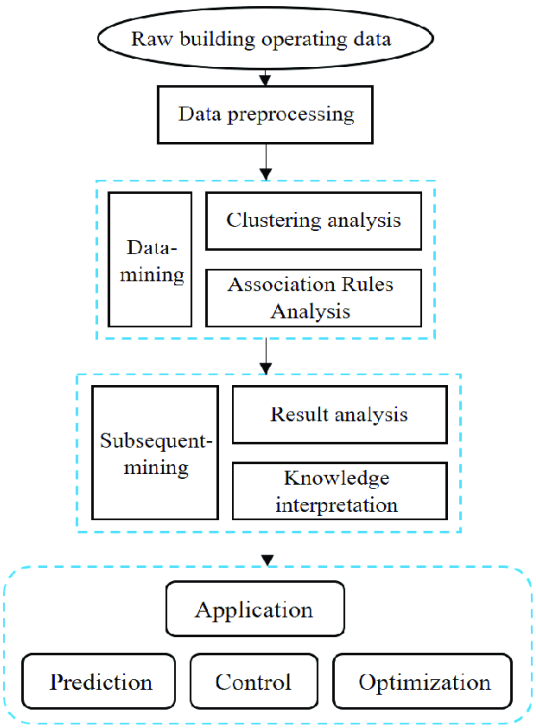

Figure 3: The flow chart of data analytics in building

2. Feature Relationships: Two models were used to explore the connections between the features and the target (diabetes), these include scatter plots and correlation heat maps. Least some of the graphical displays included scatter plots which demonstrated the rising upward trend in incidences of diabetes as glucose levels rose [18]. Since the correlation heatmap gave current insights of health indicators such as HbA1c levels, age and blood pressure concerning diabetes. It was possible to observe some trends, for example the positive association between the HbA1c levels and diabetes.

3. Model Performance: For the assessment of the model, confusion matrices and ROC curves were used. In form of confusion matrices, the classification outcome of every model was clearly shown, by means of true positive values, false positive values, true negatives and false negatives. Receiver operating characteristics or ROC curves depicted the true positive and the false positive for different models in relation to its discriminant capacity [19]. Also the bar charts which compare the accuracies, precisions, recalls and F1-scores of each of the models made for a clear comparison of model efficiency.

4. Insights and Interpretation: Data visualization was useful in transforming figures and other data, as well as results from the model into useful information. Due to the fact that show-case presentations of key findings were not possible within the frameworks of the present work, one could illustrate the consequences of high glucose levels towards diabetes or the performance of various predictive models to the respective stakeholders [20].

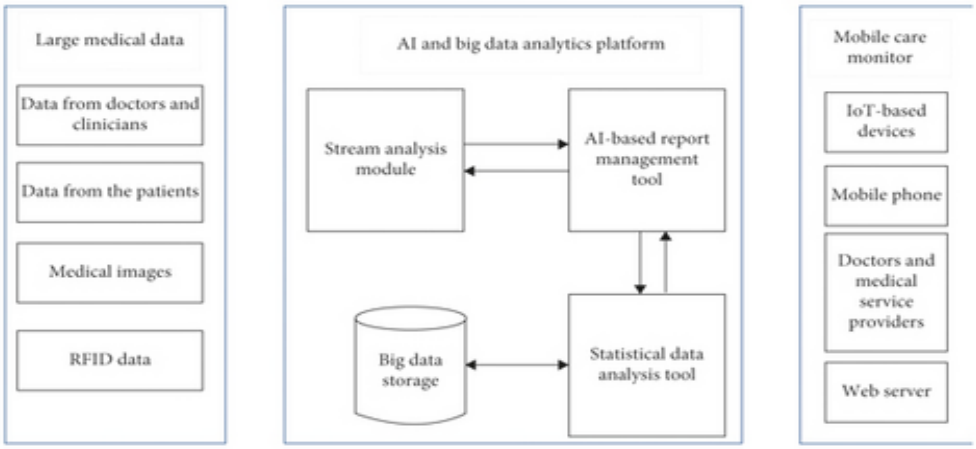

Figure 4: Applications of Artificial Intelligence and Big Data Analytics in m-Health

4. RESULTS AND DISCUSSION

4.1 Results

Diabetes detection dataset has been selected to decrease the risk of healthcare and data analytics, as well as AI helped in improving the quality of patient care. Diabetes is a common issue in recent times and it can increase severe complexity in global health. Relevant proactive management such as AI and data analytics helps in improving the early disease detection method and determining the prevention method. The results section helps to identify the factors that provoke the occurrence of diabetes. The main objective behind this analysis process is to predict diabetes based on the related factors before fatality. This may help the patient to suffer less and recover at an early stage. Machine Learning models help in data driven decision making process by adapting changing conditions [2]. Thus, several “machine learning models” have been developed for the sake of proactive healthcare risk management to predict the occurrence of diabetes based on other medical parameters.

```
# Duplicate value checking
healthcare.duplicated().sum()
```

```
3854
```

```
healthcare = healthcare.drop_duplicates()
```

```
# Null value checking
healthcare.isnull().sum()
```

	0
gender	0
age	0
hypertension	0
heart_disease	0
smoking_history	0
bmi	0
HbA1c_level	0
blood_glucose_level	0
diabetes	0

Figure 4.1: Data cleaning process

The above figure supports that the data frame is free from any of the null values. This one contains 3854 duplicated entries which are eliminated so that the distortions or violations from test results cannot be possible. It can be noticed that 96146 rows of data are used for further analysis after removing duplicate entries.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
count	96146	96146.000000	96146.000000	96146.000000	96146	96146.000000	96146.000000	96146.000000	96146.000000
unique	3	NaN	NaN	NaN	6	NaN	NaN	NaN	NaN
top	Female	NaN	NaN	NaN	never	NaN	NaN	NaN	NaN
freq	56161	NaN	NaN	NaN	34398	NaN	NaN	NaN	NaN
mean	NaN	41.794326	0.077601	0.040803	NaN	27.321461	5.532609	138.218231	0.088220
std	NaN	22.462948	0.267544	0.197833	NaN	6.767716	1.073232	40.909771	0.283616
min	NaN	0.080000	0.000000	0.000000	NaN	10.010000	3.500000	80.000000	0.000000
25%	NaN	24.000000	0.000000	0.000000	NaN	23.400000	4.800000	100.000000	0.000000
50%	NaN	43.000000	0.000000	0.000000	NaN	27.320000	5.800000	140.000000	0.000000
75%	NaN	59.000000	0.000000	0.000000	NaN	29.860000	6.200000	159.000000	0.000000
max	NaN	80.000000	1.000000	1.000000	NaN	95.690000	9.000000	300.000000	1.000000

Figure 4.2: Descriptive Statistics

From the descriptive statistics above, it has been possible to make measurements for mean, standard deviation, maximum, and minimum values. Descriptive analysis is a valuable starting point and presents a summarized characteristic that such data are scrutinized for [1]. The dataset calculates the average age of the patients as 41.79 years and their blood glucose level in the range of 80 to 300.

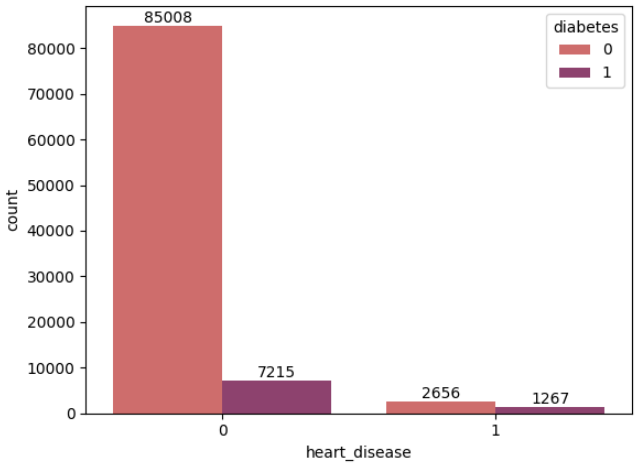

Figure 4.3: Relation of diabetes with heart disease

According to the bar plot in figure 4.3, 1267 people are suffering from diabetes among 3923 heart patients. This concludes that people suffering from heart disease are more likely to diagnosed with diabetes.

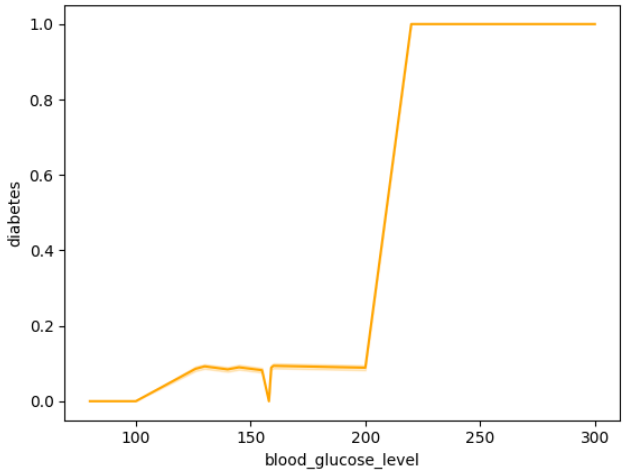

Figure 4.4: Change in the blood glucose level

A sharp increase in diabetes can be observed when the amount of glucose in the blood reaches 200. This implies a direct co-relation between the blood glucose level and diabetes.

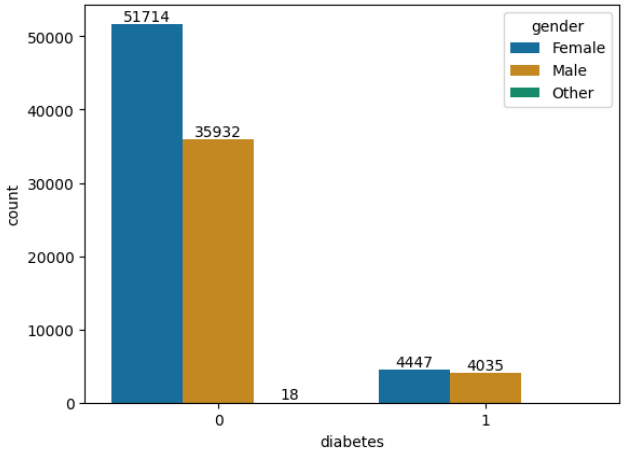

Figure 4.5: Diabetes for different gender

The above image shows that for every 4447 diabetic male patients, there are 4035 females' patients. However, the numbers for non-diabetic male and female's patients are 35932 and 51714 respectively.

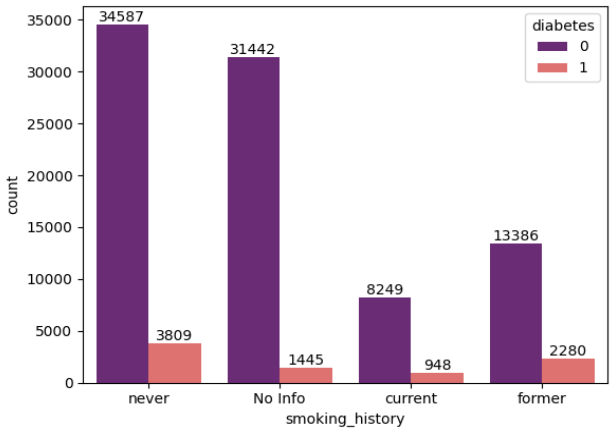

Figure 4.6: Relation between smoking and diabetes

The above plot shows that most of the non-diabetic patients are either non-smokers or their smoking information is not available. In relative terms, the former smokers might have a higher association with diabetes.

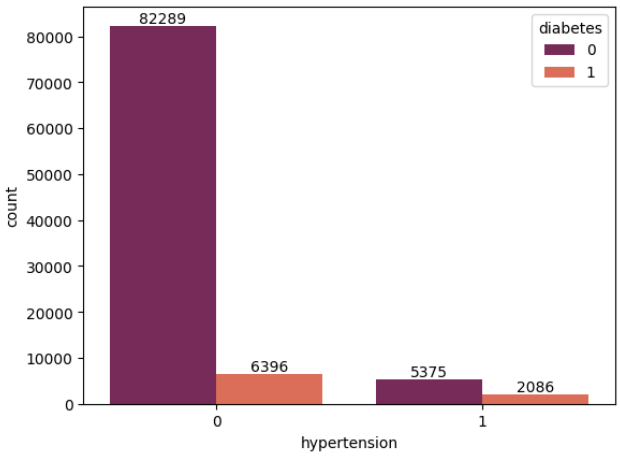

Figure 4.7: Relation between diabetes and hypertension

2088 out of 8482 people who have hypertension also suffer from diabetes. This means that there is a higher chance of having diabetes if the patient already has hypertension.

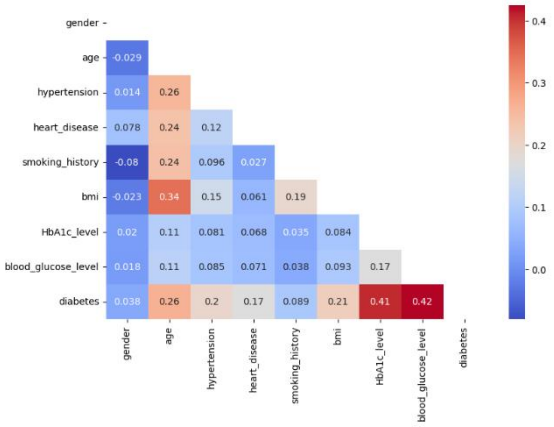

Figure 4.8: Correlation heatmap

The image above represents the correlation matrix among some of the medical parameters. Correlation of 0.42 was the one which happened to be the strongest between Diabetes and HbA1c Level, whereas Diabetes and Blood Glucose Level were just near to it, 0.41. There is a strong positive relationship between age and hypertension, as even included in the Correlation matrix is equal to 0.26. However, age seems to be inversely related as age and heart disease have a very weak negative correlation of −0.029.

```
# Logistic Regression
from sklearn.linear_model import LogisticRegression
logistic = LogisticRegression()
logistic.fit(scaled_x_train, y_train_health)
y_pred_logistic = logistic.predict(scaled_x_test)
```

```
accuracy_logistic = accuracy_score(y_test_health, y_pred_logistic)
print(f'Accuracy of Logistic Regression: {accuracy_logistic}')
```

```
Accuracy of Logistic Regression: 0.9569942797711909
```

```
class_report_logistic = classification_report(y_test_health, y_pred_logistic)
print(f'Classification Report of Logistic Regression:\n{class_report_logistic}')
```

```
Classification Report of Logistic Regression:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98     17509
           1       0.85      0.63      0.72      1721

    accuracy                           0.96     19230
   macro avg       0.91      0.81      0.85     19230
weighted avg       0.95      0.96      0.95     19230
```

Figure 4.9: Logistic Regression results

The logistic regression model performed well, with an overall accuracy of 96% on the test dataset. It yielded a good precision, recall rate, and F1-Score of about 96%, 99%, and 98% for the negative class. The model also performed quite satisfactorily in being discriminative for positive instances as well, that the precision and recall were 0.85 and 0.63, respectively.

```
# KNN Classifier
from sklearn.neighbors import KNeighborsClassifier
```

```
knn = KNeighborsClassifier(n_neighbors=7)
knn.fit(scaled_x_train, y_train_health)
y_pred_knn = knn.predict(scaled_x_test)
```

```
acc_knn = accuracy_score(y_test_health, y_pred_knn)
print(f'Accuracy of KNN Classifier model: {acc_knn}')
```

```
Accuracy of KNN Classifier model: 0.9604784191367655
```

```
class_report_knn = classification_report(y_test_health, y_pred_knn)
print(f'Classification Report of KNN Classifier model:\n{class_report_knn}')
```

```
Classification Report of KNN Classifier model:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98     17509
           1       0.91      0.62      0.74      1721

    accuracy                           0.96     19230
   macro avg       0.94      0.81      0.86     19230
weighted avg       0.96      0.96      0.96     19230
```

Figure 4.10: Results of KNN classifier model

The KNN classifier model returned an accuracy of 96% on the test set. Similar to logistic regression, it worked quite well classifying negative cases with a precision, recall, and F1-score of 0.96, 0.99, and 0.98, respectively. Compared to logistic regression, the KNN model performed slightly better at detecting positive cases with a precision of 0.91 and recall of 0.62.

```
# Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
random_forest = RandomForestClassifier()
random_forest.fit(scaled_x_train, y_train_health)
y_pred_forest = random_forest.predict(scaled_x_test)
accuracy_forest = accuracy_score(y_test_health, y_pred_forest)
class_report_forest = classification_report(y_test_health, y_pred_forest)
print(f'Accuracy of Random Forest Classifier: {accuracy_forest}')
print(f'Classification Report of Random Forest Classifier:\n{class_report_forest}')
```

```
Accuracy of Random Forest Classifier: 0.9682267290691627
Classification Report of Random Forest Classifier:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     17509
           1       0.94      0.69      0.80      1721

    accuracy                           0.97     19230
   macro avg       0.95      0.84      0.89     19230
weighted avg       0.97      0.97      0.97     19230
```

Figure 4.11: Random Forest Classifier model

The test set accuracy with the help of "Random Forest Classifier" model was 96.82%. In the case of negative instances, it had perfect recall, with a precision and F1-score of 1.00 and 0.98, respectively. Even as this model performed well in classifying the negative class, positive class instances were not that perfectly predicted; their precision was 0.94, and their recall was 0.69.

```
# Desision Tree Classifier
from sklearn import tree
dt = tree.DecisionTreeClassifier()
dt.fit(scaled_x_train, y_train_health)
y_pred_dt = dt.predict(scaled_x_test)
accuracy_dt = accuracy_score(y_test_health, y_pred_dt)
class_report_dt = classification_report(y_test_health, y_pred_dt)
print(f'Accuracy Decision Tree Classifier: {accuracy_dt}')
print(f'Classification Report Decision Tree Classifier:\n{class_report_dt}')
```

```
Accuracy Decision Tree Classifier: 0.9471138845553823
Classification Report Decision Tree Classifier:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     17509
           1       0.69      0.74      0.71      1721

    accuracy                           0.95     19230
   macro avg       0.83      0.85      0.84     19230
weighted avg       0.95      0.95      0.95     19230
```

Figure 4.12: Decision Tree Classifier result

The Decision Tree Classifier model yield a good accuracy on the test dataset of 95%. The classifier turned in values for precision, recall, and F1-score equal to 0.97, 0.97, and 0.97 respectively. Its performance in terms of classifying instances for the positive class was less impressive, with its precision being 0.69 and its recall being 0.74.

4.2 Discussion

The entire process took place in terms of exploratory data analysis and predictive model development. The exploratory data analysis shows that men are more likely to be diabetic. Apart from this, diabetes depends on blood glucose and HbA1c level based on the dataset.

Metrics	**Logistic Regression**	**KNN Classifier model**	**Random Forest Classifier**	**Decision Tree Classifier**
Accuracy	0.96	0.96	0.97	0.94
Precision	0.85	0.91	0.94	0.69

Recall	0.63	0.62	0.69	0.74
f-1 score	0.72	0.74	0.80	0.71

Table 1: Comparative analysis of the machine learning models

All ML models revealed an overall high accuracy as well as in the negative class predictions (non-diabetic). Random Forest ended up being good both in overall accuracy, 96.82%, and had a perfect recall for negative cases. It still was superseded by KNN while doing well on the positive class precision. Logistic Regression and Decision Tree have good accuracy but poor recall for positive class instances. According to the metrics given, Random Forest is the best model to be recommended for diabetes prediction because of its overall high accuracy, excellent performance on negative cases, and reasonable precision on positive cases.

5. CONCLUSION

The result of this study has established high use of machine learning and data analytics in the effective management of health especially in risk assessment of diabetes. Using Logistic Regression, k-Nearest Neighbors (KNN), Random Forest Classifier, and Decision Tree Classifier models, it is possible to improve the efficiency of early diagnosis and prevention of diabetes. Each model is computation friendly and has its advantages in the results obtained. It is found that the Random Forest Classifier is the most overfulfilling model with the highest general accuracy and good prediction of the negative and positive classes. This essentially speaks to the fact that the model is able to model intricate data patterns and accurately predict on them. The KNN model also good as expect, especially in terms of the accuracy of the positive cases, but we had similar results with using the Logistic Regression and Decision Tree models, although they provide us with different perspectives, yet suffers from some sort of limitations in the accurate measurement. The finding for analysis and interpretation was supported with the help of data visualization which helped identify correlations between diabetes and other Health related parameters including blood glucose and HbA1c. Based on the result of correlation analysis and performance measures, it was possible to understand the advantages and the disadvantages of each predictive models for selecting the best tool to use for risk prediction of diabetes. Thus, the application of ML and DA is the greatest breakthrough in allowing for active and effective management of health. The above technologies help in early detection of patients who are at high risk and therefore early management is done on them to increase their lifespan. This paper has highlighted the following limitations which should be dealt with in future research; Firstly, the generalizability of the model Secondly, incorporation of data other than website activity data in developing recommendation models. By developing these fields, health care systems can progress the work on the determination of different aspects and accurate prediction necessary for fair and efficient treatment of chronic diseases such as diabetes. In sum, the results stress that the AI-supported technologies can open up a myriad of positive changes in the sphere of patients'

REFERENCES

[1] Gadde, S.S. and Kalli, V.D.R., 2020. Descriptive analysis of machine learning and its application in healthcare. Int J Comp Sci Trends Technol, 8(2), pp.189-196.

[2] Jayatilake, S.M.D.A.C. and Ganegoda, G.U., 2021. Involvement of machine learning tools in healthcare decision making. Journal of healthcare engineering, 2021(1), p.6679512.

[3] ABHISHEK, C., CHOUBEY, S.B., PRAFULL, K., DAULATABAD, V.S. and NITIN, J., 2024. Healthcare Transformation: Artificial Intelligence Is the Dire Imperative of the Day. Cureus, 16(6),.

[4] ARJMANDNIA, F. and ALIMOHAMMADI, E., 2024. The value of machine learning technology and artificial intelligence to enhance patient safety in spine surgery: a review. Patient Safety in Surgery, 18, pp. 1-6.

[5] BHAGAT, S.V. and DEEPIKA, K., 2024. Navigating the Future: The Transformative Impact of Artificial Intelligence on Hospital Management- A Comprehensive Review. Cureus, 16(2),.

[6] CHANDRA, P., DUBEY, A., SHARMA, S.K. and KARSOLIYA, S., 2024. A novel Conceptualization of AI Literacy and Empowering Employee Experience at Digital Workplace Using Generative AI and Augmented Analytics: A Survey. Journal of Electrical Systems, 20(2), pp. 2582-2603.

[7] GIUFFRÈ, M. and SHUNG, D.L., 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digital Medicine, 6(1), pp. 186.

[8] JEE, Y.K., HASAN, A., KELLOGG, K.C., RATLIFF, W., MURRAY, S.G., SURESH, H., VALLADARES, A., SHAW, K., TOBEY, D., VIDAL, D.E., LIFSON, M.A., PATEL, M., INIOLUWA, D.R., GAO, M., KNECHTLE, W., TANG, L., BALU, S. and SENDAK, M.P., 2024. Development and preliminary testing of Health Equity Across the AI Lifecycle (HEAAL): A framework for healthcare delivery organizations to mitigate

the risk of AI solutions worsening health inequities. PLOS Digital Health, 3(5),.

[9] KALOGIANNIDIS, S., KALFAS, D., PAPAEVANGELOU, O., GIANNARAKIS, G. and CHATZITHEODORIDIS, F., 2024. The Role of Artificial Intelligence Technology in Predictive Risk Assessment for Business Continuity: A Case Study of Greece. Risks, 12(2), pp. 19.

[10] KARRAS, A., GIANNAROS, A., KARRAS, C., THEODORAKOPOULOS, L., MAMMASSIS, C.S., KRIMPAS, G.A. and SIOUTAS, S., 2024. TinyML Algorithms for Big Data Management in Large-Scale IoT Systems. Future Internet, 16(2), pp. 42.

[11] LUKKIEN, D.R.M., STOLWIJK, N.E., ASKARI, S.I., HOFSTEDE, B.M., HENK, H.N., BOON, W.P.C., PEINE, A., MOORS, E.H.M. and MINKMAN, M.M.N., 2024. AI-Assisted Decision-Making in Long-Term Care: Qualitative Study on Prerequisites for Responsible Innovation. JMIR Nursing, 7.

[12] LUO, Y., MAO, C., SANCHEZ-PINTO, L., AHMAD, F.S., NAIDECH, A., RASMUSSEN, L., PACHECO, J.A., SCHNEIDER, D., MITHAL, L.B., DRESDEN, S., HOLMES, K., CARSON, M., SHAH, S.J., KHAN, S., CLARE, S., WUNDERINK, R.G., LIU, H., WALUNAS, T., COOPER, L., FENG, Y., WEHBE, F., FANG, D., LIEBOVITZ, D.M., MARKL, M., MICHELSON, K.N., MCCOLLEY, S.A., GREEN, M., STARREN, J., ACKERMANN, R.T., D'AQUILA, R.,T., ADAMS, J., LLOYD-JONES, D., CHISHOLM, R.L. and KHO, A., 2024. Northwestern University resource and education development initiatives to advance collaborative artificial intelligence across the learning health system. Learning Health Systems, 8(3),.

[13] MENG-LEONG HOW and SIN-MEI CHEAH, 2024. Forging the Future: Strategic Approaches to Quantum AI Integration for Industry Transformation. Ai, 5(1), pp. 290.

[14] NAIR, M., SVEDBERG, P., LARSSON, I. and NYGREN, J.M., 2024. A comprehensive overview of barriers and strategies for AI implementation in healthcare: Mixed-method design. PLoS One, 19(8),.

[15] ONYEJEKWE, E.R., PHD., SHERIFI, DASANTILA, PHD,M.B.A., R.H.I.A. and CHING, HUNG,PHD., D.A.B.R., 2024. Perspectives on Big Data and Big Data Analytics in Healthcare. Perspectives in Health Information Management, 21(1), pp. 1-19.

[16] PDF, 2024. Framework for Organization of Medical Processes in Medical Institutions Based on Big Data Technologies. International Journal of Advanced Computer Science and Applications, 15(3),.

[17] SHIVA, M.V. and FOROUZANFAR, M., 2024. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. Bioengineering, 11(4), pp. 337.

[18] WILLIAMSON, S.M. and PRYBUTOK, V., 2024. Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. Applied Sciences, 14(2), pp. 675.

[19] YAZDI, M., ZAREI, E., ADUMENE, S. and BEHESHTI, A., 2024. Navigating the Power of Artificial Intelligence in Risk Management: A Comparative Analysis. Safety, 10(2), pp. 42.

[20] ZAMANI, E.D., SMYTH, C., GUPTA, S. and DENNEHY, D., 2023. Artificial intelligence and big data analytics for supply chain resilience: a systematic literature review. Annals of Operations Research, 327(2), pp. 605-632.

[21] ZAVALETA-MONESTEL ESTEBAN, QUESADA-VILLASEÑOR, R., ARGUEDAS-CHACÓN SEBASTIÁN, GARCÍA-MONTERO, J., BARRANTES-LÓPEZ MONSERRAT, SALAS-SEGURA, J., ANCHÍA-ALFARO ADRIANA, NIETO-BERNAL, D. and DIAZ-JUAN, D., 2024. Revolutionizing Healthcare: Qure.AI's Innovations in Medical Diagnosis and Treatment. Cureus, 16(6),.

[22] ZHU, Y., SALOWE, R., CHOW, C., LI, S., BASTANI, O. and JOAN M O'BRIEN, 2024. Advancing Glaucoma Care: Integrating Artificial Intelligence in Diagnosis, Management, and Progression Detection. Bioengineering, 11(2), pp. 122."