How to Cite:
Devi, K., Kumar, J. S., Poovizhi, S., & Krishnan, M. (2022). Analytical study on diabetes prediction-using random forest classifier. *International Journal of Health Sciences*, *6*(S4), 6404–6413. https://doi.org/10.53730/ijhs.v6nS4.9617

Analytical study on diabetes prediction-using random forest classifier

Dr. Kabirdoss Devi
Assistant Professor, Department of Commerce, College of Science and Humanities, SRM Institute of Science and Technology, Ramapuram Campus, Chennai

Mr. J. Sathish Kumar
Assistant Professor, Department of Commerce, College of Science and Humanities, SRM Institute of Science and Technology, Ramapuram Campus, Chennai

Dr. S. Poovizhi
Assistant Professor, Department of Management Studies, St. Joseph's College of Engineering, Chennai

Mr. M. Krishnan
Associate Business Intelligence, 4i Apps Solutions Private Ltd.

***Abstract*---**Diabetes predictions have gained major attention due to its consequences on the healthy well-being of an individual. When glucose levels go high due to non- availability of the hormone called insulin which digest glucose, together with other side effects like frequent urination, excessive thirst, and hunger with sudden weight reduction, one can be confirmed of suffering from diabetes. This requires a consistent treatment and monitoring of its complications which are considered fatal in some cases. There are various ways to keep a tract of the glucose level in blood to adjust the diet and dosage of insulin. However, predicting it as early as possible is a challenging task due to its inter-dependency factor that causes trouble to human organs like viscera, peripherals, nervous system, cardiovascular, eyes and excretory system. This research paper aims to provide five different machine learning methods for the prediction of diabetes such as SVM, Logistics regression, KNN Classifier, Random Forest and Logistic algorithm. These proposed methods are effective techniques for earlier detection of the diabetes.

***Keywords*---**SVM, KNN, classifier, diabetes, random forest, algorithm.

Introduction

Diabetes is one the most troubling disorder which mandates regular check-ups for the awareness of insulin level and to keep up the intake. It sometimes considered to be fatal when ignored as it damages the important vital organs of cardiovascular and respiratory organs. Diabetes affects all the important organs including eye causing cataract and severe cases blindness. It also affects the peripheral nervous system, skin and feet. Gangrene is the another major issue resulting in tissue necrosis and mandatory surgical intervention. There are two ways diabetes in patients can be calculated. Firstly, data pre- processing where all the attributes are identified and secondly, the predictive models which are constructed based on the decision tree methods. There are many available tools for the purpose like WEKA mining tools, Decision tool algorithm and Artificial Neural Network using kappa statistics, mean absolute error and relative squared error.

Objectives of the study

The objective of the study is:

1. To predict the presence of diabetes in an individual based on insulin level, age and BMI.
2. To analyse the preset data obtained from Kaggle for the presence of diabetes.
3. To test all the five algorithm for the best fit algorithm identification.

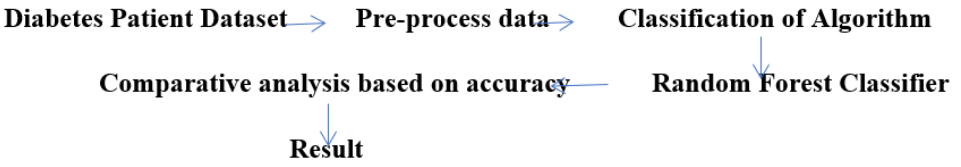
Review of Literature

Diabetes is made as a chronic disease as there is no permanent cure leading to many consequences and demanding life long conservative management for keeping the sugar level at low. (Roshi Saxena et.al 2022) conducted a analysis on models for diabetic predictions and reviewed many researches to find the best one. They brought out that the analysis of these diabetic datasets were quite challenging because these were non linear and complex in nature and they proved that these shortages could be overcome by the machine learning based stratification of risk system for the classification of patients into diabetic and control. (Quan Zou et.al 2018) used decision tree, random forest and neural network to predict DM. Their devised models were used five-fold cross validation for the purpose of examination. They conducted independent experiments to check the methods that had the probability for the better performance. They did a random selection of 64994 healthy people and diabetic datasets for the training set. The outcome projected high accuracy with the random forest when all the other attributes used. (More, Rana 2017) they summarized all the literature survey between the years 2000 and 2016 with respect to the RFC and resolving class imbalance. The research paper (Gerald Biau 2012) on random forest discussed about the set of decision trees that was growing between the spaces of the datasets. He also discussed about the statistical and mathematical properties that was driving the algorithm of the random forest. This paper they discussed the in-depth analysis of this model proposed by Breiman in 2004. This was very close to the original algorithm. Not only diabetic predication and some of the health care contributions but also environmental issues like ground water yield

was predicted with the help of self-learning Random Forest (Maher Ibrahim, Biswajith, Saro Lee 2018). They conducted this study in South Africa having inventory data segmented into two groups where the first group was made 70% of training and 30% of testing. They used Bayesian Optimization Method and compared their model with other ML models such as Support Vector Machine, Artificial Networks and Decision trees.

Research Methodology

This methodology uses Random Forest which is Machine Learning and part of Artificial Intelligence enabling computer ability to learn. Reinforcement leaning is one of the category of machine learning which learns through past experiences. However, Random forest uses supervised learning algorithm that uses both regression and classification. Random forest has many decision tress similar to real forest that has many trees. These various subsets improves the accuracy of the outcome. This research paper aims to predict presence of diabetes through five different supervised machine learning methods like SVM, Logistics Regression, KNN Classifier, Random Forest and Logistic algorithm.
The architectural design of the research methodology includes -



Data Design

Datasets for this study has been taken from the secondary source where the details of the past records of diabetes predictions stored. These datasets were obtained from the third party source known as Kaggle and it consisted of seven features and two thousand plus rows. The variables such as glucose, insulin, BMI, BP included in the model. The information related to the study alone were fitted to the database model. This database model was designed such a way that it must avoid redundant to prevent wasting of space which would further increase the faults and discrepancies within the database. There are many features being used in the machine learning model which used to predict the output accurately. These datasets has a major drawback; where certain features in the datasets might be high value and some might be low value to make it more feasible.

Data Collection

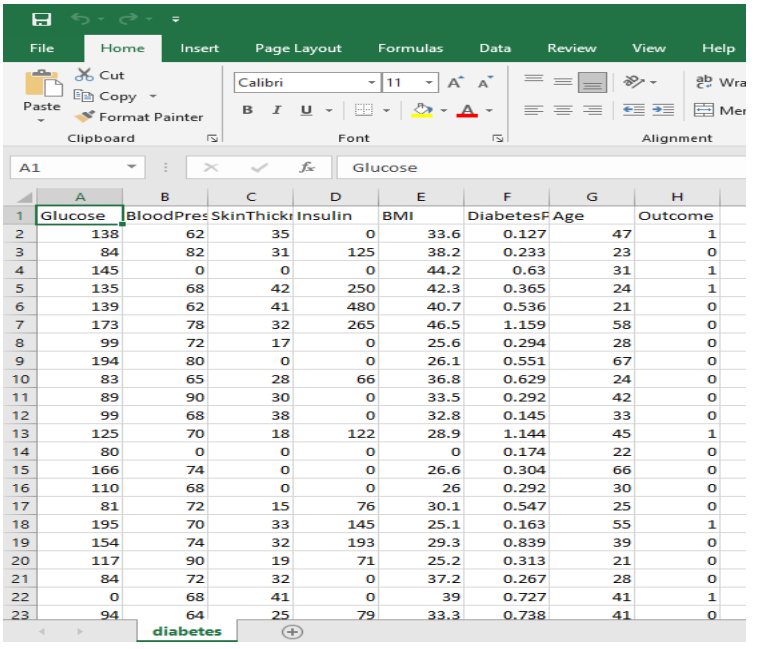
The multiple diabetic datasets were obtained from Kaggle and Merged together. The variables for the study includes glucose level, BMI, prediction function and insulin etc.

Data pre – processing

There are four steps involved in the date pre- processing mainly to avoid errors, inconsistency and incomplete status of the data. Errors do occur when the raw data being transferred to an understandable format for the analysis. To qualify for the data analysis, the datasets should be split into training and validation sets, should be rectified for all the missing values, should be divided based on the categorical features and the datasets should be normalized.

Tools used for analysis

The whole computation process was made on the web- based application for the ease of capturing entire calculation. These calculation process includes developing, documenting and coding and communicating outcomes. Another component used was browser based where the documents could be combined for explanatory texts with rich media output computations such as HTML, LaTeX, PNG, SVG etc.



	A	B	C	D	E	F	G	H
1	Glucose	BloodPres	SkinThick	Insulin	BMI	DiabetesF	Age	Outcome
2	138	62	35	0	33.6	0.127	47	1
3	84	82	31	125	38.2	0.233	23	0
4	145	0	0	0	44.2	0.63	31	1
5	135	68	42	250	42.3	0.365	24	1
6	139	62	41	480	40.7	0.536	21	0
7	173	78	32	265	46.5	1.159	58	0
8	99	72	17	0	25.6	0.294	28	0
9	194	80	0	0	26.1	0.551	67	0
10	83	65	28	66	36.8	0.629	24	0
11	89	90	30	0	33.5	0.292	42	0
12	99	68	38	0	32.8	0.145	33	0
13	125	70	18	122	28.9	1.144	45	1
14	80	0	0	0	0	0.174	22	0
15	166	74	0	0	26.6	0.304	66	0
16	110	68	0	0	26	0.292	30	0
17	81	72	15	76	30.1	0.547	25	0
18	195	70	33	145	25.1	0.163	55	1
19	154	74	32	193	29.3	0.839	39	0
20	117	90	19	71	25.2	0.313	21	0
21	84	72	32	0	37.2	0.267	28	0
22	0	68	41	0	39	0.727	41	1
23	94	64	25	79	33.3	0.738	41	0

Data Structure

Data Analysis

Below is the datasets that has been included for the study highlighting all the variables. The values are arranged in an array for the easy calculation and for the inclusion in the study.

6408

Data Set

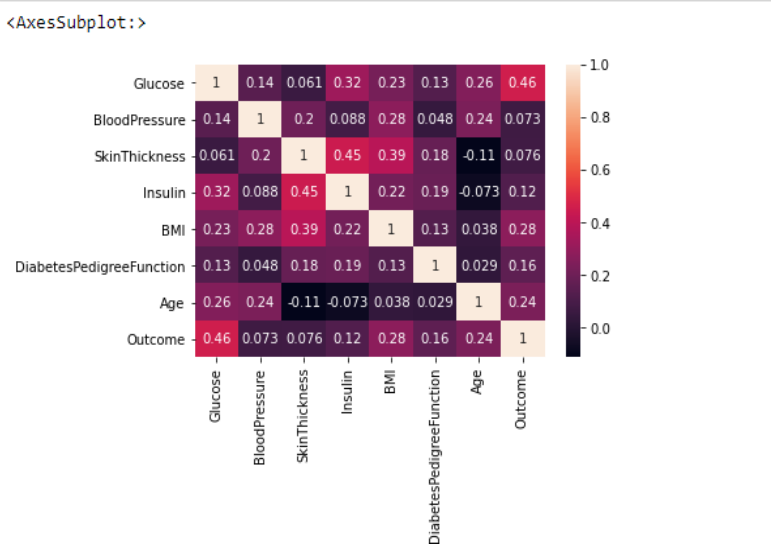
Out[2]:

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	138	62	35	0	33.6	0.127	47	1
1	84	82	31	125	38.2	0.233	23	0
2	145	0	0	0	44.2	0.630	31	1
3	135	68	42	250	42.3	0.365	24	1
4	139	62	41	480	40.7	0.536	21	0
...	...	...	...	...	...	...	...	...
2763	101	76	48	180	32.9	0.171	63	0
2764	122	70	27	0	36.8	0.340	27	0
2765	121	72	23	112	26.2	0.245	30	0
2766	126	60	0	0	30.1	0.349	47	1
2767	93	70	31	0	30.4	0.315	23	0

2768 rows × 8 columns

Detecting correlation between input features

High and low correlation between the variables are plotted in the correlation matrix. This brings out the fact that the changes in direction of the one variable with respect the another variable in specific manner. Revealing the relationship between the variable is useful because value of one variable helps to predict the value of another variable.

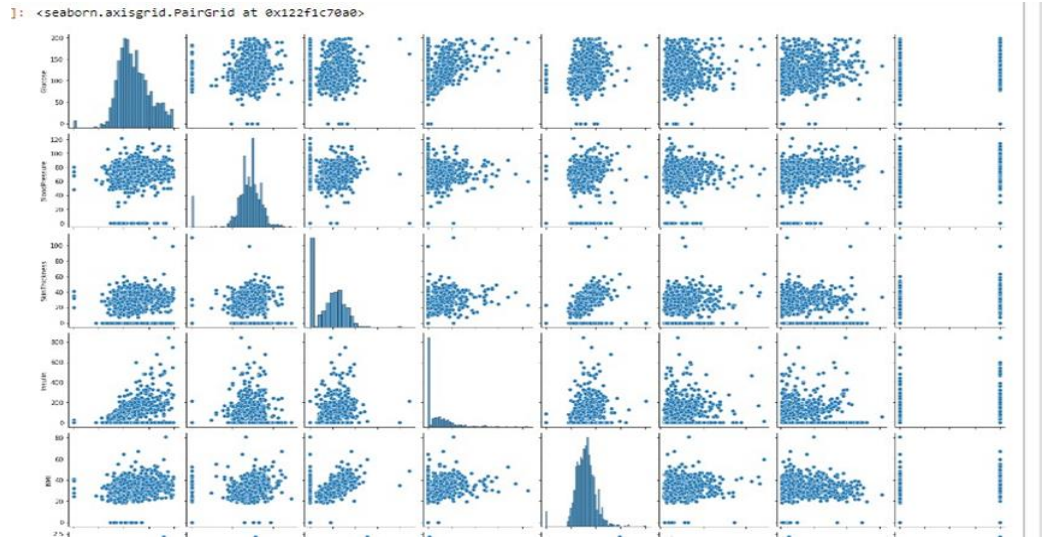


Pair plot of the attributes

Pairwise relationship is being made through the help of attributes of pair plot methodology. Pair plot is made with variables across the columns and rows, pairplot() function is used to plot multiple pairwise bivariate distributions. This

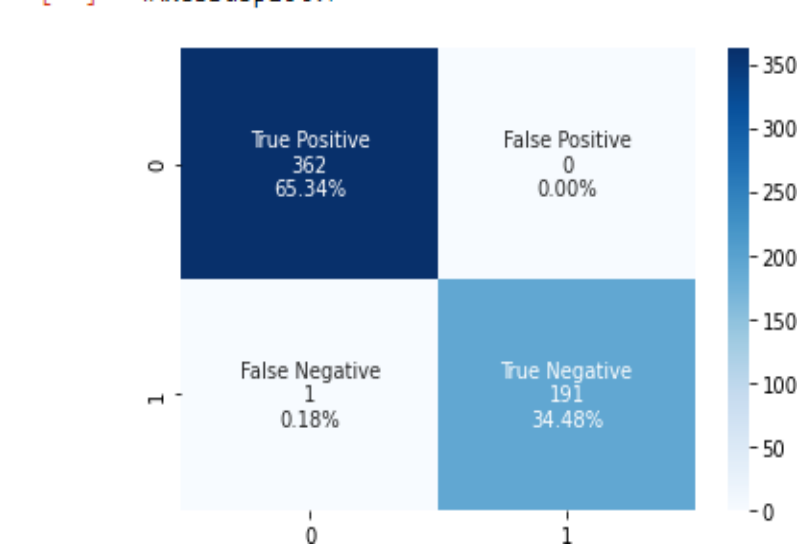
pair plot shows the relationship for (n,2) in a DataFrame as matrix with the combination of variables as well as diagonal plot of univariate. This pair plot is very useful in identifying the relationship between both single and two variables. Follow-up analysis are better made with the pair plots and which can also easily be implemented with Python.

Pair plot



Confusion Matrix

The first step after data has been cleaned, pre- processed and wrangled is to feed it to a model and get the probability output. For the better effectiveness and performance the datasets requires confusion matrix.



Confusion Matrix

Performance measurements for machine learning comes with the four different values of predicted and actual values.
True Positive: Interpretation: You predicted positive and it's true.
True Negative: Interpretation: You predicted negative and it's true.
False Positive: (Type 1 Error) Interpretation: You predicted positive and it's false.
False Negative: (Type 2 Error) Interpretation: You predicted negative and it's false.

Results and Discussion

Chronic diseases like diabetes requires early detection to prevent irrecoverable damages to the vital organs. This random forest technique not only helps researchers to create veracious and effective tool but also clinician to make the best decision about the disease condition. Availability of surplus datasets of epidemiological and genetic risk and the advanced computational methods makes the machine learning with an enhanced way of predictions. There are different data mining techniques and its applied algorithm on different medical data sets and these have strong power of machine learning on datasets.

Accuracy Score

ALGORITHM	ACCURACY SCORE
Logistic Regression	77.29%
KNN Classifier	78.73%
Support Vector Machine	75.28%
Naïve Bayes	75.84%
Random Forest Classifier	96.87%

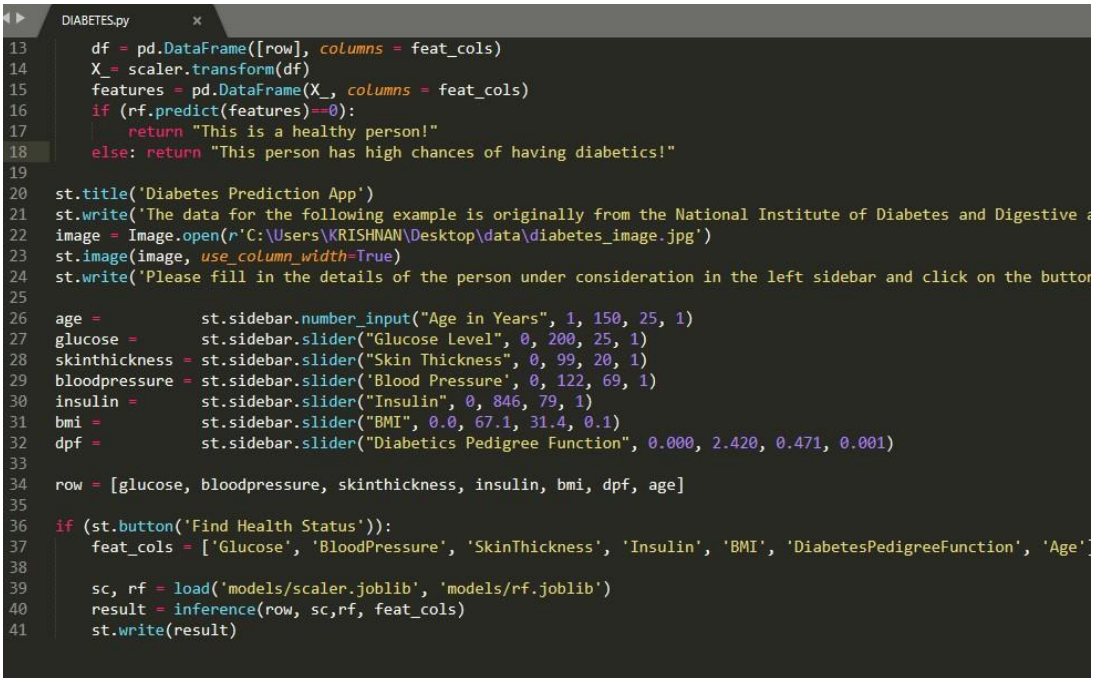
Accuracy score works well when there are equal number of samples in each class. If there are two classes with 98% in one class and 2% in another class, then the model considers the 98% training accuracy and omits the other class. Similarly, when the same model is being utilized for 60% and 40%, then it would drop to 60%. Hence, classification accuracy gives the false sense while achieving high accuracy.

F_Score

The highest possible value of an F-core is 1.0. This indicates the perfect precision and the lowest value is 0 and if either then the precision or the recall is zero.

ALGORITHM	F - SCORE
Logistic Regression	61.5%
KNN Classifier	67.48%
Support Vector Machine	56.5%
Naïve Bayes	61.26%
Random Forest Classifier	99.74%

Graphical User Interface



```
DIABETES.py
    df = pd.DataFrame([row], columns = feat_cols)
    X_= scaler.transform(df)
    features = pd.DataFrame(X_, columns = feat_cols)
    if (rf.predict(features)==0):
        return "This is a healthy person!"
    else: return "This person has high chances of having diabetics!"

st.title('Diabetes Prediction App')
st.write('The data for the following example is originally from the National Institute of Diabetes and Digestive a
image = Image.open(r'C:\Users\KRISHNAN\Desktop\data\diabetes_image.jpg')
st.image(image, use_column_width=True)
st.write('Please fill in the details of the person under consideration in the left sidebar and click on the buttor

age =           st.sidebar.number_input("Age in Years", 1, 150, 25, 1)
glucose =       st.sidebar.slider("Glucose Level", 0, 200, 25, 1)
skinthickness = st.sidebar.slider("Skin Thickness", 0, 99, 20, 1)
bloodpressure = st.sidebar.slider('Blood Pressure', 0, 122, 69, 1)
insulin =       st.sidebar.slider("Insulin", 0, 846, 79, 1)
bmi =           st.sidebar.slider("BMI", 0.0, 67.1, 31.4, 0.1)
dpf =           st.sidebar.slider("Diabetics Pedigree Function", 0.000, 2.420, 0.471, 0.001)

row = [glucose, bloodpressure, skinthickness, insulin, bmi, dpf, age]

if (st.button('Find Health Status')):
    feat_cols = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age']

    sc, rf = load('models/scaler.joblib', 'models/rf.joblib')
    result = inference(row, sc,rf, feat_cols)
    st.write(result)
```

Python Structure

Many GUIs comes with touchscreen and voice-command and allow the users to interact with different electronic devises. GUI are found to be great when compared to the command line interface. GUI which was commercially deployed into Apple Macintosh and Windows Android designed to respond to the problem of inefficient usability. The text - based command- line interfaces for the average user.

There are different kinds of algorithms of machine learning being used to discover patterns in machine language like python which directs to the actionable visual precepts. Based on the learning pattern this algorithm can be classified into two groups such as supervised and unsupervised learning. Predicting the class for the given data point is known as classification. This predictive model does a mapping function on the X input variables to Y discrete output variables. Under the supervised learning, it is possible to predict because computer learns the data input given. This data has the ability to identify bi-class like male or female or multi-class where it includes many variables.

DIABETES.py

```python
import streamlit as st
import joblib
import pandas as pd
from PIL import Image

@st.cache(allow_output_mutation=True)
def load(scaler_path, rf_path):
    sc = joblib.load(r"C:\Users\KRISHNAN\Desktop\models\scaler.joblib")
    rf = joblib.load(r"C:\Users\KRISHNAN\Desktop\models\rf.joblib")
    return sc , rf

def inference(row, scaler,rf, feat_cols):
    df = pd.DataFrame([row], columns = feat_cols)
    X_= scaler.transform(df)
    features = pd.DataFrame(X_, columns = feat_cols)
    if (rf.predict(features)==0):
        return "This is a healthy person!"
    else: return "This person has high chances of having diabetics!"

st.title('Diabetes Prediction App')
st.write('The data for the following example is originally from the National Institute of Diabetes and Digestive and Kidney Diseas
image = Image.open(r'C:\Users\KRISHNAN\Desktop\data\diabetes_image.jpg')
st.image(image, use_column_width=True)
st.write('Please fill in the details of the person under consideration in the left sidebar and click on the button below!')

age =           st.sidebar.number_input("Age in Years", 1, 150, 25, 1)
glucose =       st.sidebar.slider("Glucose Level", 0, 200, 25, 1)
skinthickness = st.sidebar.slider("Skin Thickness", 0, 99, 20, 1)
bloodpressure = st.sidebar.slider('Blood Pressure', 0, 122, 69, 1)
insulin =       st.sidebar.slider("Insulin", 0, 846, 79, 1)
bmi =           st.sidebar.slider("BMI", 0.0, 67.1, 31.4, 0.1)
dpf =           st.sidebar.slider("Diabetics Pedigree Function", 0.000, 2.420, 0.471, 0.001)

row = [glucose, bloodpressure, skinthickness, insulin, bmi, dpf, age]
```

Feature Interpretation

Stream Lit GUI
Stream lit is an open source python library. It makes it easy to create and share beautiful web apps for data analysis. This particular app does not require any knowledge on web development. It automatically saves and updates all the codes and the codes runs from top to bottom and does not require a callback.

Suggestions

Clinicians with the help of analysts should explore the possibilities of predicting the glucose level in the blood using these models. HTML and FLASK can also be used to predict well compare to steam lit web application. HTML would provide better web application compare to GUI. HTML and FLASK can also be easily uploaded into cloud storage compare to SLA.

Conclusion

Patients stage and the grade has to be found and the parameters like accuracy, classification report, confusion matric by SML algorithm method for the analytical process after the data cleaning, pre-processing and wrangling done. Finding out the level of glucose through machine learning and predicting at a minimum effort is the need of an hour. When the advanced technology grows exponentially, the problem should be addressed and explicitly used. LDA and SVM balances with 92% accuracy. RNN balances with 94%.

References

1. A. S. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017, pp. 72-78, doi: 10.1109/ICISIM.2017.8122151.
2. S. Devika,V.Gopikamani,S.Mahima,Rejini, GUI based prediction of diabetic stages using machine learning approach, International Journal of Advanced Research and Innovative Ideas in Education, Vol 6, Issue 2, 2020 pp- 787-796.
3. Gerald Biau, Analysis of a Random Forests Model, The Journal of Machine Learning Research, Vol 13, 2012, pp- 1063-1095 In text; L. Breiman. Random forests. *Machine Learning*, 45:5-32, 2001.
4. Roshi Saxena, Sanjay Kumar Sharma, Manali Gupta, G. C. Sampada, "A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey", Journal of Healthcare Engineering, vol. 2022, Article ID 8100697, 15 pages, 2022. https://doi.org/10.1155/2022/8100697
5. Sameen, M.I., Pradhan, B. & Lee, S. Self-Learning Random Forests Model for Mapping Groundwater Yield in Data-Scarce Areas. *Nat Resour Res* **28,** 757–775 (2019). https://doi.org/10.1007/s11053-018-9416-1
6. Zou Quan, Qu Kaiyang, Luo Yamei, Yin Dehui, Ju Ying, Tang Hua, Predicting Diabetes Mellitus With Machine Learning Techniques, Frontiers in Genetics, Vol 9 2018. DOI=10.3389/fgene.2018.00515
7. https://link.springer.com/article/10.1007/s11053-018-9416-1
8. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.735.5824&rep=rep1&type=pdf
9. https://projecteuclid.org/journals/annals-of-statistics/volume-43/issue-4/Consistency-of-random-forests/10.1214/15-AOS1321.full
10. http://proceedings.mlr.press/v32/denil14.html
11. Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). Health and treatment of diabetes mellitus. *International Journal of Health Sciences*, *5*(1), i-v. https://doi.org/10.53730/ijhs.v5n1.2864
12. Saraswati, P. A. I. ., Gunawan, I. M. K., & Budiyasa, D. G. A. (2021). Overview of glomerulus filtration in type 2 of diabetes mellitus at Sanjiwani Gianyar hospital year of 2018-2019. *International Journal of Health & Medical Sciences*, *4*(1), 50-55. https://doi.org/10.31295/ijhms.v4n1.726