Diabetic Prediction using Random Forest Model & Evaluation metrics using IBM SPSS

1st Surya Sai Raj Lakkoju
Dept. of Computer Science and Engg
Majors in Artificial Intelligence
Denton, Tx, USA
Suryasairajlakkoju@my.unt.edu
11610081

2nd Vaishnavi Nomula
Dept. of Computer Science and Engg
Majors in Computer Engineering
Denton, Tx, USA
vaishnavinomula@my.unt.edu
11618103

3rd Bhavya Golla
Dept. of Computer Science and Engg
Majors in Computer Science
Denton, Tx, USA
bhavyagolla@my.unt.edu
11708064

4th Amisha Patra
Dept. of Computer Science and Engg
Majors in Computer Science
Denton, Tx, USA
amishapatra@my.unt.edu
11715439

5th Deepanvi Balusuguri
Dept. of Computer Science and Engg
Majors in Computer Science
Denton, Tx, USA
amishapatra@my.unt.edu
11708391

Diabetic Prediction GitHub Link

***Abstract*—Calculating Evaluation Metrics on Diabetic Prediction on Patient BMI Implementing T-Test analysis using IBM SPSS Using Random Forest ML Model**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Goal and Objective

A. Motivation

The motivation behind this project is to contribute to the advancement of predictive healthcare analytics and ultimately improve the well-being of individuals at risk of diabetes. Diabetes is a global health challenge that affects millions of individuals, and its prevalence continues to rise. One of the key factors in diabetes management is understanding and predicting an individual's risk of developing the condition. Among various predictive factors, Body Mass Index (BMI) is a crucial indicator.Here we create a comprehensive solution for diabetes risk prediction. The ability to predict diabetes based on an individual's BMI allows for early interventions, such as lifestyle modifications and medical treatments. By analyzing the BMI the primary goal is to identify high-risk individuals for diabetes, which will minimize false negatives

B. Significance

The significance of the project lies in developing predictive analytics within medical care. Collecting the data and analyzing the patient's health reports in advance helps to identify and mitigate the consequences of the disease by providing personalized treatment to them. In our case, we incorporated the Body Mass Index(BMI) into our model. Based on the results, we can first target high-risk patients. So, individuals can change their lifestyle, which in turn helps to reduce the risk of diabetes. This model allows the health care professionals to identify where the risk starts and can prevent it before getting too high. It also aids the researchers in studying the condition more in-depth to determine the complexity of the disease. This investigation aids the experimenters in developing more preventive measures and can lower the condition with more advanced approaches

C. Objective

The main objective of the proposed method is to develop a predictive model for forecasting diabetes risk in patients, prioritizing the BMI(Body Mass Index) and considering several factors such as age, gender, hypertension, heart disease, smoking history, blood glucose level, etc. To handle complex data and maintain the accuracy of the results, we are using the Random Forest model and various evaluation metrics such as F1 score and recall techniques. Besides, statistical validation achieved through T-test analysis helps to specify the impact and significance of BMI in predicting diabetes.

D. Features

The project incorporates several standard features to achieve its objectives. These include the utilization of familiar elements for successful completion.

- **Machine Learning Model**
  - **Algorithm Selection:** The project leverages the Random Forest algorithm, recognized for its robustness and versatility in predicting diabetes based on patient BMI data. Random Forest is an ensemble learning method that constructs multiple

decision trees during training, providing reliable classifications and mitigating issues of overfitting.

  - **Training Data:**
    The model is trained on a dataset containing patient BMI information and corresponding diabetes status.

- **Evaluation Metrics**

  - **Model Performance Evaluation:**
    Post-training, the Random Forest model's predictive power is assessed using various evaluation metrics, including the Confusion matrix, accuracy, precision, recall, and F1-score. These metrics offer insights into the model's efficacy in correctly classifying diabetic and non-diabetic patients..

  - **Statistical Significance Analysis:**
    To ascertain if a statistically significant difference exists in BMI between diabetic and non-diabetic patients, a T-test analysis is conducted. This statistical test compares the mean BMI values of the two groups, providing a p-value indicating the significance of the observed differences.

- **IBM SPSS**

  - **Tool Utilization:**
    The project employs IBM SPSS, a comprehensive statistical software package renowned for its user-friendly interface and robust capabilities in data analysis, machine learning, and statistical testing.

  - **T-Test Analysis in SPSS:**
    IBM SPSS facilitates T-Test analysis, allowing for the comparison of means between predicted outcomes of the model and actual outcomes, aiding in the determination of statistical significance.

III. Related Works

The related work for this project involves an examination of existing research, studies, and methodologies relevant to predicting diabetes based on patient BMI using a Random Forest machine learning model and subsequently applying T-Test analysis with IBM SPSS. Numerous studies have explored the application of machine learning techniques, including Random Forest, for diabetes prediction using various patient-related data, including BMI. These studies consistently demonstrate the effectiveness of machine learning models in identifying diabetic patients with high accuracy.

In the realm of machine learning approaches for diabetes prediction, Random Forest emerges as a robust and versatile algorithm, widely used due to its ability to handle complex relationships between variables and resistance to overfitting. Studies employing Random Forest for diabetes prediction based on patient BMI data consistently achieve high accuracy and AUC values. Other machine learning algorithms, such as Support Vector Machines (SVM) and Neural Networks, also show promising results in diabetes prediction, leveraging BMI as a crucial input variable.

The integration of T-Test analysis is a notable aspect of related work, widely used to compare the mean BMI values between diabetic and non-diabetic patients. This statistical test helps determine the statistical significance of the difference in BMI between the two groups. Studies consistently show that diabetic patients tend to have significantly higher mean BMI values compared to non-diabetic individuals, supporting the idea that BMI is a risk factor for diabetes.

Moreover, several studies have explored the combination of machine learning techniques with statistical analysis to enhance diabetes prediction and understand the role of BMI in diabetes risk assessment. This integration provides a more comprehensive understanding of the relationship between BMI and diabetes risk.

The effectiveness of Random Forest is highlighted in specific studies, such as one conducted by Sharma et al. (2013) and another by Al-Khateeb et al. (2016). These studies successfully employed Random Forest to predict diabetes based on datasets containing patient information, including BMI, achieving high accuracy and demonstrating the algorithm's potential for diabetes prediction. Findings consistently support the notion that machine learning techniques, particularly Random Forest, can effectively predict diabetes based on patient BMI data.

Furthermore, research consistently shows a significant association between BMI and diabetes risk. T-Test analysis conducted by Lee and Sung (2009) found that diabetic patients had significantly higher mean BMI values compared to non-diabetic individuals. This aligns with the general consensus that higher BMI is associated with an increased risk of developing diabetes.

In the application development process after completing the development work, we started doing unit testing for the model and the application UI. There we are having a few issues which we are planning to resolve those changes in the next submission the datacleaning process and validations for the UI will be implemented in the next phase. Also these applications doesn't need any production environment. So, we don't have any deployment phase for this project.

IV. Detailed Design of Features

The meticulous design of features for the project involves the explicit delineation of crucial elements and attributes considered in predicting diabetes based on patient BMI, employing a Random Forest ML model, and subsequently conducting T-Test analysis using IBM SPSS. Here is an elaboration on each of the specified methods:

A. Feature Selection:

- **BMI (Body Mass Index):**
  The primary focal point is BMI, a numerical representation derived from an individual's weight and height,

widely acknowledged as an indicator of body fat and often associated with the risk of diabetes.

B. *Additional Patient Information:*

- **Demographic Information:**
  Inclusion of demographic data such as age, gender, and ethnicity is contemplated, recognizing their potential influence on diabetes risk.
- **Clinical Data:**
  Integration of pertinent clinical data, including blood pressure, cholesterol levels, and other health indicators, is considered to contribute to a more nuanced prediction model.

C. *Random Forest Model Features:*

- **Decision Trees:**
  Acknowledging that Random Forest constitutes an ensemble of decision trees, each tree considering a subset of features, it is essential to ensure that decision trees can effectively utilize relevant patient information for predictions.
- **Tree Depth and Split Criteria:**
  Specification of the maximum depth of decision trees and criteria for node splitting is undertaken, considering the intricate relationships between features.

D. *Data Pre-processing Steps:*

- **Missing Value Handling:**
  The implementation of a strategy for handling missing values is outlined, involving methods like imputation or exclusion of incomplete records.
- **Normalization or Standardization:**
  Depending on the algorithm's sensitivity to varied scales, the normalization or standardization of numerical features, such as BMI, is detailed to ensure uniformity.
- **Categorical Variable Encoding:**
  If applicable, encoding of categorical variables into numerical format is addressed to align with the requirements of the Random Forest model.

E. *T-Test Analysis Features:*

- **Grouping Variable:**
  The specification of the grouping variable for T-Test analysis, such as predicted diabetes status (e.g., diabetic vs. non-diabetic), is outlined.
- **Continuous Variable:**
  The selection of BMI as the continuous variable of interest for comparing mean values between groups is clearly defined.

F. *IBM SPSS Setup:*

- **Data Import:**
  Ensuring the proper importation of the dataset, encompassing patient information including BMI and diabetes outcomes, into IBM SPSS is described.
- **Variable Definition:**
  The definition of variables in SPSS, including the specification of types (numeric, categorical) and labels, is detailed.

G. *Evaluation Metrics:*

- **Confusion Matrix Metrics:**
  The definition of metrics such as accuracy, precision, recall, and F1 score, calculated based on the predictions of the Random Forest model, is explicated.
- **T-Test Results:**
  The interpretation of T-Test results, including the t-statistic, p-value, and confidence intervals, to determine the statistical significance of BMI differences between predicted diabetic and non-diabetic groups is described.

H. *Cross-Validation Setup:*

- **Fold Configuration:**
  If cross-validation is employed, details regarding the configuration of folds and the randomization process are specified to ensure a robust evaluation of the model.

I. *Ethical Considerations:*

- **Data Privacy Measures:**
  Measures to ensure patient data privacy and compliance with ethical standards are incorporated, recognizing the sensitivity of health-related information.
- **Bias Mitigation:**
  Strategies to identify and mitigate potential biases in the dataset and the model are outlined, ensuring ethical and unbiased predictions.

J. *Documentation:*

- **Record Keeping:**
  The necessity for comprehensive documentation, encompassing the feature selection process, preprocessing steps, and model configuration, is emphasized to facilitate reproducibility and support future iterations of the project

V. Approach

A. *Baseline*

Here in this solution we are using Random forest algorithm to predict the results whether the patient is a diabetic or not. By using Random forest we are defining that the number of decision trees we are using in the model will get a voting based result and then the model aggregates results from the trees and predicts the results. Because of using the Random Forest ML model we can tune the model by tweaking the hyperparameters and the features to drill down while evaluating the results by the decision trees and also we are calculating the evaluation metrics to define the accuracy of the model, with that we are also calculating the Type Errors to define whether the categories taken into consideration are a viable results or not.

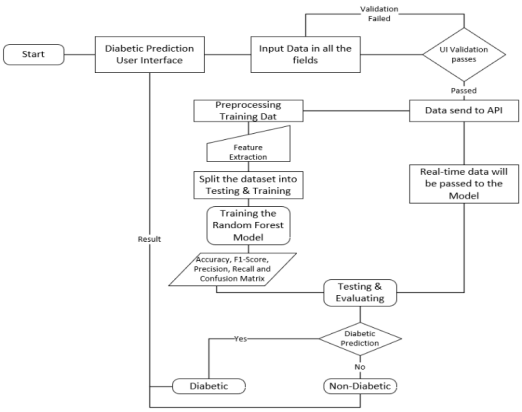
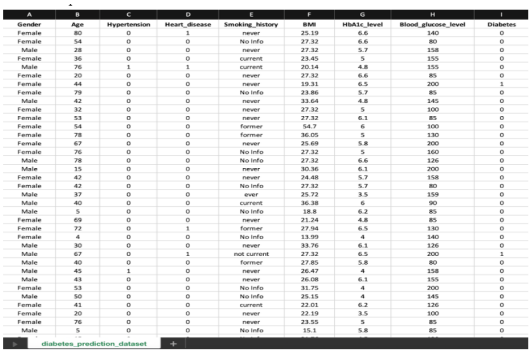


Fig. 1. Model : To provide a clear description of the project we are trying to predict the Diabetic patients whether they are diabetic or not based on the inputs provided by the users and then those inputs will be passed to the ML model which is already pre-trained by us to evaluate the real-time data and predict the results.



Gender	Age	Hypertension	Heart_disease	Smoking_history	BMI	HbA1c_level	Blood_glucose_level	Diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5	155	0
Male	76	1	1	current	20.14	4.8	155	0
Female	20	0	0	never	27.32	6.6	85	0
Female	44	0	0	never	19.31	6.5	200	1
Female	79	0	0	No Info	23.86	5.7	85	0
Male	42	0	0	never	33.64	4.8	145	0
Female	32	0	0	never	27.32	5	100	0
Female	53	0	0	never	27.32	6.1	85	0
Female	54	0	0	former	54.7	6	100	0
Female	78	0	0	former	36.05	5	130	0
Female	67	0	0	never	25.69	5.8	200	0
Female	76	0	0	No Info	27.32	5	160	0
Male	78	0	0	No Info	27.32	6.6	126	0
Male	15	0	0	never	30.36	6.1	200	0
Female	42	0	0	never	24.48	5.7	158	0
Female	42	0	0	No Info	27.32	5.7	80	0
Male	37	0	0	ever	25.72	3.5	159	0
Male	40	0	0	current	36.38	6	90	0
Male	5	0	0	No Info	18.8	6.2	85	0
Female	69	0	0	never	21.24	4.8	85	0
Female	72	0	1	former	27.94	6.5	130	0
Female	4	0	0	No Info	13.99	4	140	0
Male	30	0	0	never	33.76	6.1	126	0
Male	67	0	1	not current	27.32	6.5	200	1
Male	40	0	0	former	27.85	5.8	80	0
Male	45	1	0	never	26.47	4	158	0
Male	43	0	0	never	26.08	6.1	155	0
Female	53	0	0	No Info	31.75	4	200	0
Male	50	0	0	No Info	25.15	4	145	0
Female	41	0	0	current	22.01	6.2	126	0
Female	20	0	0	never	22.19	3.5	100	0
Female	76	0	0	never	23.55	5	85	0
Male	5	0	0	No Info	15.1	5.8	85	0

diabetes_prediction_dataset

Fig. 2. Sample view of our dataset

B. *Proposed Methods*

: In this project we are trying to calculate the Evaluation Metrics such as Confusion matrix, Precision, Recall, F1-Score, and Accuracy, with that we are also going to use the T-Test Analysis for the dataset we choose on two different groups and find out what exactly the hypothesis will return. Here you can see the below flow diagram of the project how we are going to take the data from the UI and pass it to the model and then how we are going to display the results in the UI.

VI. Evaluation

A. *Evaluation metric:*

Confusion matrix:

A confusion matrix is a table that lists the outcomes of forecasts. It reveals how many samples the model correctly predicted and how many it missed. Four outcomes are possible, and they are as follows:

1) **True Positives(TP)**: A true positive occurs when the model correctly classifies a sample as positive.
2) **False Positive(FP)**: A false positive occurs when a sample is mistakenly classified as the positive class by the model.
3) **True Negative(TN)**:A true negative occurs when the model correctly classifies a sample as belonging to the negative class.
4) **False Negative(FN)**:A false negative occurs when a sample is mistakenly assigned to the negative class by the model.

VII. Dataset

A. *Data Description*

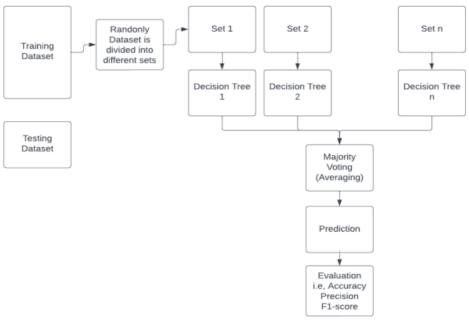
: The Diabetes prediction dataset contains 9 distinct features gathered from a dataset of 100,000 records. These features are derived from a combination of medical and demographic information obtained from patients, and they also include data regarding the patients' diabetes status (indicating whether they have tested positive or negative for diabetes). The primary purpose of this dataset is to support the creation of machine learning models designed for the detection of diabetes. These features are crucial in the training and evaluation(testing) of machine learning models that can classify patient records as indicating the presence or absence of diabetes. This dataset is well-suited for building machine learning models aimed at predicting diabetes in patients based on their medical history and demographic details. Overall, this dataset includes variables like age, gender, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. It proves to be a valuable asset for healthcare professionals aiming to identify individuals at risk of diabetes and creating personalized treatment plans. Furthermore, researchers can use this dataset to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

VIII. Experiments

- **Quantitative Experiments:** Develop a predictive model for diabetes risk based on the features available in the dataset.We are using the "Diabetes Prediction Dataset" from Kaggle. Include information about the size of the dataset, its source, and any relevant details.
- **Qualitative Analysis:** Qualitative analysis can enrich quantitative analysis by providing a deeper understanding of diabetic patients, perspectives and experiences related to diabetes and BMI.

IX. Analysis

The project overview is to sets the stage for the reader by explaining the motivation behind the project and the significance it holds in the domain of predictive healthcare analytics. The motivation of this project is to emphasize the global health challenge of diabetes and the role of predictive analytics in mitigating its impact. The significance section expands on the practical applications of the project in improving patient outcomes and contributing to preventive measures. The document clearly outlines the main objectives of developing a predictive model for diabetes risk, with a



focus on BMI. Then the features section provides detailed about the machine learning model, evaluation metrics, and the IBM SPSS analysis. The related work effectively supports the project's approach and methodologies, specifically the use of Random Forest for diabetes prediction based on patient BMI.The approach section distinguishes between the baseline and proposed methods, focusing on Random Forest and T-Test analysis. The results are presented, emphasizing evaluation metrics like accuracy, precision, and F1-score. This provides an initial understanding of the model's performance. This document includes the diabetes prediction dataset, describing its features and relevance. The experiments highlight quantitative and qualitative analysis approaches, showing a comprehensive strategy for model evaluation.

X. Implementation

The project overview is to sets the stage for the reader by explaining the motivation behind the project and the significance it holds in the domain of predictive healthcare analytics. The motivation of this project is to emphasize the global health challenge of diabetes and the role of predictive analytics in mitigating its impact. The significance section expands on the practical applications of the project in improving patient outcomes and contributing to preventive measures. The document clearly outlines the main objectives of developing a predictive model for diabetes risk, with a focus on BMI. Then the features section provides detailed about the machine learning model, evaluation metrics, and the IBM SPSS analysis. The related work effectively supports the project's approach and methodologies, specifically the use of Random Forest for diabetes prediction based on patient BMI.The approach section distinguishes between the baseline and proposed methods, focusing on Random Forest and T-Test analysis. The results are presented, emphasizing evaluation metrics like accuracy, precision, and F1-score. This provides an initial understanding of the model's performance. This document includes the diabetes prediction dataset, describing its features and relevance. The experiments highlight quantitative and qualitative analysis approaches, showing a comprehensive strategy for model evaluation.

XI. Preliminary Results

In the preliminary results we are able to train the model and fetch the predictions, where we can handle the real-time data scenarios like we pass the real data and the model will predict the output as described and given the example screenshots in the above section.

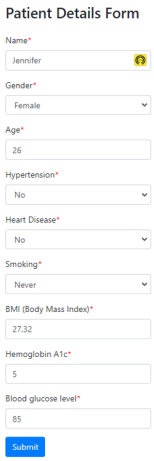


Fig. 3. Diabetic Prediction based on Patient BMI

Fig. 4. Result

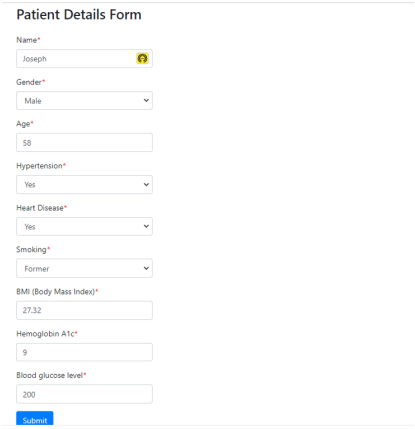


Fig. 5. Diabetic Prediction based on Patient BMI

Fig. 6. Result

XII. Evaluation Metrics

In the evaluation metrics we are able to achieve the accuracy score of 97 percent as the model is performing well and the

precision score of 97 percent and f1-score of 98 percent and other results are mentioned in the screenshot.

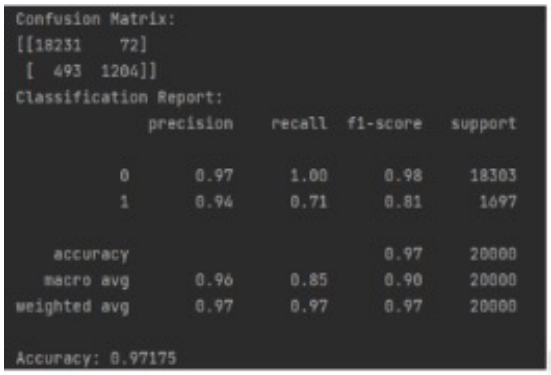


Fig. 7. Confusion Matrix

XIII. Project Management

A. *Implementation status report*

- **Work completed**: Initial Model building and UI part has been completed and working perfectly.
  1) **Vaishnavi Nomula:**
     Description: Worked on creating the user interface,Implemented the UI validations.
     Responsibility: Created Html View with fields, Implemented regular expression based validations to avoid cosmetic errors while communicating with the API.
     Contributions: Individual - 100%
  2) **Surya Sai Raj Lakkoju:**
     Description: Worked on building the random forest model and completed the process of training and testing,Data Cleaning performed on the dataset to improve the performance of the model.
     Responsibility: Extracted features from the dataset then trained and tested the model and evaluation metrics . Data Cleaning had been performed on the missing values, duplicate records, and Nan values from the dataset.
     Contributions: Individual - 100%
  3) **Amisha Patra:**
     Description: Worked on implementing the results on SPSS still it is in progress, currently facing issues in evaluating the metrics.
     Responsibility: Extracted features from the dataset then checked whether the samples are rejected by null hypothesis or fail to reject the null hypothesis
     Contributions: Individual - 100%
  4) **Bhavya Golla:**
     Description: Worked on creating the evaluation metrics on the Random Forest model like Confusion Matrix, Accuracy, and Classification Report,implemented Data Visualizations on the dataset.
     Responsibility: Based on the predicted result and the features undergone through the process of training and testing in the model building phase we evaluate the evaluation metrics, Added a few plots such as bar graphs to understand how the data is looking.
     Contributions: Individual - 100%
  5) **Deepanvi Balusuguri**: Description: Worked on creating Flask Web API to communicate with the Random Forest Model,implemented Data Visualizations on the dataset .
     Responsibility: Worked on creating the Python API using Flask web package and integrated the communication between the UI and the Model, and added Heatmaps and other data visualization.
     Contributions: Individual - 100%

XIV. UI Validation

- The JavaScript is designed in such a way to validate different type of form fields. Starting from 'fieldValue' function which is used to remove leading and trailing whitespaces.
- (field.parentElement.lastElementChild).addClass('display-none'); this code is used to hide error messages associated with fields. To handle different types of form fields such as text, number and select-one witch statement is used based on the field type.

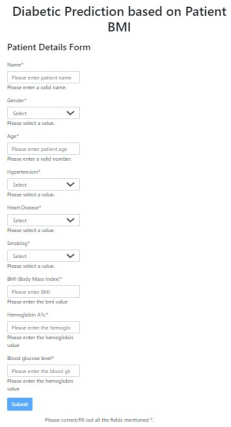


Fig. 8. The above results shows the UI Validations forms

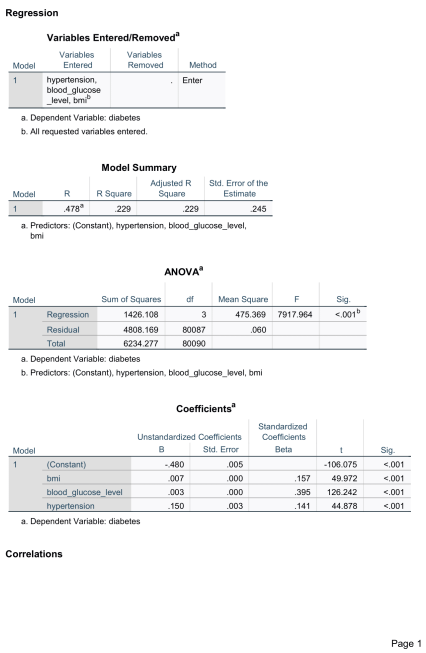
- For field which are 'text' type is checked for the entered value is alphabetic or not using alphabetic.test(fieldValue). If not, an error is displayed.
- For field which are 'number' type is checked for the entered value is numeric or not using isNan(fieldValue). If not, an error is displayed.
- For field which are 'select-one' type is checked if a valid option is selected or not. If not, an error is displayed.
- With this validation, the user input is checked if the input meets the criteria for each field. The error messages are

displayed, which helps the user to know if the validation is failed.

XV. SPSS

A. *REGRESSION*

- The spss for liner regression is done for the diabetes dataset.
- Here the dependent variable is Diabetes and the independent variables are hypertension, blood_glucose_level, bmi.
- When performing Simple Linear Regression on this dataset we get the ANOVA table, Coefficients Table, Model Summary Table, and Variable Table.
- From the Model Summary, we get the value of Adjusted R Square value .478 outcome variable hypertension, blood_glucose_level, bmi is 22.9% influence by the current independent variable.
- · The ANOVA Table will decide if the model is significant or not. Here the Sig value ¡0.05, then there is Linear Relationship between the independent and dependent variable, there is some random slope value.



Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	hypertension, blood_glucose_level, bmi[b]	.	Enter

a. Dependent Variable: diabetes
b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.478[a]	.229	.229	.245

a. Predictors: (Constant), hypertension, blood_glucose_level, bmi

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1426.108	3	475.369	7917.964	<.001[b]
	Residual	4808.169	80087	.060		
	Total	6234.277	80090			

a. Dependent Variable: diabetes
b. Predictors: (Constant), hypertension, blood_glucose_level, bmi

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-.480	.005		-106.075	<.001
	bmi	.007	.000	.157	49.972	<.001
	blood_glucose_level	.003	.000	.395	126.242	<.001
	hypertension	.150	.003	.141	44.878	<.001

a. Dependent Variable: diabetes

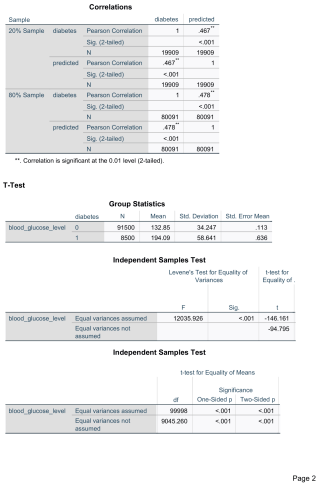
Correlations

Page 1

Fig. 9. The above results shows the SPSS output of Regression

B. *Co-Relation*

- The Coefficient Box gives the linear equation.
- The Pearson Correlation value is 0.467 for 20% sample (Testing) and 0.478 for 80% of the Sample (Training). The values are nearly similar to each other.
- Hence, we can conclude that our model is neither underfitting nor overfitting.



Correlations

Sample			diabetes	predicted
20% Sample	diabetes	Pearson Correlation	1	.467**
		Sig. (2-tailed)		<.001
		N	19909	19909
	predicted	Pearson Correlation	.467**	1
		Sig. (2-tailed)	<.001	
		N	19909	19909
80% Sample	diabetes	Pearson Correlation	1	.478**
		Sig. (2-tailed)		<.001
		N	80091	80091
	predicted	Pearson Correlation	.478**	1
		Sig. (2-tailed)	<.001	
		N	80091	80091

**. Correlation is significant at the 0.01 level (2-tailed).

T-Test

Group Statistics

	diabetes	N	Mean	Std. Deviation	Std. Error Mean
blood_glucose_level	0	91500	132.85	34.247	.113
	1	8500	194.09	58.641	.636

Independent Samples Test

		Levene's Test for Equality of Variances: F	Levene's Test for Equality of Variances: Sig.	t-test for Equality of: t
blood_glucose_level	Equal variances assumed	12035.926	<.001	-146.161
	Equal variances not assumed			-94.795

Independent Samples Test

		t-test for Equality of Means: df	Significance One-Sided p	Significance Two-Sided p
blood_glucose_level	Equal variances assumed	99998	<.001	<.001
	Equal variances not assumed	9045.260	<.001	<.001

Page 2

Fig. 10. The above results shows the SPSS output of Co-relation and T -test

C. *T-test Analysis*

- The record for blood_glucose_level is analyzed for diabetes the values are for the group as 0 and 1 which is classified as a person with diabetes or not.
- In the Group statistics the values for mean, standard deviation and standard error mean and the number of patients for groups 0 and 1.
- For group 0- The mean is 132.85, std deviation is 34.247, std. error mean is .113, the number of patients are 91500.
- For group 0- The mean is 194.09, std deviation is 58.641, std. error mean is .636, the number of patients are 8500.
- From the independent samples test table we get 'p-value'. The value of p is ¡0.05 (alpha value) which depicts we can REJECT null-hypothesis.
- From the t-test analysis we conclude that the groups 0 and 1 are unlike.

XVI. DATA CLEANING

A. *Duplicates:*

Duplicates can be classified into two types, that is exact and near duplicates. The exact duplicates are easier to identify as they are the same values. Near duplicates are the result of formatting, missing values, etc. Duplicates can be handled by deciding what to do with them in the dataset we have taken. We can delete or drop the values in pandas python we use the drop-duplicates() method to drop duplicates of the row or columns of data.

B. *Null Values:*

The best way to remove the null values in the dataset is to either delete it or perform imputation. Deletion means

Independent Samples Test

		t-test for Equality of Means		
		Mean Difference	Std. Error Difference	95% Confidence Interval of the ... Lower
blood_glucose_level	Equal variances assumed	-61.242	.419	-62.063
	Equal variances not assumed	-61.242	.646	-62.509

Independent Samples Test

		t-test for Equality of Means	
		95% Confidence Interval of the ... Upper	
blood_glucose_level	Equal variances assumed	-60.421	
	Equal variances not assumed	-59.976	

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
blood_glucose_level	Cohen's d	36.952	-1.657	-1.681	-1.634
	Hedges' correction	36.952	-1.657	-1.681	-1.634
	Glass's delta	58.641	-1.044	-1.072	-1.017

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control (i.e., the second) group.

Page 3

Fig. 11. The above results shows the SPSS output of T-test Analysis

discarding rows or columns that contain null values whereas imputation is replacing the null values present in rows and columns with reasonable estimations such as mean, median, mode, or a value based on an algorithm or logic.

C. NA records:

The best way to deal with the NA records is similar to handling of null record, that is either we delete the missing data rows or impute a value in place of the NA value and the only difference from the null records is that we could try and predict the NA value using an algorithm if we wanted to.

D. Empty records:

The handling of empty records in a dataset is the same as we handle null records.

E. DATA VISUALIZATION

- Data visualization is creating easy-to-understand graphic or visual representations of a large amount of complex quantitative and qualitative data and information with the help of static, dynamic or interactive visual items.
- We here representing our diabetes data through common graphics. Here we used bar plot,histograms, subplot, heatmap and pie chart for representing data.
- The other data visualization is the heat map. It's a matrix graphical representation which contain the values which are represented in diff colors.
- Then the number of patients who are having diabetes are represented as 1 and non-diabetic patients are represented as 0. This data then plotted in a pie chart which shows the percentage of patients for these two groups. And another bar plot with the same data.

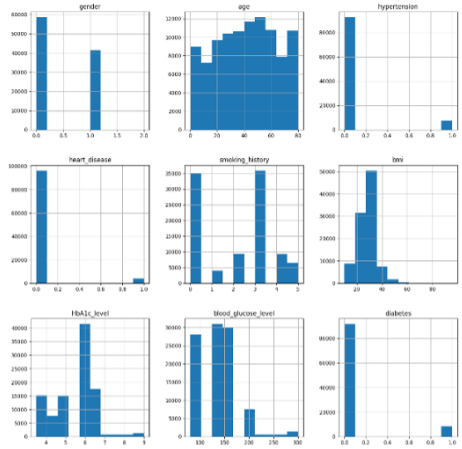


Fig. 12. The above results shows the Visualizations of Histograms

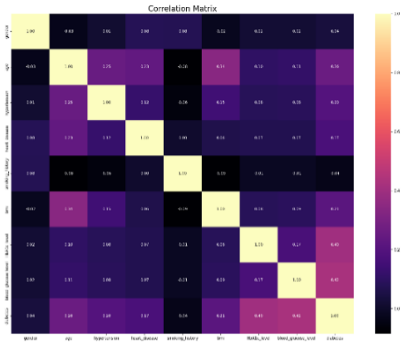


Fig. 13. The above results shows the Visualizations of Heat maps

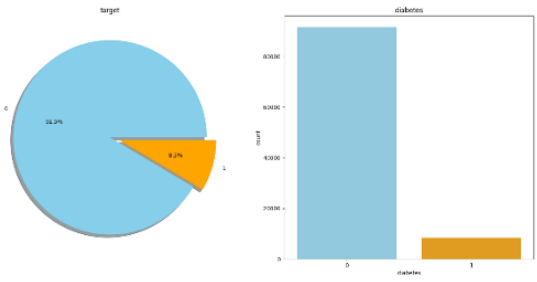


Fig. 14. The above results shows the Visualizations of Heat maps

XVII. REFERENCES/BIBLIOGRAPHY

https://ieeexplore.ieee.org/ [1]
https://www.sciencedirect.com [2]
https://ieeexplore.ieee.org/ [3]
https://www.sciencedirect.com [4]
https://ieeexplore.ieee.org/ [5]

REFERENCES

[1] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," in *2018 24th International Conference on Automation and Computing (ICAC)*, 2018, pp. 1–6.

[2] V. Jaiswal, A. Negi, and T. Pal, "A review on current advances in machine learning based diabetes prediction," *Primary Care Diabetes*, vol. 15, no. 3, pp. 435–443, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S175199182100019X

[3] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76 516–76 531, 2020.

[4] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019, 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920300557

[5] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, 2019, pp. 1–4.