International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 3 Ver. I || March 2018 || PP. 76-83

An Exploratory Data Analysis Approach to Predict Gestational Diabetes Mellitus in Pregnant Women

Praveen Kumar Misra, Shreya Gupta
Dr.Shakuntala Misra National Rehabilitation University, Lucknow

Abstract
Exploratory Data Analysis (EDA) is crucial for improving the accuracy of predictive models in Gestational Diabetes Mellitus (GDM) prediction for pregnant women. This paper outlines the significance of EDA in utilizing diverse datasets to enhance predictive models for GDM risk assessment. It begins by highlighting the complex nature of GDM and its impact on maternal and fetal health, stressing the need for accurate prediction and early intervention. AI and ML techniques are recognized for their potential in leveraging extensive datasets containing maternal characteristics, medical history, and biochemical markers. EDA's pivotal role in preprocessing raw data, detecting outliers, handling missing values, and identifying relevant features for GDM prediction is emphasized. Techniques such as data visualization and statistical analysis aid in understanding data distribution and relationships, informing feature selection and model refinement. Additionally, EDA complements advanced ML algorithms, facilitating feature engineering and model interpretability. Integrating heterogeneous data sources further enriches predictive models and improves generalizability across diverse patient populations. EDA's implications extend to optimizing personalized intervention strategies and preventive measures for high-risk pregnant women. Continued research and interdisciplinary collaborations are advocated to harness the full potential of AI and ML in GDM prediction and management. [2],[13]

I. Introduction

The Pima Indians Diabetes Dataset stands as a cornerstone in healthcare and machine learning research, providing invaluable insights into diabetes diagnosis, notably gestational diabetes mellitus (GDM) risk assessment in pregnant women. Compiled by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) during the 1980s, this dataset captures clinical and demographic profiles of Pima Indian women, a population known for its high prevalence of type 2 diabetes, offering a unique lens into the disease's epidemiology.

Containing attributes such as age, pregnancies, glucose levels, blood pressure, skinfold thickness, insulin levels, BMI, and diabetes pedigree function, alongside a binary outcome indicating diabetes presence within five years, the dataset serves as a goldmine for predictive modeling. Its richness enables exploration of intricate relationships between demographic and clinical features and diabetes occurrence, facilitating the development of robust predictive models for early diagnosis and intervention.[1]

Moreover, the dataset serves as a benchmark for evaluating machine learning algorithms' performance in diabetes prediction. Techniques ranging from logistic regression to ensemble methods have been scrutinized, contributing to an extensive repertoire of predictive models derived from this dataset.

Despite its vintage, the dataset remains pertinent amidst the persistent global diabetes challenge. Its accessibility and well-documented nature render it an ideal resource for educational, research, and algorithmic development endeavours in healthcare analytics and machine learning domains.

In this paper, we endeavour to harness the insights gleaned from the Pima Indians Diabetes Dataset to delve into the potential of artificial intelligence and machine learning techniques in diabetes predictive modelling, with a specific emphasis on implications for pregnant women susceptible to GDM. By elucidating this potential, we aim to advance ongoing efforts in combating the diabetes epidemic and enhancing healthcare outcomes for vulnerable populations.[1],[15]

II. Objective

The primary objective of this study is to explore the utility of the Pima Indians Diabetes Dataset in developing predictive models for gestational diabetes mellitus (GDM) among pregnant women. Specifically, the study aims to achieve the following objectives:

- Analyze the demographic and clinical attributes contained within the Pima Indians Diabetes Dataset to identify factors associated with the onset and progression of GDM.
- Investigate the predictive power of various machine learning algorithms, including logistic regression, decision trees, support vector machines, neural networks, and ensemble methods, in accurately classifying pregnant women at risk of developing GDM based on the dataset attributes.

- Evaluate the performance of developed predictive models through metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC), ensuring robustness and generalizability.
- Explore the impact of feature selection techniques and model optimization strategies on enhancing the predictive accuracy and interpretability of the developed models for GDM prediction.
- Assess the clinical relevance and feasibility of integrating the developed predictive models into healthcare practice to facilitate early identification, intervention, and personalized management of GDM among pregnant women.

By achieving these objectives, this study aims to contribute to the advancement of predictive modelling techniques for GDM using the Pima Indians Diabetes Dataset, ultimately improving healthcare outcomes for pregnant women and their offspring at risk of this significant health condition.

III. Dataset Description

Detailed Description of the Pima Indians Diabetes Dataset:

The Pima Indians Diabetes Dataset is a collection of clinical and demographic data originally gathered by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in the 1980s. It comprises information from Pima Indian women, a population known for having one of the highest prevalence rates of type 2 diabetes in the world. This dataset has since become a cornerstone in diabetes research and predictive modeling, serving as a benchmark for evaluating algorithms aimed at predicting diabetes onset.

3.1 Attributes:

- Pregnancies: Number of pregnancies the individual has had.
- Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test.
- BloodPressure: Diastolic blood pressure (mm Hg).
- SkinThickness: Triceps skinfold thickness (mm).
- Insulin: 2-Hour serum insulin (mu U/ml).
- BMI: Body mass index (weight in kg/(height in m)^2).
- DiabetesPedigreeFunction: Diabetes pedigree function, a measure of the hereditary risk.
- Age: Age of the individual in years.
- Outcome: Binary variable indicating the presence (1) or absence (0) of diabetes within a five-year follow-up period.

1. Data Format:

- The dataset is typically provided in a tabular format, with each row representing an individual and each column representing an attribute.
- Values may be numerical or categorical, with numerical attributes ranging from continuous to discrete values.
- Missing values may be present and need to be handled appropriately during data preprocessin

3.2 Characteristics:

- Size: The dataset typically consists of around 768 instances or observations.
- Balance: The dataset is often imbalanced, with a higher proportion of individuals without diabetes compared to those with diabetes.
- Diversity: The dataset encompasses a diverse range of attributes, including demographic factors (age, pregnancies), clinical measurements (glucose, blood pressure), and anthropometric data (BMI, skinfold thickness).
- Follow-up Period: The outcome variable indicates the presence or absence of diabetes within a five-year follow-up period, providing longitudinal insights into disease progression.

3.3 Utilization:

- Research: The dataset has been extensively used in research studies exploring various aspects of diabetes prediction, risk factors, and intervention strategies.
- Education: Due to its accessibility and well-documented nature, the dataset serves as a valuable resource for educational purposes in data science and healthcare analytics courses.
- Algorithm Development: Researchers and data scientists utilize the dataset to develop and evaluate machine learning algorithms for diabetes prediction, ranging from traditional statistical methods to advanced deep learning techniques.

3.4 Importance:

- Benchmarking: The dataset serves as a benchmark for evaluating the performance of predictive models aimed at diabetes diagnosis and prediction.
- Real-world Relevance: The high prevalence of diabetes among Pima Indian populations lends real-world relevance to the insights derived from analyzing this dataset, with implications for healthcare interventions and public health policies.
- Generalizability: Findings from studies utilizing this dataset may inform predictive modeling efforts for diabetes across diverse populations, contributing to the development of globally applicable healthcare solutions.

```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
```

IV. Exploratory Data Analysis

Data Exploration

- The initial phase involves delving into the dataset's structure, dimensions, and attribute types. This aids in gaining a comprehensive understanding of the data under scrutiny.
- Furthermore, the identification and handling of missing values play a pivotal role in ensuring the integrity and reliability of subsequent analyses.[14]

```
# Display the first few rows of the dataset
print("First few rows of the dataset:")
print(data.head())
```

```
First few rows of the dataset:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
```

Summary Statistics

- Subsequently, descriptive statistics are computed for each attribute, encompassing metrics such as mean, median, standard deviation, minimum, and maximum values. These statistics offer crucial insights into the central tendencies and variability within the dataset.

```
print(data.describe())
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	\
count	768.000000	768.000000	768.000000	768.000000	768.000000	
mean	3.845052	120.894531	69.105469	20.536458	79.799479	
std	3.369578	31.972618	19.355807	15.952218	115.244002	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	99.000000	62.000000	0.000000	0.000000	
50%	3.000000	117.000000	72.000000	23.000000	30.500000	
75%	6.000000	140.250000	80.000000	32.000000	127.250000	
max	17.000000	199.000000	122.000000	99.000000	846.000000	

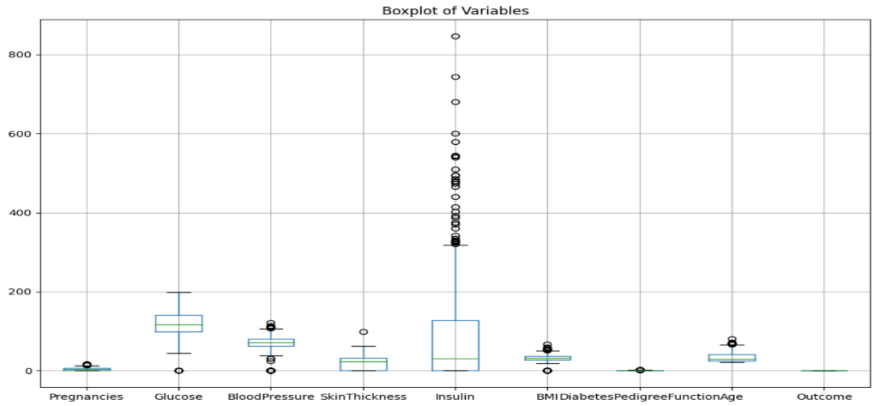
	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000
mean	31.992578	0.471876	33.240885	0.348958
std	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.078000	21.000000	0.000000
25%	27.300000	0.243750	24.000000	0.000000
50%	32.000000	0.372500	29.000000	0.000000
75%	36.600000	0.626250	41.000000	1.000000
max	67.100000	2.420000	81.000000	1.000000

Data Visualization

- Harnessing the power of visualization, histograms and boxplots are employed to shed light on the distributions of numeric variables and identify potential outliers.
- Bar plots are instrumental in visually representing the distribution of categorical variables, such as Outcome, thereby facilitating a deeper understanding of class distribution.
- Moreover, the construction of a correlation matrix and heatmap serves to unveil underlying relationships between variables, pinpointing potential predictors of diabetes.
- Pairwise scatterplots further augment the analysis by enabling the visualization of bivariate relationships between numeric variables, offering valuable insights into potential correlations and trends.[2][4][14]

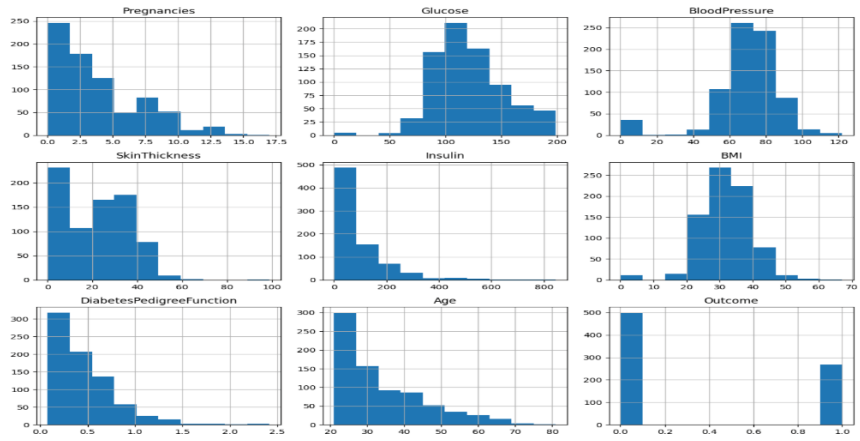
BOXPLOTS

```
# Boxplot for each variable
data.boxplot(figsize=(12, 8))
plt.title('Boxplot of Variables')
plt.show()
```



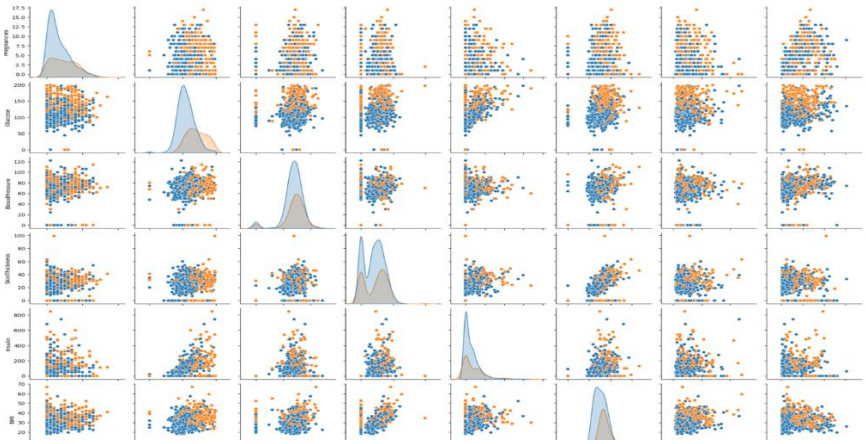
HISTOGRAM

```
# Histograms for each variable
data.hist(figsize=(12, 10))
plt.tight_layout()
plt.show()
```



PAIRPLOT

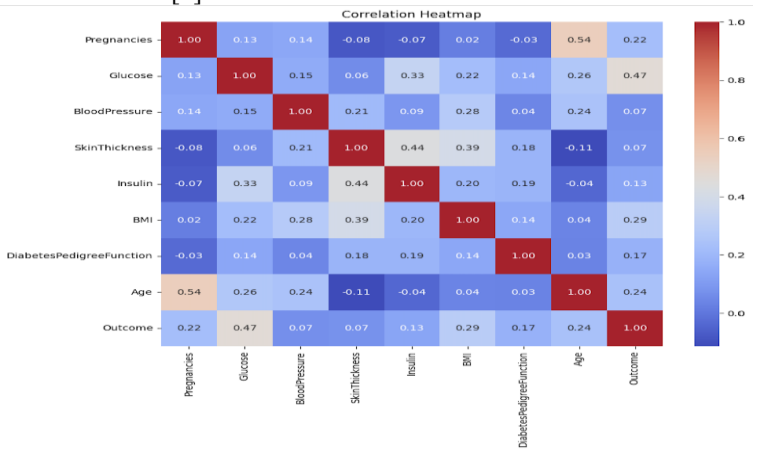
```
sns.pairplot(data, hue='Outcome', diag_kind='kde')
plt.show()
```



V. Findings And Insights

Correlation Analysis

- The analysis reveals a notable positive correlation between Glucose levels and the likelihood of diabetes. This underscores the pivotal role played by Glucose levels in diabetes diagnosis.
- Additionally, Age and BMI exhibit significant correlations with diabetes outcome, accentuating the multifactorial nature of diabetes.[4]



```
# Importing necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"
names = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
data = pd.read_csv(url, names=names)

# Compute the correlation matrix
correlation_matrix = data.corr()

# Plotting the heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```

Outlier Detection

- Through meticulous examination, outliers are identified in certain variables such as Insulin and Blood Pressure. These outliers warrant further scrutiny and may harbor valuable insights into underlying trends or anomalies within the dataset.
- Outlier detection is a crucial step in data preprocessing and analysis, especially in predictive modeling tasks. Outliers are data points that significantly deviate from the rest of the dataset and can distort statistical analyses and machine learning models. Here's how you can perform outlier detection on the Pima Indians Diabetes Dataset using the Isolation Forest algorithm, which is a popular method for detecting outliers:[4]

```
# Importing necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import IsolationForest

# Load the dataset
url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"
names = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
data = pd.read_csv(url, names=names)

# Splitting data into features (X) and target (y)
X = data.drop('Outcome', axis=1)
y = data['Outcome']

# Standardizing the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Isolation Forest for outlier detection
clf = IsolationForest(contamination=0.1, random_state=42)
y_outliers = clf.fit_predict(X_scaled)

# Outlier detection results (-1: outlier, 1: inlier)
outliers_df = data.copy()
outliers_df['Outlier'] = y_outliers
print(outliers_df[outliers_df['Outlier'] == -1])
```

```
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
4              0      137             40             35      168  43.1
7             10      115              0              0        0  35.3
8              2      197             70             45      543  30.5
9              8      125             96              0        0   0.0
12            10      139             80              0        0  27.1
..           ...      ...            ...            ...      ...   ...
715            7      187             50             33      392  33.9
744           13      153             88             37      140  40.6
753            0      181             88             44      510  43.3
759            6      190             92              0        0  35.5
763           10      101             76             48      180  32.9

     DiabetesPedigreeFunction  Age  Outcome  Outlier
4                       2.288   33        1       -1
7                       0.134   29        0       -1
8                       0.158   53        1       -1
9                       0.232   54        1       -1
12                      1.441   57        0       -1
..                        ...  ...      ...      ...
715                     0.826   34        1       -1
744                     1.174   39        0       -1
753                     0.222   26        1       -1
759                     0.278   66        1       -1
763                     0.171   63        0       -1

[77 rows x 10 columns]
```

Class Imbalance

Class imbalance refers to the situation where the distribution of classes (or labels) in a dataset is skewed, meaning that one class is significantly more prevalent than the others. In the context of the Pima

Indians Diabetes Dataset, class imbalance would occur if there are substantially more instances of one outcome (e.g., no diabetes) compared to the other outcome (e.g., diabetes).[8]

To address class imbalance in the dataset, several techniques can be employed:

1. **Resampling Methods**:

- **Undersampling**: Randomly remove samples from the majority class to balance the class distribution. This may result in loss of information.
- **Oversampling**: Randomly duplicate samples from the minority class or generate synthetic samples to match the size of the majority class. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) can be used for this purpose.[9]

Algorithmic Techniques

- **Cost-sensitive Learning**: Assign different costs to misclassification errors for different classes to penalize misclassifications of the minority class more heavily.
- **Ensemble Methods**: Use ensemble techniques like bagging or boosting with techniques that are inherently robust to class imbalance, such as Balanced Random Forest Classifier or Easy Ensemble Classifier.[10]

Algorithm-specific Methods

- Some algorithms have parameters or techniques specifically designed to handle class imbalance, such as class weight parameter in scikit-learn's Logistic Regression or Random Forest Classifier.

Evaluation Metrics

- Use evaluation metrics that are robust to class imbalance, such as precision, recall, F1-score, or area under the ROC curve (AUC-ROC). These metrics provide a more comprehensive assessment of model performance compared to accuracy, especially in imbalanced datasets.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

# Assuming X_train, X_test, y_train, y_test are already defined

# Initialize Logistic Regression model with class weights
model = LogisticRegression(class_weight='balanced')

# Fit the model
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.81      0.69      0.74        99
           1       0.56      0.71      0.62        55

    accuracy                           0.69       154
   macro avg       0.68      0.70      0.68       154
weighted avg       0.72      0.69      0.70       154
```

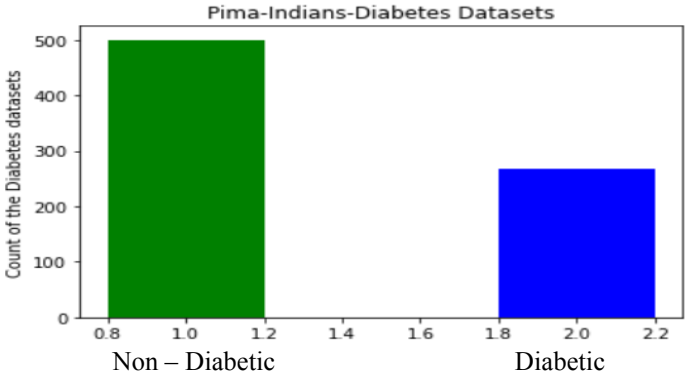
Testing & Comparing the Algorithms

After applying the algorithms, we will compute the accuracy of each and every model i.e., Logistic Regression, K-Nearest Neighbours, Support Vector Machine and Random Forest Classifier. And after analysing and comparing the accuracy for training and Testing dataset, we will write the result in tabular format as shown below. And We can conclude that Random Forest Classifier Model is the best Algorithm for Prediction of the Diabetes Disease.[10][12]

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	78.478261	77.597403
1	K-nearest neighbors	79.782609	69.155844
2	Support Vector Machine	100.000000	65.584416
3	Random Forest Classifier	100.000000	74.025974

Accuracy Results for Diabetes Prediction

The graph shown below shows the No. of persons who are Diabetic and Non-Diabetic. The green Colour bar shows the Non-Diabetic person and blue colour bar shows the Diabetic person.



VI. Conclusion

The comprehensive Exploratory Data Analysis conducted on the Pima Indians Diabetes Dataset has yielded invaluable insights into the intricate interplay between diagnostic measurements and the likelihood of diabetes among female patients of Pima Indian heritage. By discerning patterns, correlations, and potential predictors of diabetes, this analysis contributes to the collective understanding of this complex medical condition. These algorithms are used in building up the model for diabetes prediction. With the advancement of technology, machine learning will be very effective in our health care sector. It can help in saving human's life in any or the other way. Early prediction in this type of disease will be a boon for our generation and upcoming generation. But People should also change their lifestyle and should adopt a proper routine in order to live a healthy life forever.

The identified findings serve as a catalyst for future research endeavours aimed at developing predictive models and innovative interventions for diabetes diagnosis and management.

References

[1]. National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). Diabetes Data. Retrieved from https://www.kaggle.com/uciml/pima-indians-diabetes-database.
[2]. https://www.researchgate.net/publication/308007227_Exploratory_Data_Analysis
[3]. https://www.ijitee.org/wp-content/uploads/papers/v8i12/L35911081219.pdf
[4]. Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, and Chad Marston, "A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets," Visual Informatics, Volume 2, Issue 4, December 2018, pp. 235-253.
[5]. John T. Behrens, "Principles and Procedures of Exploratory Data Analysis," Psychological Methods, 1997, Vol. 2, No. 2, pp.131-160.
[6]. Chokey Wangmo, "An Exploratory Study On Bank Lending To SME Sector In Bhutan," International Journal of Scientific & Technology Research, volume 6, issue 11, November 2017, pp. 47-51.
[7]. Matthew Ntow-Gyamfi and Sarah Serwaa Boateng, "Credit Risk and Loan Default among Ghanaian Banks: An Exploratory Study," Management Science Letters, Vol. 3, 2013, pp.753–762.
[8]. He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." IEEE Transactions on Knowledge and Data Engineering 21.9 (2009): 1263-1284.
[9]. X. Francis Jency, V. P. Sumathi, Janani Shiva Sri, "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients," International Journal of Recent Technology and Engineering (IJRTE), Volume-7 Issue-4S, November 2018, pp.176-179.
[10]. K. Ulaga Priya, S. Pushp, K. Kalaivani, A. Sartiha, "Exploratory Analysis on Prediction of Loan Privilege for Customers using Random Forest," International Journal of Engineering & Technology, Vol. 7, Issue 2.21, 2018, pp. 339-341.
[11]. Introduction To Machine Learning using Python [Online], Available: https://www.geeksforgeeks.org/introduction-machine-learning-using-p ython/
[12]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5796274/
[13]. Exploratory data analysis – From Wikipedia, the free encyclopedia [Online], Available: https://en.wikipedia.org/wiki/Exploratory_data_ analysis
[14]. https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf
[15]. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.