# **Exploratory Data Analysis**

The end outcome of our research is to build a predictive model which can classify diabetes patients based on different predictors. To improve the Accuracy of the model, performing an exploratory data analysis is very important for selecting the significant features, preprocessing the data, handling the outliers.

The diabetes data BFSS 2015 has 21 variables and the ‘Diabetes’ column can be considered as response variables. From the histogram [Fig 2.1 and Fig 2.2], we can infer that some of the variables are discrete like “High BP”, “HighChol” etc. and continuous variables as well like “BMI”, “Age” etc.,

Apart from the same we have visualized multiple variables, and tried to visualize the relation between various predictors and the response variable.

A group of blue bars

Description automatically generated with medium confidence

Fig: 2.1 Histograms of variables of the dataset

A screenshot of a graph

Description automatically generated

Fig: 2.2 Histograms of variables of the dataset

A blue circle with orange and black numbers

Description automatically generated A graph of diabetes

Description automatically generated

Fig 3: Distribution of diabetic and non-diabetic Fig 4: Diabetes frequency across High BP Flag

We have tried to visualize the diabetic and non-diabetic distribution across the dataset which indicates that the dataset is pretty imbalanced as almost ~84% of the response variable is consisted of non-diabetics where as only 16% of the dataset is of diabetic [Fig 3].

[Fig 4] Now we have tried to visualize the frequency of HighBP flag across diabetic and non-diabetic flag. For HighBP flag 0 denotes the records without HighBP where 1 denotes the records with HighBP.The Non-diabetic patients are less prone to HighBP where the diabetic people are more vulnerable to HighBP. This aligns to other heart related variables “HighChol” and “HeartDisease” flag.

[Fig 5] depicts frequency of Diabetic and Non-diabetics across cohort based on High BP and High Cholesterol. In the following paragraph we have mentioned the findings

A graph of diabetes and diabetes

Description automatically generated A graph of diabetes

Description automatically generated

Fig 5: Distribution of diabetic & non-diabetic across high BP Fig 4: Diabetes frequency across Phys activity variable

high chol

This figure depicts that Physical activity reduces the vulnerability of diabetes by the number of diabetes patients count, when we talk about number of non-diabetics the non-diabetics are significantly higher in the cohort where the subjects are involved in physical activity. [Fig6]

A graph of diabetes and smoker

Description automatically generated

Fig 7: Diabetes frequency across Smoker

[Fig 7] depicts frequency of Diabetic and Non-diabetics across smoker and non-smoker cohort. In both the cohorts the number of diabetic is comparable where as the non-diabetic records are higher in non-smoker cohort.

A graph of a diabetes disease

Description automatically generated with medium confidence

Fig 8: Diabetes frequency for Veggies Consumer cohort

[Fig 8] depicts the count of Diabetic and Non-diabetic patients across Veggies consumer vs non-consumer cohort.

The number of non-diabetics is significantly larger in the veggies consumer cohort. We have investigated other variables like fruit consumption. The number of diabetics is also higher in the veggies consumer cohort, therefore we would further investigate this variables.

Thereafter the correlation was checked between the variables and also with the response variables.

[Fig 9] depicts some of the variables are highly correlated; such as the correlation heatmap show GenHlth and PhysHlth are highly correlated with each other.(positive relation) also GenHlth and Income are highly correlated with each other .(negative relation).

A chart with numbers and letters

Description automatically generated with medium confidence

Fig 9: Correlation of the variables

A graph with orange bars

Description automatically generated

Fig 10: Correlation with Diabetes Binary

[Fig 10] depicts there are some highly correlated variables with Diabetes flag like Genital health, HighBP etc., and the least correlated variables are like “Anyhealthcare” and “sex” etc.,

Also, there are variables like “Income”, “Education”, “PhysActivity” which are negatively correlated with Diabetes variable.

## Result intrepretetion

TABLE III

| Summary of Result | | | | | |
| --- | --- | --- | --- | --- | --- |
| Model | Class | Precision | Recall | F1 Score | Accuracy |
| LR | 0 | 0.80 | 0.93 | 0.86 | 0.85 |
| 1 | 0.92 | 0.76 | 0.83 |
| DT | 0 | 0.78 | 0.96 | 0.86 | 0.85 |
| 1 | 0.94 | 0.74 | 0.83 |
| KNN | 0 | 0.75 | 0.94 | 0.83 | 0.81 |
| 1 | 0.92 | 0.68 | 0.78 |
| Random Forest | 0 | 0.80 | 0.96 | 0.87 | 0.86 |
| 1 | 0.95 | 0.77 | 0.85 |
| SVM | 0 | 0.8 | 0.96 | 0.87 | 0.86 |
| 1 | 0.95 | 0.76 | 0.85 |
| XgBoost | 0 | 0.75 | 0.97 | 0.85 | 0.83 |
| 1 | 0.96 | 0.68 | 0.8 |
| RNN | 0 | 0.82 | 0.92 | 0.87 | 0.86 |
| 1 | 0.91 | 0.8 | 0.85 |
| FNN | 0 | 0.82 | 0.96 | 0.88 | 0.87 |
| 1 | 0.95 | 0.79 | 0.86 |
| Ensemble Neural Network | 0 | 0.75 | 0.82 | 0.78 | 0.77 |
| 1 | 0.8 | 0.73 | 0.76 |

A blue squares with white text

Description automatically generated A blue squares with white text

Description automatically generated

Fig 15: Confusion matrix -LR Fig 16: Confusion matrix -DT

A blue squares with white text

Description automatically generated A blue squares with white text

Description automatically generated

Fig 17: Confusion matrix - KNN Fig 18: Confusion matrix -RF

A blue squares with white text

Description automatically generated A blue squares with white text

Description automatically generated

Fig 19: Confusion matrix -SVM Fig 20: Confusion matrix -XgBoost

A black text on a white background

Description automatically generated A close-up of numbers

Description automatically generated

Fig 21: Confusion Matrix for RNN Fig 22: Confusion Matrix for FNN

A black text on a white background

Description automatically generated A graph with numbers and text

Description automatically generated

Fig 22: Confusion Matrix for EnsembleNN Fig 23: Accuracy across different models

* Logistic Regression (LR):It shows decent performance for both classes, but the Feedforward Neural Network (FNN) seems to outperform LR, especially for diabetic (Class 1) classification.
* Decision Tree (DT):Like LR, it provides reasonable results, but FNN, Random Forest and other ensemble methods might offer better overall performance.
* K-Nearest Neighbors (KNN):KNN exhibits good recall for non-diabetic instances (Class 0) but has lower precision and recall for diabetic instances. FNN may be a better choice for a more balanced performance.
* Random Forest (RF): RF shows strong performance, especially in diabetic classification (Class 1). It's a competitive option, and alongside SVM, it stands out among the traditional ML methods.
* Support Vector Machine (SVM):Similar to RF, SVM performs well across the board. It's a robust choice for this classification task, with high precision and recall for both classes.
* XGBoost: While XGBoost demonstrates good recall for non-diabetic instances, its overall performance might be slightly lower compared to FNN, RF, and SVM.
* Recurrent Neural Network (RNN): RNN performs well, especially in diabetic classification, but FNN seems to have a slight edge in terms of overall metrics.
* Feedforward Neural Network (FNN): FNN stands out as the top performer, with the highest precision, recall, F1 Score, and accuracy for diabetic classification. It's a strong candidate for this task.
* Ensemble Neural Network (ENN): ENN has lower overall metrics compared to FNN, indicating that the individual models or their combination might not be as effective as FNN for this specific task.

A graph of a graph

Description automatically generated

Fig 25: Variable Importance plot for Random Forest

## Discussion:

Best Performer: Feedforward Neural Network (FNN) stands out as the top performer, offering a good balance between precision and recall for both diabetic and non-diabetic instances which gives 87% accuracy.

Competitive Options: Random Forest (RF) and Support Vector Machine (SVM) also demonstrate strong performance and can be considered as alternatives, depending on specific requirements and constraints [Fig 24].

As per the variable importance plot of Random Forest we can draw the insight that physical health, mental, health, genital health followed by difficulty in walking are crucial predictors as they play an important role determining the metabolic disorder of Diabetes is associated with every sort of health-related issues [Fig 25]. Also, BMI is an important variable for the prediction which infers to the fact that obesity and weight management is crucial to restrict diabetes across the common m**ass.**