Diabetes Prediction with Machine Learning

Background

Diabetes is a significant health concern worldwide, disproportionately affecting women of color. According to the Centers for Disease Control and Prevention (CDC), among Hispanic, African American, and American Indian/Alaska Native women, diabetes is the third leading cause of death, following heart disease and cancer. For white women, diabetes ranks as the fourth leading cause of death. This alarming trend highlights the disparities in diabetes prevalence and outcomes among different racial and ethnic groups (*CDC*).

Beyond the human toll, diabetes also imposes a substantial economic burden. The International Diabetes Federation (IDF) estimates that the direct medical costs and lost productivity due to illness and death related to diabetes exceed $245 billion annually. This figure underscores the need for effective prevention and management strategies to mitigate the impact of diabetes on both individuals and the global economy (*IDF*).

Methods

Data Source

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. It can be downloaded from (*Diabetes Dataset*, 2020)

Features

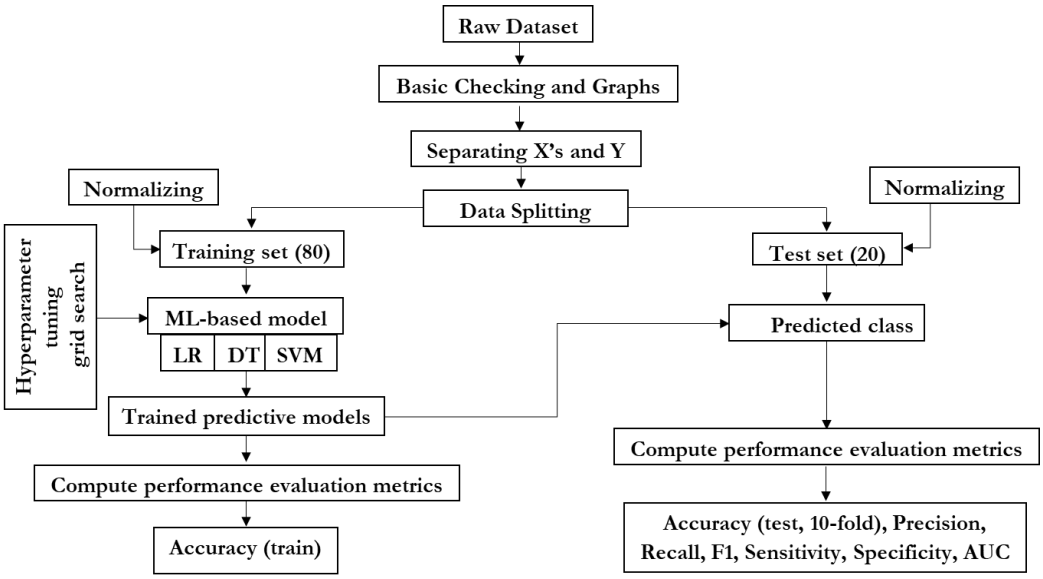
The dependent feature is whether the patient have diabetes (1) or not (0).

Independent features are:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
- Blood Pressure: Diastolic blood pressure (mm Hg)
- Skin Thickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)

- BMI: Body mass index (weight in kg/ (height in m)^2)
- Diabetes Pedigree Function: Diabetes pedigree function
- Age: Age (years)

Conceptual Framework



Result

Table 1: Evaluation Matrices for different classifiers.

Classifiers	Accuracy			AUC	Prec		Rec		F1		Sen	Spe
	Training	*Test*	*10-Fold*		*No*	*Yes*	*No*	*Yes*	*No*	*Yes*		
	Imbalanced Data											
LR	76.71	**75.97**	76.00	74.04	82.00	66.00	81.00	67.00	81.00	67.00	67.27	80.80
SVM	**76.87**	75.32	**77.00**	73.13	81.00	65.00	81.00	65.00	81.00	65.00	65.45	80.80
DT	74.26	71.42	73.00	62.02	71.00	76.00	85.00	29.00	91.00	42.00	29.09	94.94
	Balanced Data											
LR	76.55	71.42	77.00	72.25	84.00	58.00	69.00	76.00	69.00	66.00	76.36	68.68
SVM	**81.67**	68.18	**79.00**	67.97	79.00	54.00	69.00	67.00	74.00	60.00	67.27	68.68
DT	75.93	**74.02**	72.00	76.16	88.00	60.00	69.00	84.00	77.00	70.00	83.63	68.68

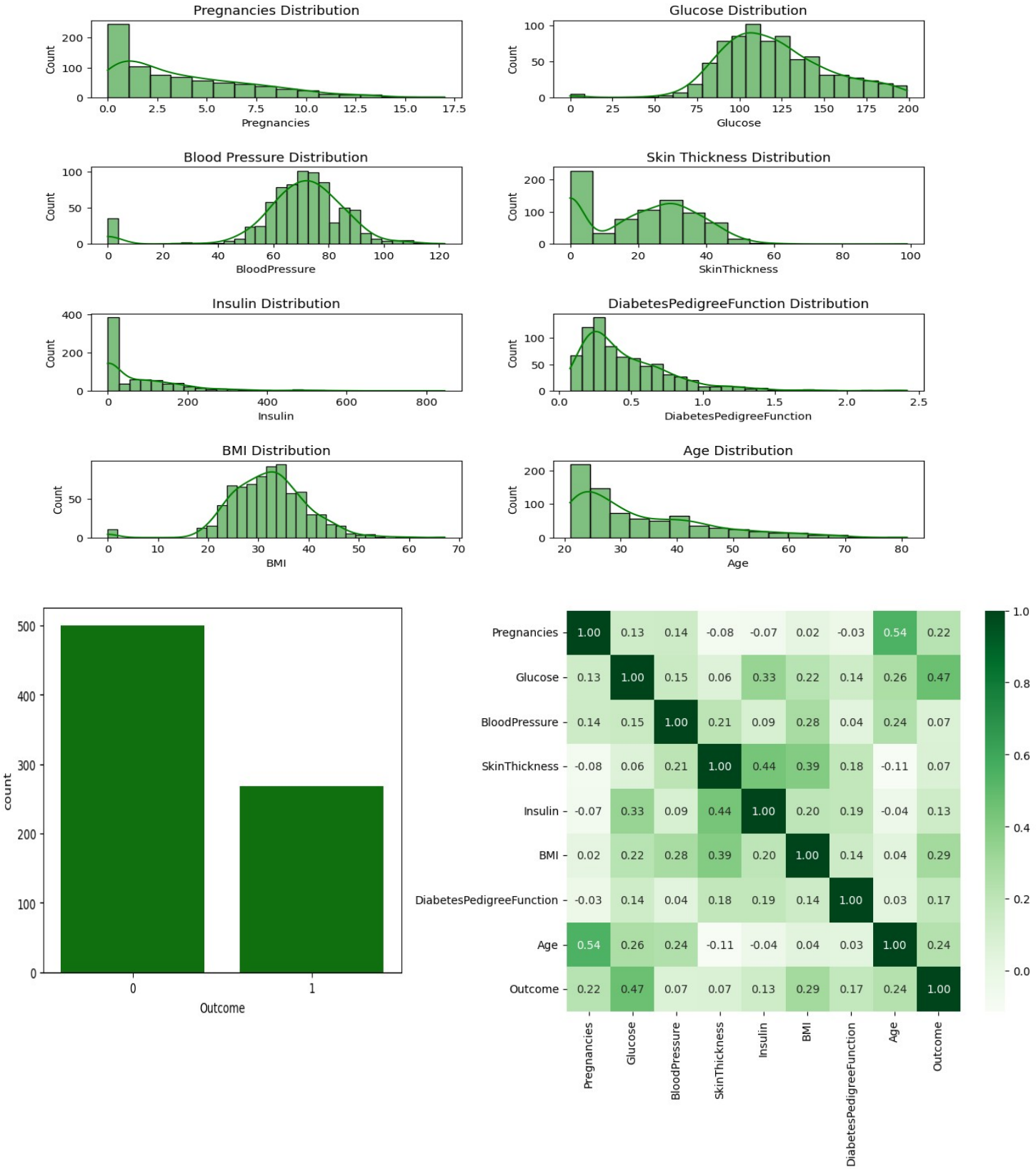


Figure 1: Baseline Graphs

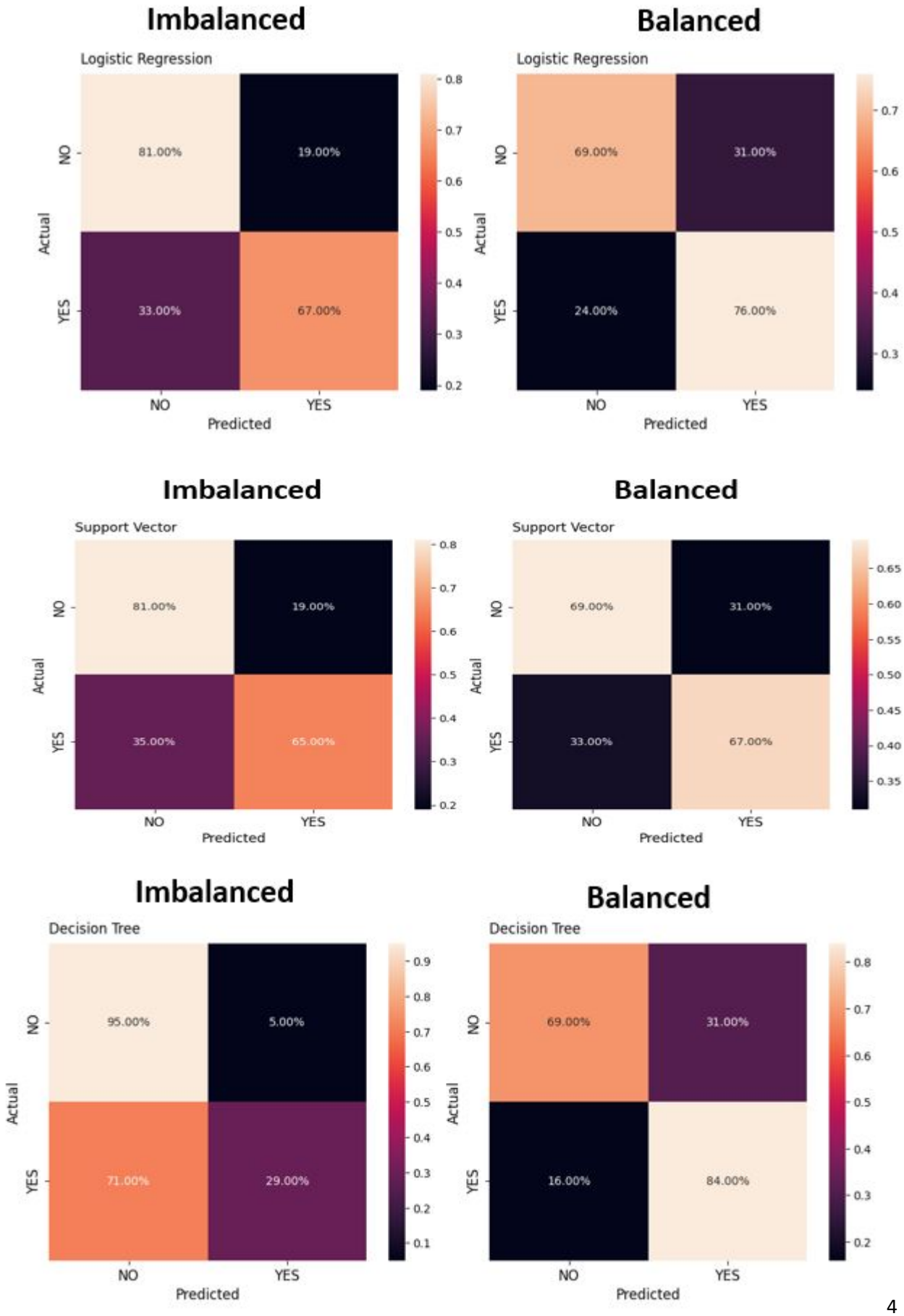


Figure 2: Confusion Matrix

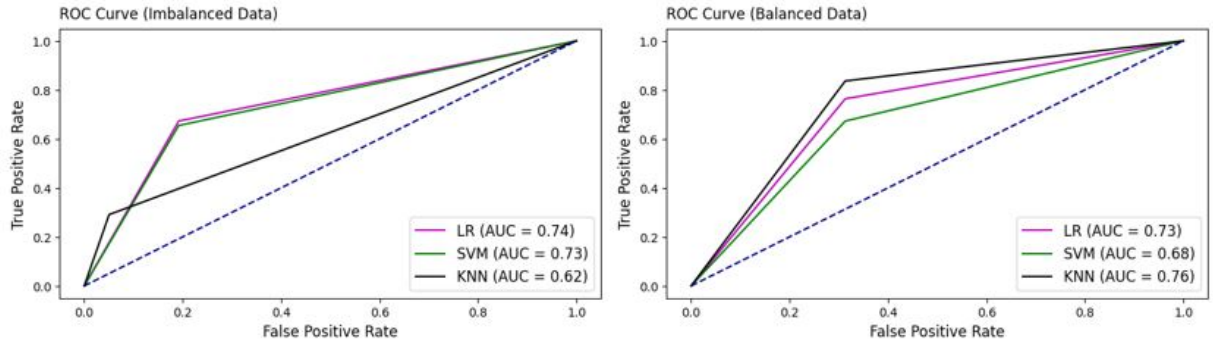


Figure 3: ROC Curve

Conclusion

In conclusion, the Logistic Regression performed better (acc:76%) for imbalanced data and Decision Tree performed better (acc:74%) for balanced data for test set. On the other hand, Support Vector Machine performed better for both imbalanced (77%) and balanced (79%) data. The balanced set have the better confusion matrix and AUC.

Reference

Centers for Disease Control and Prevention (CDC): https://www.cdc.gov/diabetes/php/data-research/index.html

Diabetes Dataset. (2020, August 5). Kaggle. https://www.kaggle.com/datasets/mathchi/diabetes-data-set

International Diabetes Federation (IDF): https://diabetesatlas.org/

Abbreviation

- **LR:** Logistic Regression
- **SVM:** Support Vector Machine
- **DT:** Decision Tree
- **AUC:** Area Under the Curve
- **Prec:** Precision
- **Rec:** Recall Score
- **F1:** F1-Score
- **Sen:** Sensitivity
- **Spe:** Specificity
- **ROC:** Receiver-Operating Characteristic Curve