WICHITA STATE UNIVERSITY

IME780 AN – Final Project

Deepansh Arora

May 03, 2021

About the Dataset

- Pima Indians Diabetes
- Diabetes is a chronic disease with potential to damage the essential organs.
- There are 768 records in the dataset
- Data was collected from the female patients
- All the patients were of Pima Indian heritage.

Table: Variable's description

Number	Variable	Description	Data Type
1	Pregnant	Number of times pregnant	Numeric
2	Glucose	Plasma glucose concentration (glucose tolerance test)	Numeric
3	Pressure	Diastolic blood pressure (mm Hg)	Numeric
4	Triceps	Triceps skin fold thickness (mm)	Numeric
5	Insulin	2-Hour serum insulin (mu U/ml)	Numeric
6	Mass	Body mass index (weight in kg/(height in m)\^2)	Numeric
7	Pedigree	Diabetes pedigree function	Numeric
8	Age	Age (years)	Numeric
9	Diabetes	Class variable (test for diabetes)	Categorical

columns	missing_values_count
Pregnant	0
Glucose	0
Pressure	0
Triceps	0
Insulin	0
Mass	0
Pedigree	0
Age	0
Diabetes	0

Fig: Missing Value Count

Fig: Project Flow

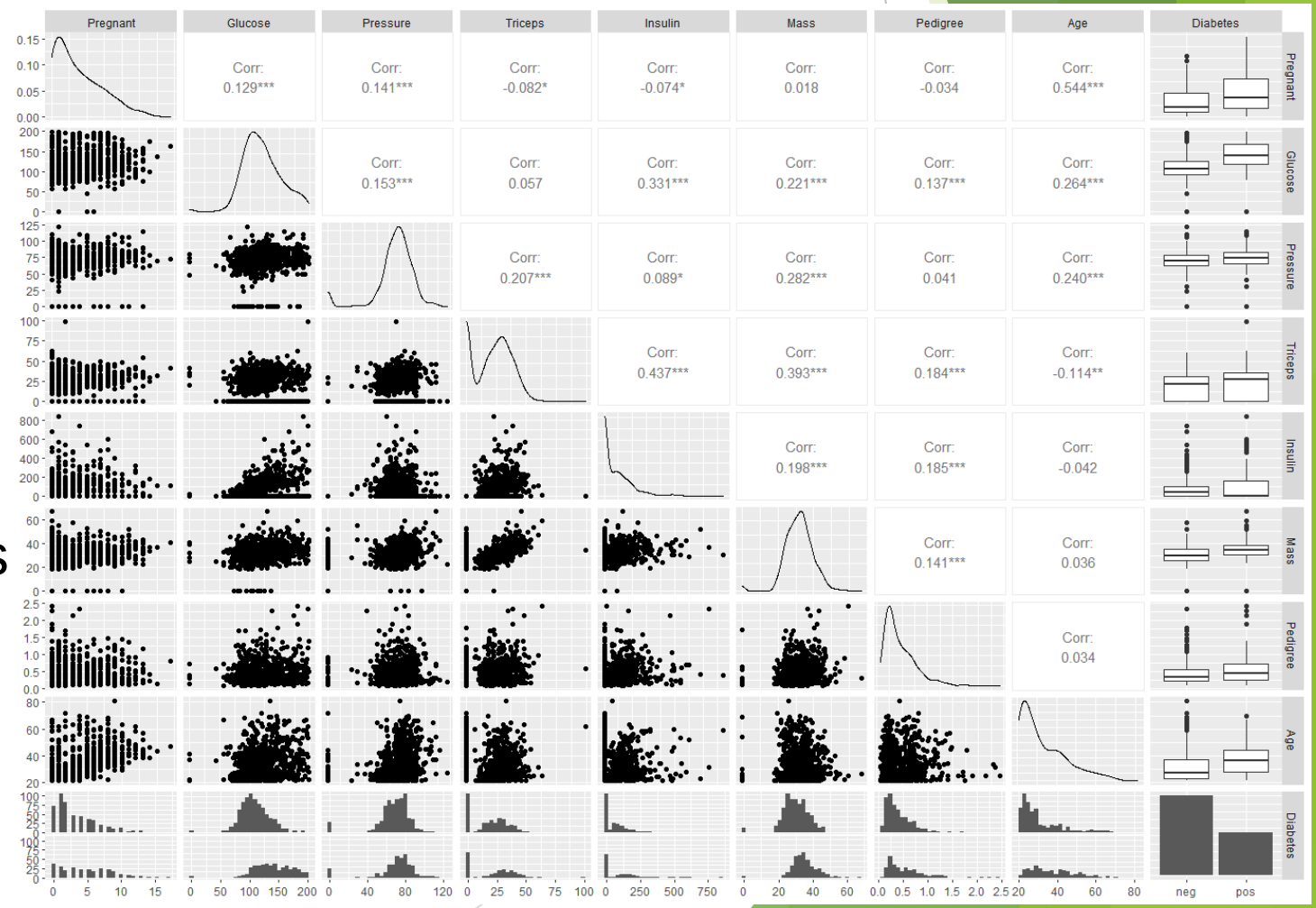
Exploratory Data Analysis

	freq	percentage
neg	500	65.10417
pos	268	34.89583

Fig: Class Distribution

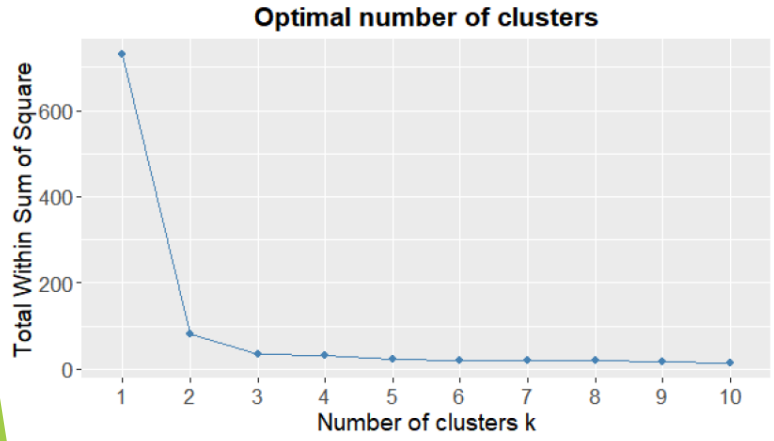
- Imbalanced Dataset
- No strong correlation
- Number of Asterisks represents the level of significance
- Most variables are positively skewed

Fig: Scatter Matrix Plot



Unsupervised Machine Learning

Fig: Elbow Plot



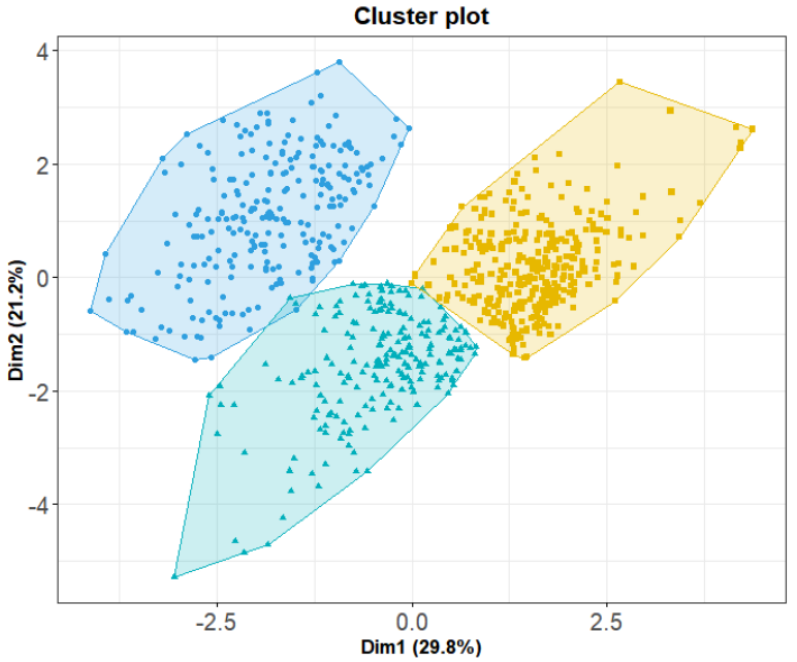


Fig: Visualizing the clusters using PCA

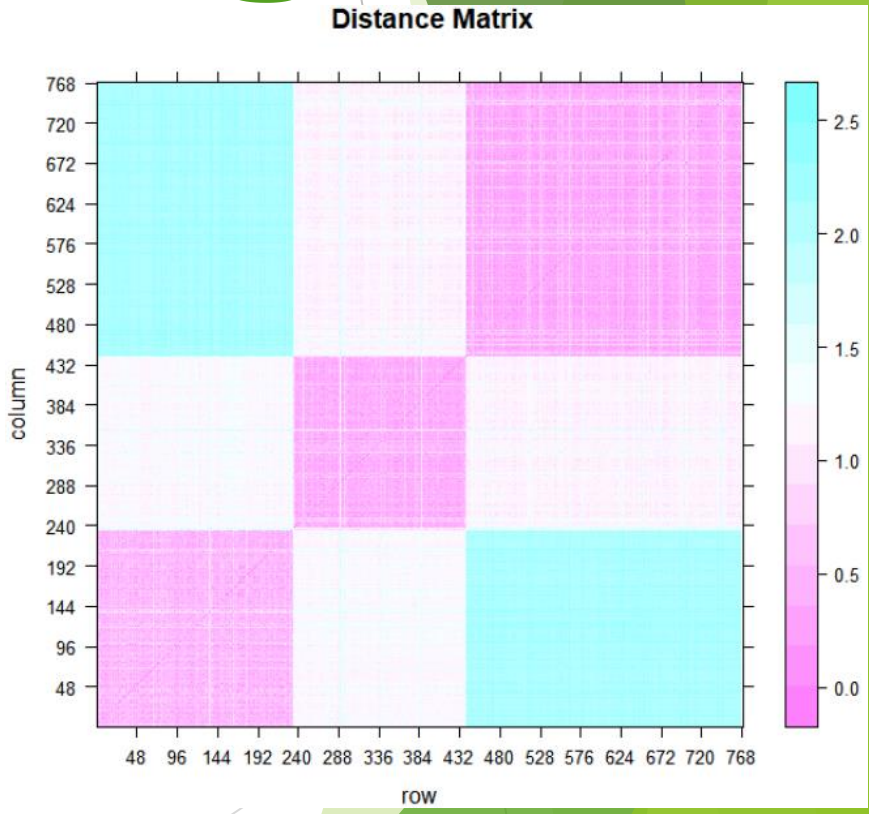


Fig: Distance Matrix

- K-means Clustering
- All numerical variables were chosen
- Clusters are cohesive in nature
- Clusters are of good quality

Supervised Machine Learning

- Data was randomly divided into training and testing subset
- 75% split ratio
- **Target Variable:** Diabetes
- **Predictor Variables:** Pregnant, Glucose, Pedigree, and Mass
- Output of Target variable can be either Positive or Negative
- Classification Machine Learning Models were implemented
- Models were then evaluated

Fig: Classification Machine Learning Model

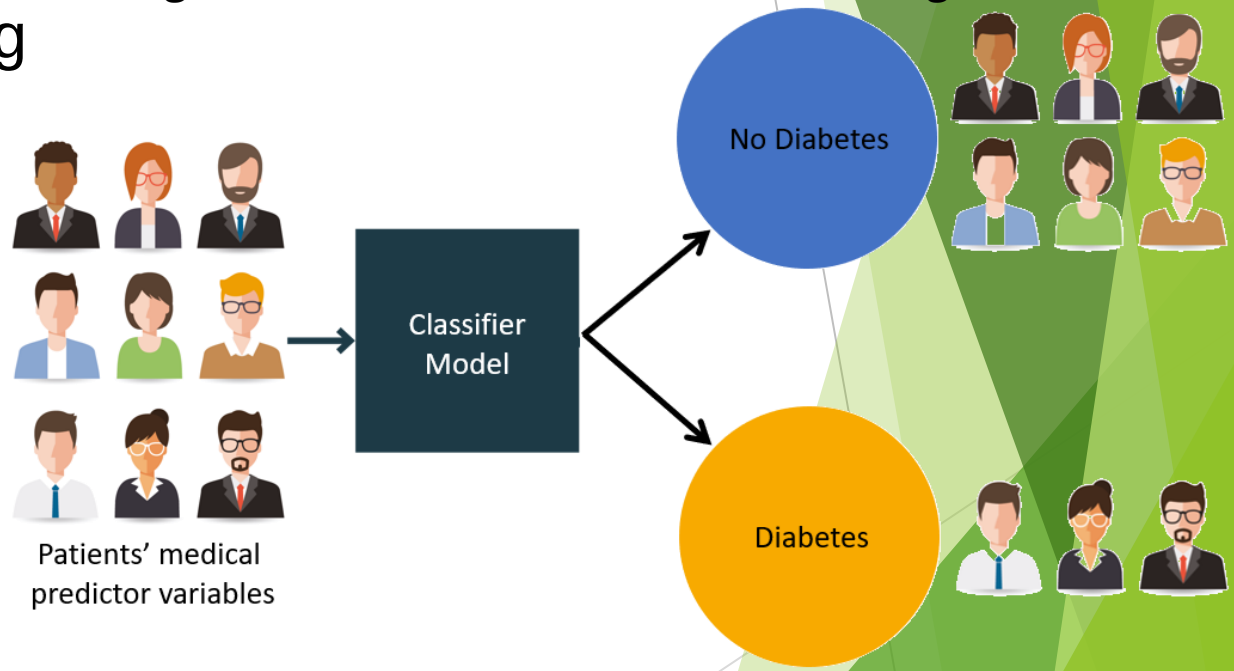


Fig source: https://towardsdatascience.com/building-a-machine-learning-classifier-model-for-diabetes-4fca624daed0

Logistic Regression

```
glm(formula = Diabetes ~ Pregnant + Glucose + Pedigree + Mass,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7574  -0.7076  -0.3675   0.6713   2.4909

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.516192   0.886670 -10.733  < 2e-16 ***
Pregnant     0.188030   0.034722   5.415 6.12e-08 ***
Glucose      0.037514   0.004434   8.461  < 2e-16 ***
Pedigree     1.147900   0.384413   2.986  0.00283 **
Mass         0.088841   0.017316   5.130 2.89e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 663.60  on 512  degrees of freedom
Residual deviance: 466.43  on 508  degrees of freedom
AIC: 476.43

Number of Fisher Scoring iterations: 5
```

Fig: Logistic Regression Output

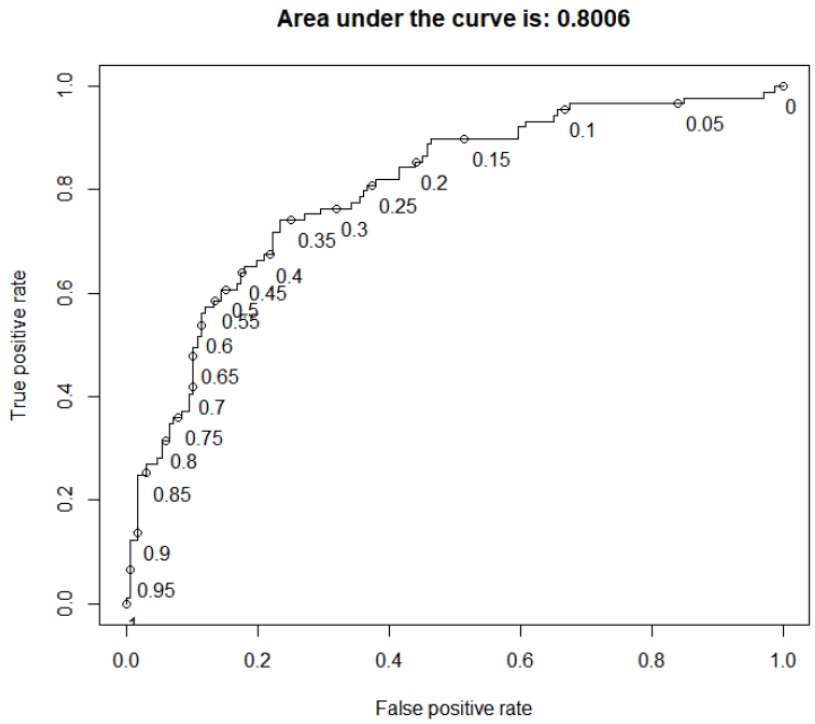


Fig: ROC Curve

```
Confusion Matrix and Statistics


df_class neg pos
     neg 113  21
     pos  53  68

               Accuracy : 0.7098
                 95% CI : (0.6499, 0.7647)
    No Information Rate : 0.651
    P-Value [Acc > NIR] : 0.0271240

                  Kappa : 0.4105

 Mcnemar's Test P-Value : 0.0003137

            Sensitivity : 0.6807
            Specificity : 0.7640
         Pos Pred Value : 0.8433
         Neg Pred Value : 0.5620
             Prevalence : 0.6510
         Detection Rate : 0.4431
   Detection Prevalence : 0.5255
      Balanced Accuracy : 0.7224

       'Positive' Class : neg
```

Fig: Confusion Matrix

- Threshold Probability = 0.3
- All the predictor variables are significant
- **Log(odds)=−9.516+$Pregnant$∗0.188+$Glucose$∗0.0375+$Pedigree$∗1.147+$Mass$∗0.088**
- Model has high specificity rate!

Decision Trees

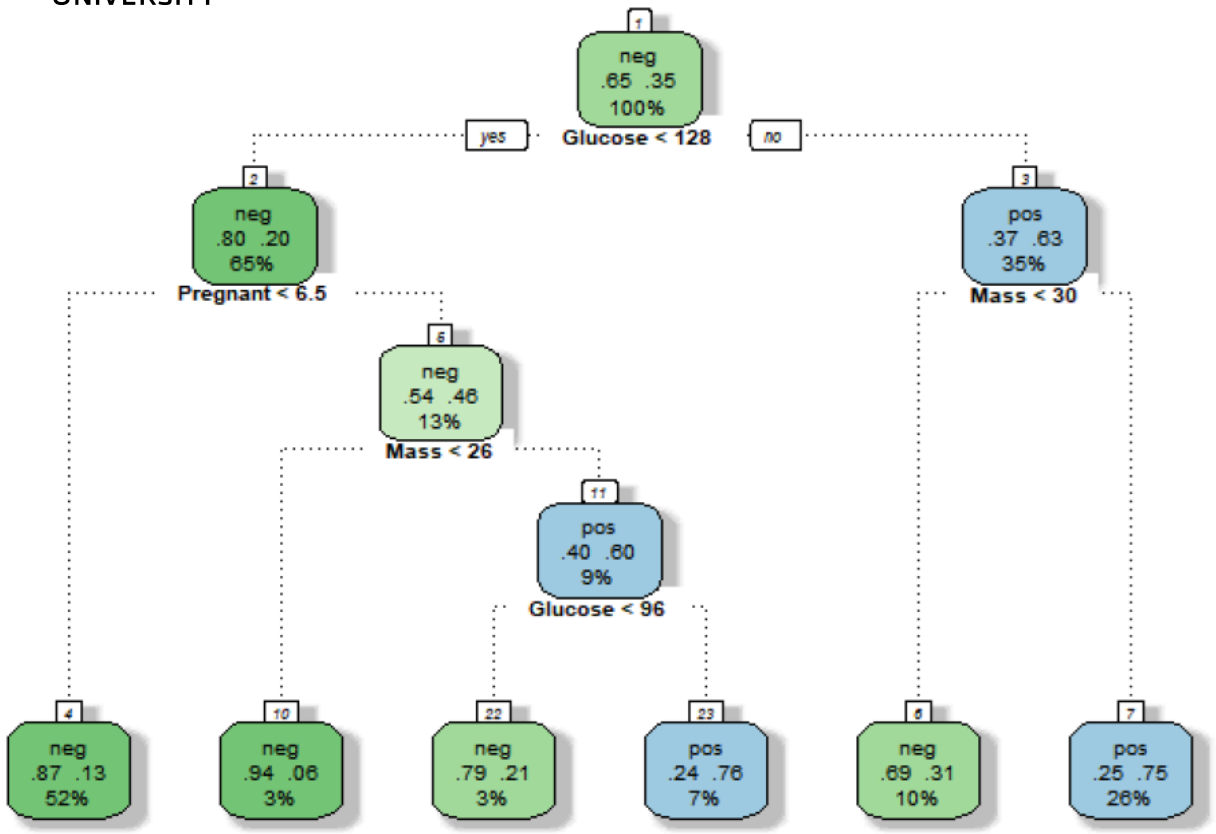


Fig: Decision Tree

```
Confusion Matrix and Statistics


pred  neg pos
  neg 131  34
  pos  35  55

               Accuracy : 0.7294
                 95% CI : (0.6705, 0.7829)
    No Information Rate : 0.651
    P-Value [Acc > NIR] : 0.004584

                  Kappa : 0.4061

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.6180
            Specificity : 0.7892
         Pos Pred Value : 0.6111
         Neg Pred Value : 0.7939
             Prevalence : 0.3490
         Detection Rate : 0.2157
   Detection Prevalence : 0.3529
      Balanced Accuracy : 0.7036

       'Positive' Class : pos
```

Fig: Confusion Matrix

- ❑ Complexity Parameter is 0.02
- ❑ Model has high accuracy!

WICHITA STATE UNIVERSITY

Conclusion

Parameters	Logistic Regression	Decision Tree
Accuracy	70.98%	72.94%
Kappa	0.4105	0.4061
Sensitivity	68.07%	78.92%
Specificity	76.40%	61.80%

Table: Model's summary

- Decision Trees has high accuracy rate
- Priority is to correctly classify the positive cases
- Kappa – level of agreement between true values and classification
- Logistic Regression Model has high value of Kappa and Specificity
- Logistic Regression is highly preferred for this dataset!!

References

- U.S. Department of Health and Human Services. (2020, January 3). National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). National Institutes of Health. https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-institute-diabetes-digestive-kidney-diseases-niddk.
- Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. Part 1: Diagnosis and Classification of Diabetes Mellitus (WHO/NCD/NCS/99.2). Geneva: World Health Organization; 2019.
- J. Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2017.
- Lantz, Brett. Machine Learning with R: Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R. Birmingham: Packt Publishing, 2015.