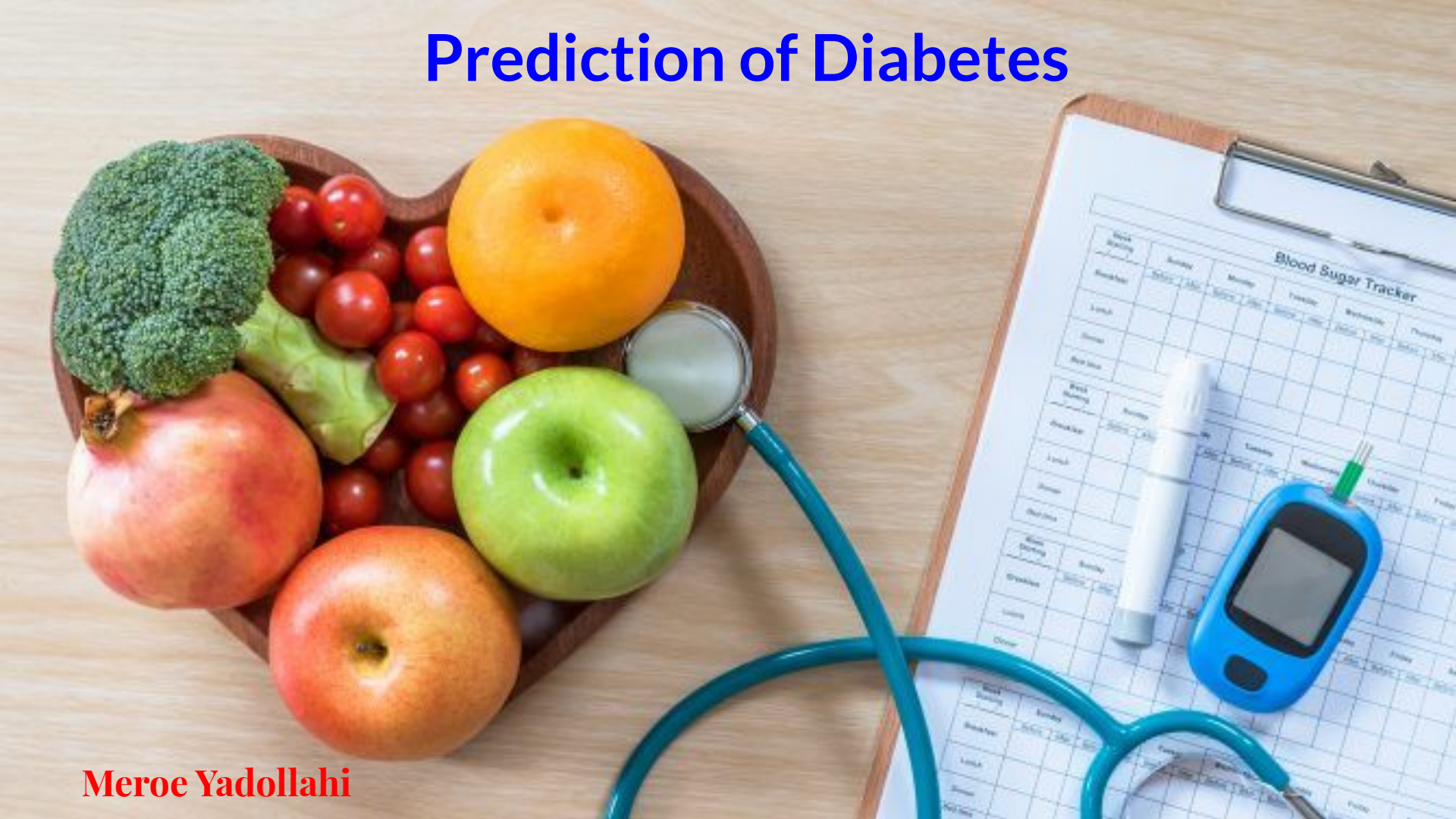
Prediction of Diabetes



Meroe Yadollahi

Introduction

- **Diabetes** is one of the deadliest diseases in the world. It is not only a disease but also creator of different kinds of diseases like heart attack, blindness etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports.
- Cause of Diabetes vary depending on the genetic makeup, family history, ethnicity, health etc.
- Diabetes & pre-diabetes is diagnosed by blood test.

Business Understanding

- **What is the profound question?**

  To identify whether the patient is having diabetes or not?

- **Goal:**

  Goal of this project is to identify the probability of diabetes in patients using data mining techniques.

- **Advantage of this project:**

  The rules derived will be helpful for doctors to identify patients suffering from diabetes. Further predicting the disease early leads to treating the patient before it becomes critical.

Data Understanding

Data Source and Details:

- This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.
- This dataset has 768 samples of diabetic and healthy individuals.
- In particular, all patients here are females of at least 21 years of age.
- The diabetes dataset is credited to UCI machine learning database repository.
- The dataset has total 9 attributes out of which 8 are independent variables and one is the dependent variable i.e. target variable which determines whether patient is having diabetes or not.

Data Understanding

Attributes Details:

- **Pregnancies:** No. of times pregnant
- **Glucose:** Plasma Glucose Concentration a 2 hour in an oral glucose tolerance test (mg/dl)

Plasma Glucose Test	Normal	Prediabetes	Diabetes
2 hour post-prandial	Below 140 mg/dl	140 to 199 mg/dl	200 mg/dl or more

A 2-hour value between 140 and 200 mg/dL is called impaired glucose tolerance. This is called "pre-diabetes." It means you are at increased risk of developing diabetes over time. A glucose level of 200 mg/dL or higher is used to diagnose diabetes.

Data Understanding

- **Blood Pressure:** Diastolic Blood Pressure(mmHg)

  If Diastolic B.P > 90 means High B.P (High Probability of Diabetes)
  Diastolic B.P < 60 means low B.P (Less Probability of Diabetes)

- **Skin Thickness:** Triceps Skin Fold Thickness (mm) –
  A value used to estimate body fat. Normal Triceps Skinfold Thickness in women is 23mm. Higher thickness leads to obesity and chances of diabetes increases.

- **Insulin:** 2-Hour Serum Insulin (mu U/ml).

Feature	Normal Insulin Level
2 Hours After Glucose	16-166 mIU/L

Values above this range can be alarming.

Data Understanding

- **BMI:** Body Mass Index (weight in kg/ height in m_2)
  - ❖ Body Mass Index of 18.5 to 25 is within the normal range
  - ❖ BMI between 25 and 30 then it falls within the overweight range. A BMI of 30 or over falls within the obese range.
- **DiabetesPedigreeFunction:**
  It provides information about diabetes history in relatives and genetic relationship of those relatives with patients. Higher Pedigree Function means patient is more likely to have diabetes.
- **Age** (in years)
- **Outcome:** Class Variable (0 or 1) where '0' denotes patient is not having diabetes and '1' denotes patient having diabetes.The **dependent variable** is whether the patient is having diabetes or not.

Data Preparation

- Data preparation stage includes data cleaning and transforming data if needed.
- Various things have to be taken into consideration for data cleaning like:
  - ❖ Handling Zero/Null Values – The zeroes shown in the table are not zeroes but null values . We have deduced this based upon our inference that certain attributes like skin thickness, insulin, BMI etc cannot be zero.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0

The dataset had a lot of zero values.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	169.5	33.6	0.627	50	1
1	1	85.0	66.0	29.0	102.5	26.6	0.351	31	0
2	8	183.0	64.0	32.0	169.5	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1
5	5	116.0	74.0	27.0	102.5	25.6	0.201	30	0
6	3	78.0	50.0	32.0	88.0	31.0	0.248	26	1
7	10	115.0	70.0	27.0	102.5	35.3	0.134	29	0
8	2	197.0	70.0	45.0	543.0	30.5	0.158	53	1
9	8	125.0	96.0	32.0	169.5	34.3	0.232	54	1
10	4	110.0	92.0	27.0	102.5	37.6	0.191	30	0

The zero values have been replaced by the median of that column.

Data Preparation

- **Select appropriate attributes for analysis:**
  As all these attributes affect diabetes so first I decided to keep all the independent variables for data mining process.

- **Data Splitting:**
  Data was divided into training and testing data into 75:25 ratio. 75% was training data and 25% was testing data.

Modeling

- As we have to classify the data into patients having diabetes or not, we used K-Nearest Neighbor(KNN), Support Vector Machine (Radial basis function kernel and linear kernel), Logistic Regression, and Decision Tree algorithms. All these four algorithms are good for classifying dependent variables based upon categorized independent variables.
- Here in this presentation, I compare the two algorithms, Decision Tree and KNN to find the one which gives the best results based upon overall accuracy and precision.

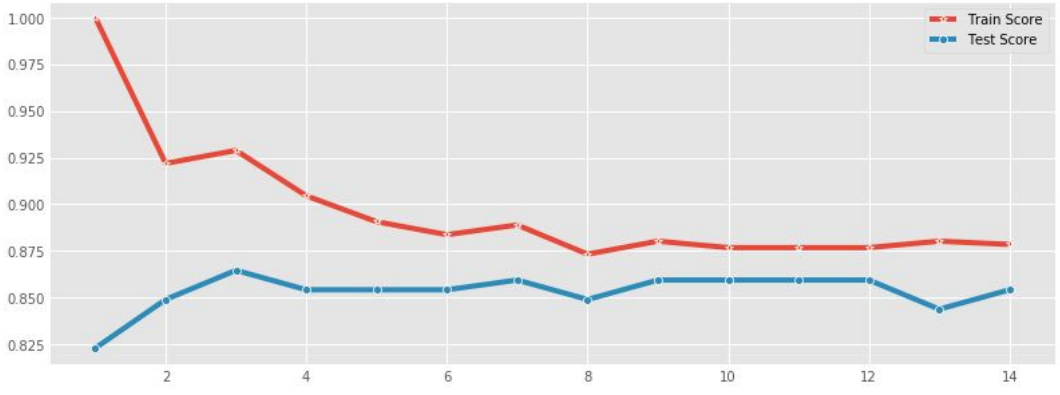
Evaluation

K_Nearest Neighbors:

The best Result is captured at K = 3, hence 3 is used for the final model.

Max Train Score	Max Test Score
100%	86.46%
K = 1	K = 3

Result Visualization:

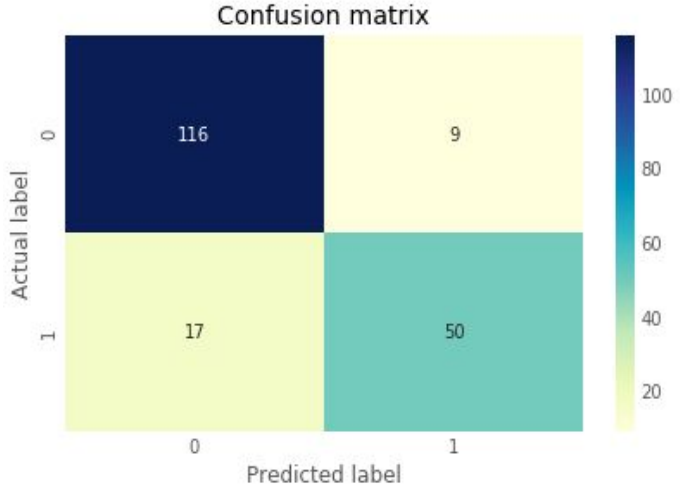


Evaluation

Confusion Matrix for the K_Nearest Neighbors:

	Predicted Yes	Predicted No
Actual Yes	tp (true positive)	fp (false positive)
Actual No	fn (false negative)	tn (true negative)

- true positives (TP): These are cases in which we correctly predicted diabetes as result.
- true negatives (TN): We correctly predicted no diabetes and they don't have the disease.
- false positives (FP): We correctly predicted no diabetes, but they actually had the disease. (Also known as a "Type I error.")
- false negatives (FN): We correctly predicted diabetes, but they actually had no disease. (Also known as a "Type II error.")

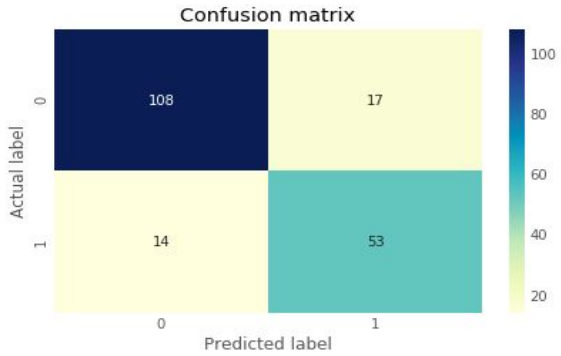


Out of 192 Testing Dataset

Alternate Model Comparison

Decision Tree VS KNN:

Confusion Matrix Decision Tree



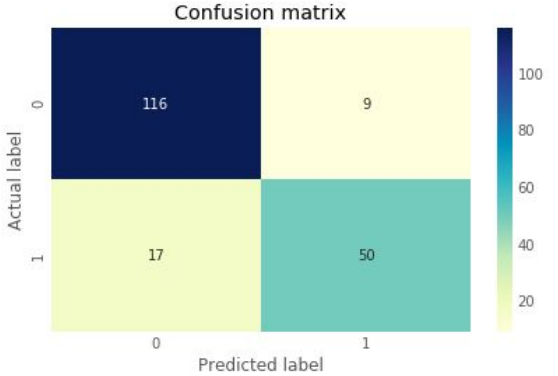
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn} \quad \text{=(sensitivity)}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

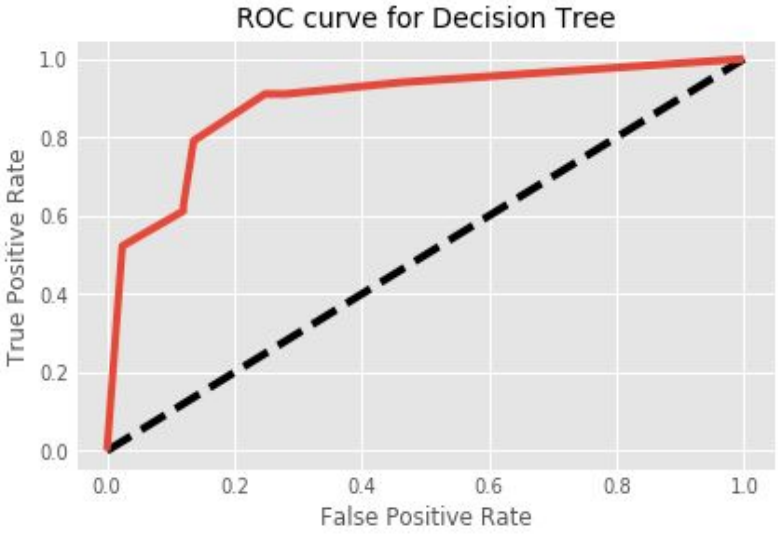
	precision	recall	f1-score	support
0	0.89	0.86	0.87	125
1	0.76	0.79	0.77	67
accuracy			0.84	192
macro avg	0.82	0.83	0.82	192
weighted avg	0.84	0.84	0.84	192

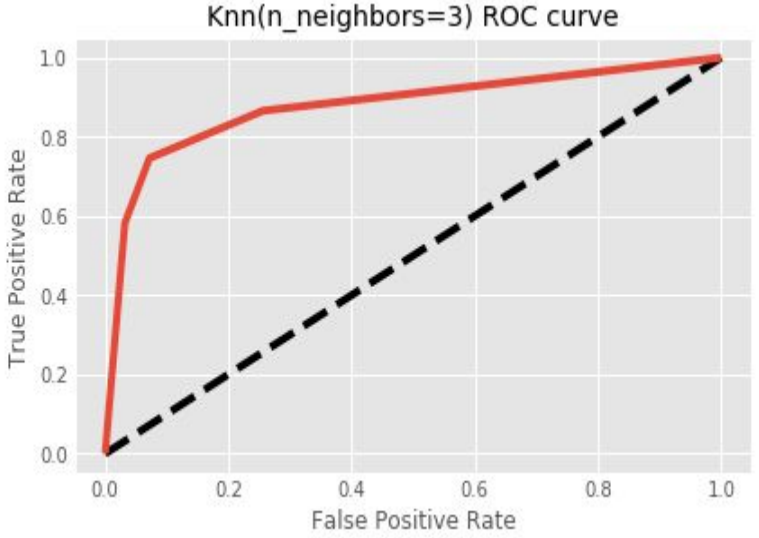
Confusion Matrix KNN



	precision	recall	f1-score	support
0	0.87	0.93	0.90	125
1	0.85	0.75	0.79	67
accuracy			0.86	192
macro avg	0.86	0.84	0.85	192
weighted avg	0.86	0.86	0.86	192

Alternate Model Comparison





The Area Under Curve(AUC) for Decision Tree is 0.89.

The Area Under Curve(AUC) for KNN is 0.88.

So by comparing both the results we infer that K_Nearest Neighbor Model is showing better results in comparison to Decision Tree Model.

Models Accuracy Before Feature Engineering

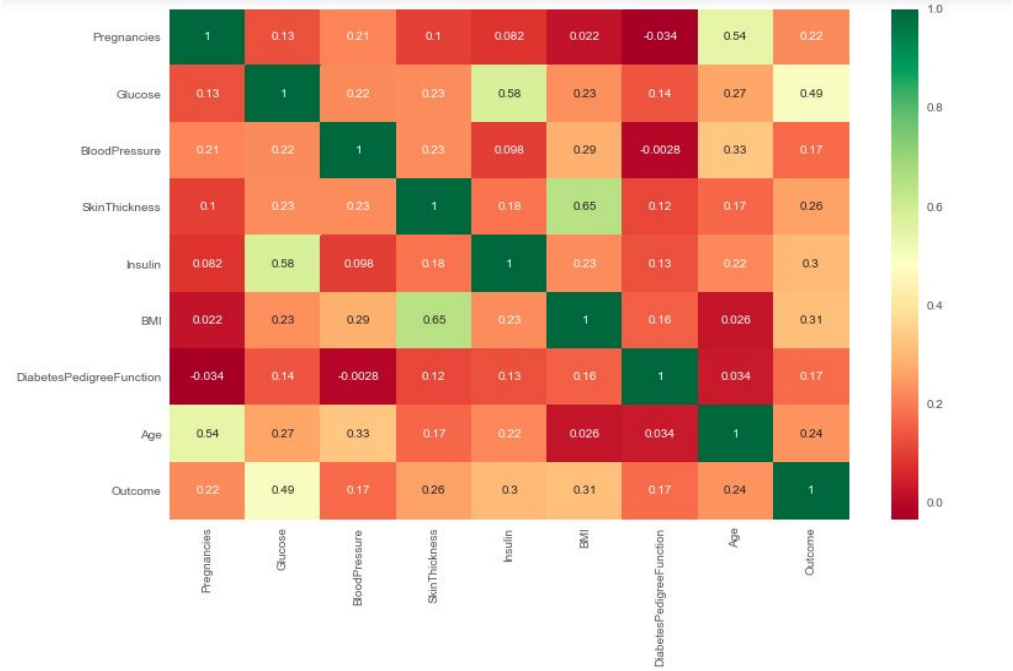
	Accuracy
Linear Svm	0.755208
Radial Svm	0.651042
Logistic Regression	0.713542
Decision Tree	0.828125
KNN	0.864583

Feature Engineering

1. Many features can affect the accuracy of the algorithm.
2. Feature Extraction means to select only the important features in-order to improve the accuracy of the algorithm.
3. It reduces training time and reduces overfitting
4. We can choose important features in 2 ways:
    a. Correlation matrix --> selecting only the uncorrelated features.
    b. RandomForestClassifier --> It gives the importance of the features

Feature Engineering

Correlation Matrix



All the features look to be uncorrelated. So we cannot eliminate any features just by looking at the correlation matrix.

Random Forest Classifier

```
Insulin                    0.363715
SkinThickness              0.146408
Glucose                    0.140869
Age                        0.092266
BMI                        0.090501
DiabetesPedigreeFunction   0.070971
Pregnancies                0.051046
BloodPressure              0.044224
dtype: float64
```

The important features are:

Insulin, SkinThickness, Glucose, Age, BMI

Comparing the Models Accuracy Before and After Feature Engineering

	New Accuracy	Accuracy	Increase
Linear Svm	0.791667	0.755208	0.036458
Radial Svm	0.822917	0.651042	0.171875
Logistic Regression	0.770833	0.713542	0.057292
Decision Tree	0.859375	0.828125	0.031250
KNN	0.807292	0.864583	-0.057292

Our Best Model:

KNN

Accuracy: 86.45 %

Including all Features

The above dataframe shows the new accuracy of the models after feature selection. We can see that the Accuarcy for Radial Svm increases by 17% and for Logistic Regression and Decision Tree the Accuracy increases around 6% and 3%. For KNN the accuracy decreases about 6%, but it is still the best model.

Cross Validation

Many times, the data is imbalanced, i.e there may be a high number of class1 instances but less number of other class instances. Thus we should train and test our algorithm on each and every instance of the dataset. Then we can take an average of all the noted accuracies over the dataset.

k=10, split the data into 10 equal parts

	CV Mean
Linear Svm	0.791667
Radial Svm	0.822917
Logistic Regression	0.770833
Decision Tree	0.859375
KNN	0.807292

Deployment

Deployment includes three important task :

- Plan Deployment – Planning basically includes the strategy to be formulated for implementing the model in real world. This model can now be used in medical organizations for easy and early detection of diabetes in patients.
- Monitor Deployment – In this, continuous monitoring of model takes place. Regular check is done to ensure model is working fine and if any error occurs can be easily detected.
- Generate Reports – Final statistical reports are generated which summarizes the overall performance of the model.

Conclusion

- The K_Nearest Neighbor Model achieved 86% accuracy.
- Different options were taken into consideration to improve the accuracy. So finally after cleaning the data, and using all the features and K = 3 for KNN we were able to achieve the desired accuracy.
- We even compared our K-Nearest Model with Decision Tree Model and SVM Model inferred that the K_Nearest Neighbor Model is the best amongst all.

References

- Google Search to get some information about Diabetes
- Slides and Lecture Notes
- UCI Machine Learning

Welcome to Diabetes Prediction Portal

Enter BMI(Body Mass Index): 33.6 Enter Glucose Level: 148.0 Enter Age: Age Enter Skin Thickness: SkinThickness Enter Insulin Level: Insulin Level

Submit

Note: *Body Mass Index (BMI) is a measure of body fat based on height and weight.Body Mass Index is a simple calculation using a person's height and weight. The formula is BMI = kg/m^2 where kg is a person's weight in kilograms and m^2 is their height in metres squared. A BMI of 25.0 or more is overweight, while the healthy range is 18.5 to 24.9. BMI applies to most adults 18-65 years.*

Important Notice: *It is for only Educational purpose.*

Your Result:

Negative!! You do not have diabetes

Thank You