|  |
| --- |
| predictive analysis using knn  (Pima Indians Diabetes Database)  21PD10(HARINI K V),21PD26(DHIKSHITHA R) |
|  |

# ABOUT DATA SET

|  |
| --- |
| Context This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. 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. |
| Content The datasets consists of several medical predictor variables and one target variable, Outcome(Class variable (0 or 1) 268 of 768 are 1, the others are 0). Predictor variables includes the number of pregnancies the patient has had, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), their BMI(Body mass index (weight in kg/(height in m)^2)), insulin level(2-Hour serum insulin (mu U/ml)), Diabetes pedigree function, age, and so on. |

# INSPIRATION

|  |
| --- |
| Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not |
|  |

# SOURCE

# 

# <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?datasetId=228&searchQuery=KNN>

# ACCURACY evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| TRAIN -TEST DATA SPLIT | K=5 | K=10 | K=15 |
| 80% - 20% | 0.79870 | 0.77922 | 0.80519 |
| 70% - 30% | 0.77056 | 0.75757 | 0.77489 |
| 60% - 40% | 0.75649 | 0.75324 | 0.75324 |

# SUMMARY

# 

* **Training accuracy** rises as model complexity increases
* **Testing accuracy** penalizes models that are too complex or not complex enough
* For KNN models, complexity is determined by the **value of K** (lower value = more complex)
* We could see that the accuracy is the highest when the K-Value is 15

# model on Breast Cancer Wisconsin (Diagnostic) Data Set

# About Dataset

# Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

# Attribute Information:

# ID number

# Diagnosis (M = malignant, B = benign)

# 3-32)

# Ten real-valued features are computed for each cell nucleus:

# radius (mean of distances from center to points on the perimeter)

# texture (standard deviation of gray-scale values)

# perimeter

# area

# smoothness (local variation in radius lengths)

# compactness (perimeter^2 / area - 1.0)

# concavity (severity of concave portions of the contour)

# concave points (number of concave portions of the contour)

# symmetry

# fractal dimension ("coastline approximation" - 1)

# The mean, standard error and "worst" or largest (mean of the three

# largest values) of these features were computed for each image,

# resulting in 30 features. For instance, field 3 is Mean Radius, field

# 13 is Radius SE, field 23 is Worst Radius.

# All feature values are recoded with four significant digits.

# source

# [https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?datasetId=228&searchQuery=KNN)

# ACCURACY evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| TRAIN -TEST DATA SPLIT | K=5 | K=10 | K=15 |
| 80% - 20% | 0.938596 | 0.964912 | 0.964912 |
| 70% - 30% | 0.935672 | 0.918128 | 0.935672 |
| 60% - 40% | 0.925438 | 0.929824 | 0.925438 |

# summary

# 

# We could see that the accuracy is the highest when the K-Value is 10