Divya Naidu

BREAST CANCER ANALYSIS
HEALTHCARE INNOVATION SUMMIT 2018

Agenda

01

Objective

02

Dataset

03

Approach

04

Modelling

05

Summary and
Next Steps

Objective

DIAGNOSIS

Mammography not enough. Need Invasive biopsy to probe further.

IMPORTANCE

One of the highest cause of death in women. Expensive Diagnosis.

GOAL

How can we effectively detect cancerous cells and reduce death rates?

Wisconsin Breast Cancer Diagnostic Dataset

- The breast cancer data includes **569 cases of cancer** biopsies, each with 32 features.

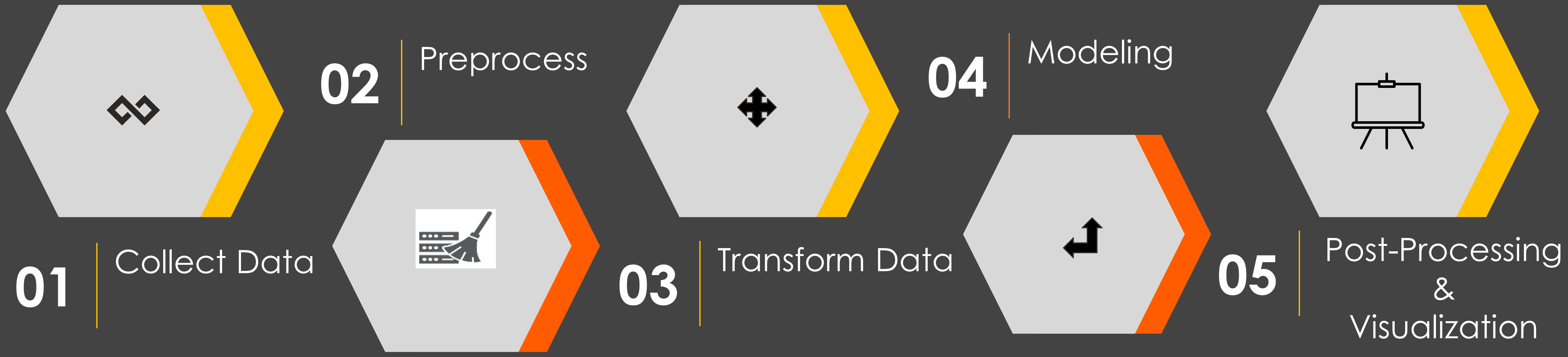
Patient ID

Cancer Diagnosis

30 numeric-valued laboratory measurements

- 30 numeric measurements comprise **the mean, standard error, and worst value** for 10 different characteristics of the digitized cell nuclei which are: **Radius, Texture, area**,etc.

Brief Solution Approach



01 Collect Data

02 Preprocess

03 Transform Data

04 Modeling

05 Post-Processing & Visualization

Preprocess & Transform

01

Missing Values

Checked for any missing values in the dataset

02

Data Type Consistency

Checked for data duplication, violation of data constraints etc.

03

Extreme Values

Dataset is checked for existence of any noisy data

04

Normalize Data

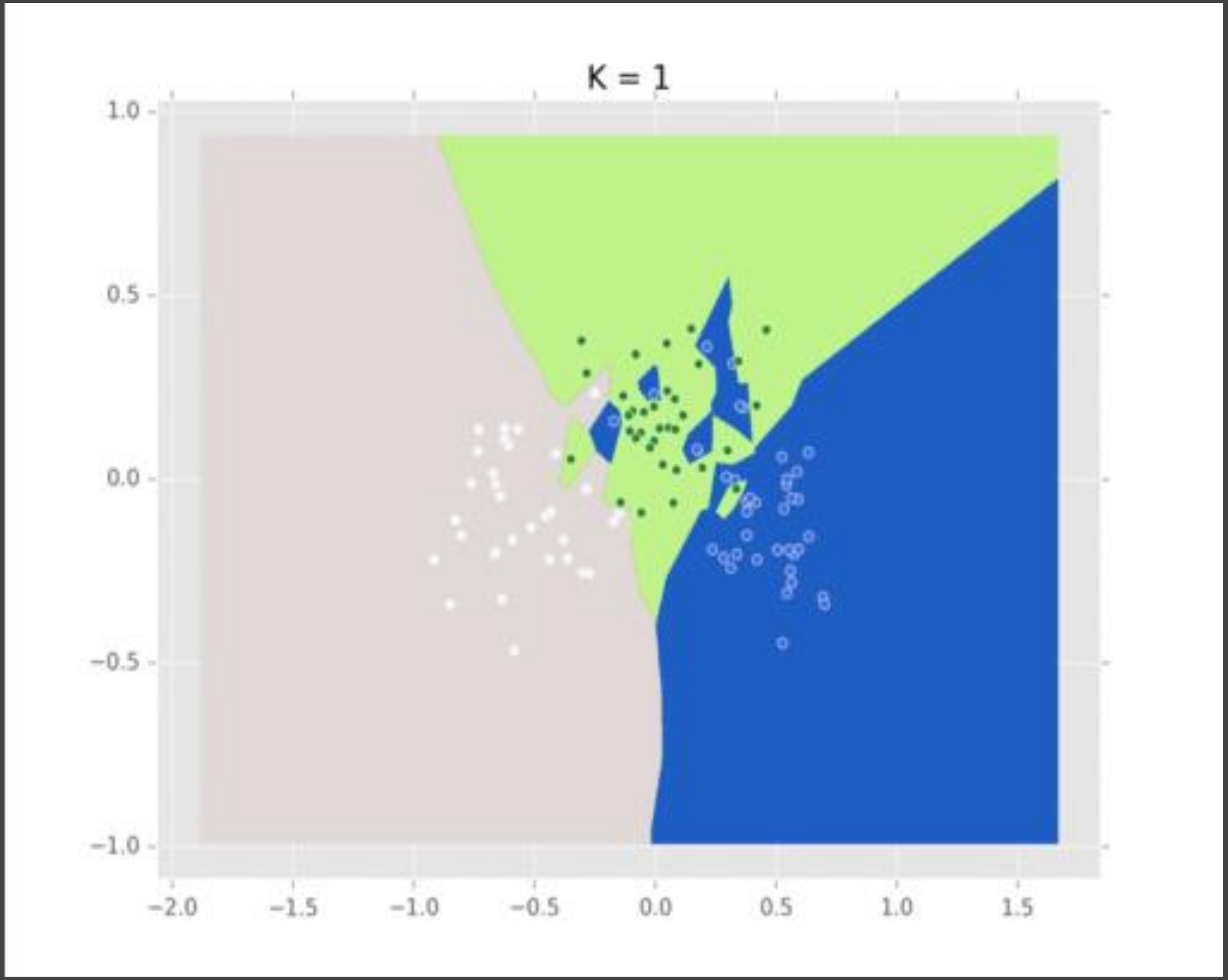
Since input variables are on different scale, it necessary to normalize data

05

Train & Test Data

Divided data to train and test 70% - 30% respectively

How kNN Algorithm works

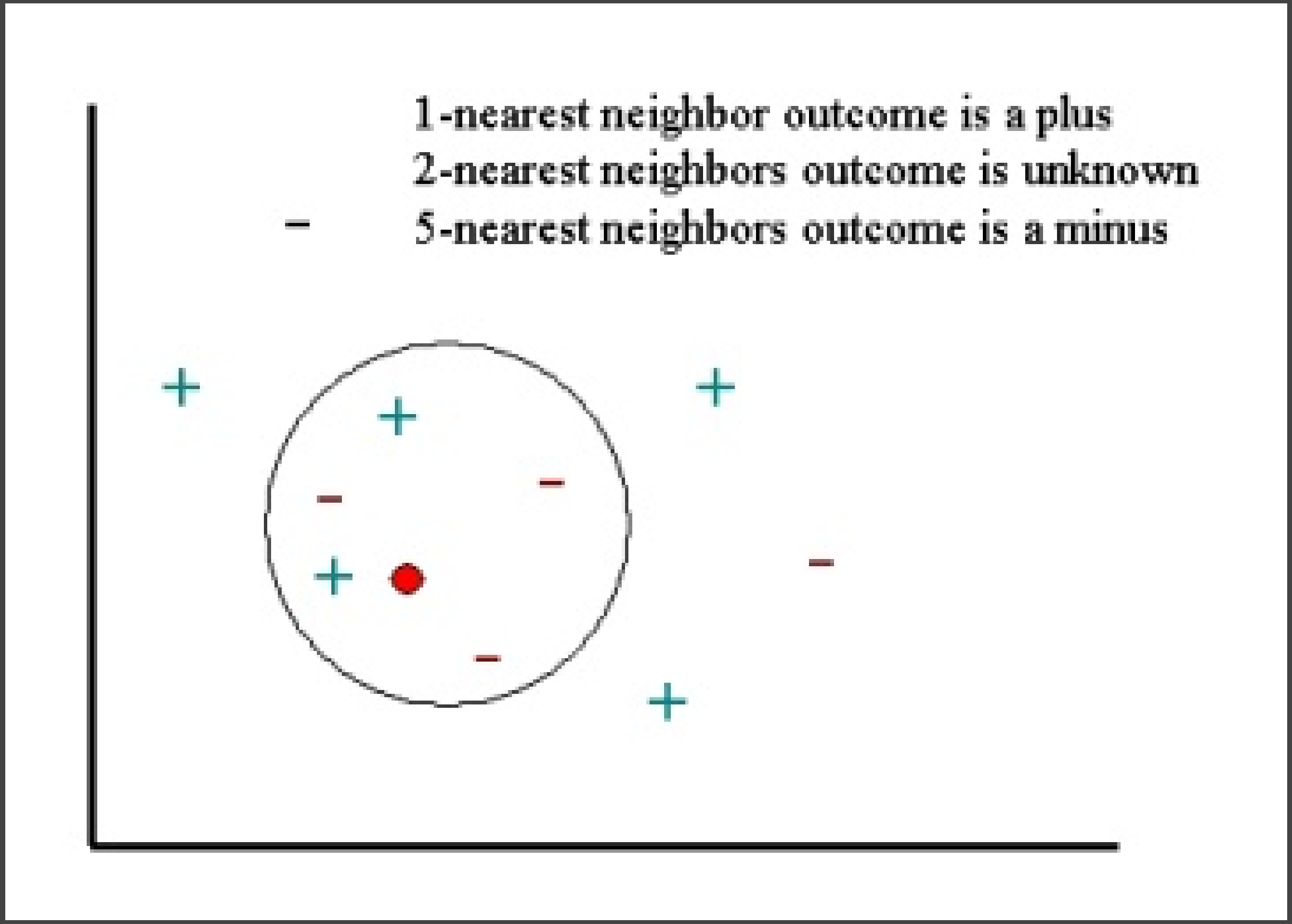


- kNN uses its previous observations to train its algorithm into classifying new observations (test dataset) by looking at already classified examples.

- Divided the data into train and test - 2/3rd and 1/3rd respectively.

- Monte Carlo Cross Validation - Used to avoid biases in dataset. Here, we randomly select fraction of data to form training set and then assign rest of the data to a test set. This is repeated multiple times.

How kNN Algorithm works


1-nearest neighbor outcome is a plus
2-nearest neighbors outcome is unknown
5-nearest neighbors outcome is a minus
+
−

- KNN Algorithm is based on feature similarity.

- Implemented kNN with Euclidean Distance formula to find k nearest neighbors.

- 97% accuracy with kNN on cancer dataset!

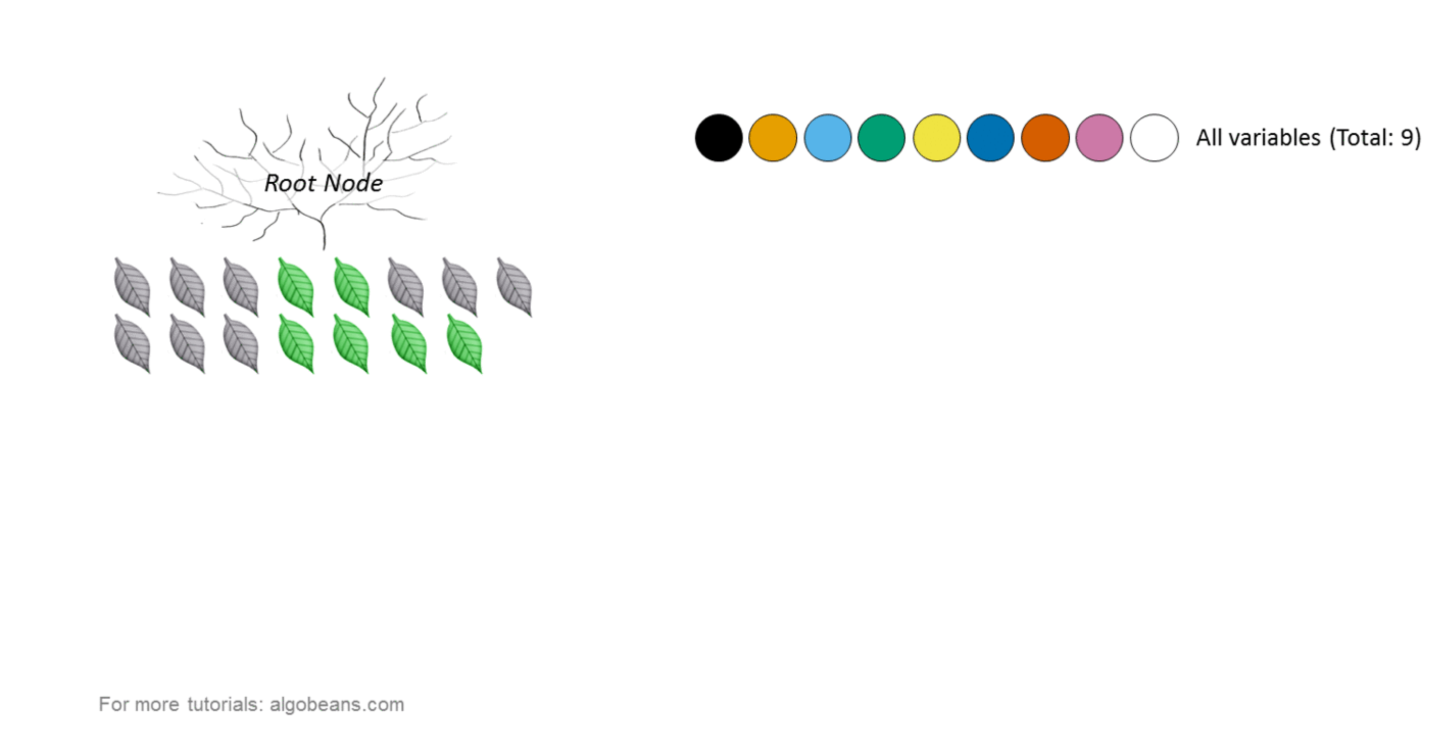
Optimization

Our Data Set uses 10 different usable variables to classify a patient's tumor as Benign or Malignant.

Three traits each for 10 measurement variables in the distance calculation introduces some redundancy. Due to this we may lose an opportunity to gain important insight into tumor biology.

We reduced the number of variables used to 27 variables and taken by the algorithm was 60% of the earlier runtime

How Random Forest works



- Random Forest operates by constructing a multitude of decision trees which only use a subset of all the predictors (input variables) and averaging their outputs to obtain a single low-variance statistical learning model.

- By implementing random forest, we got an accuracy of 96%

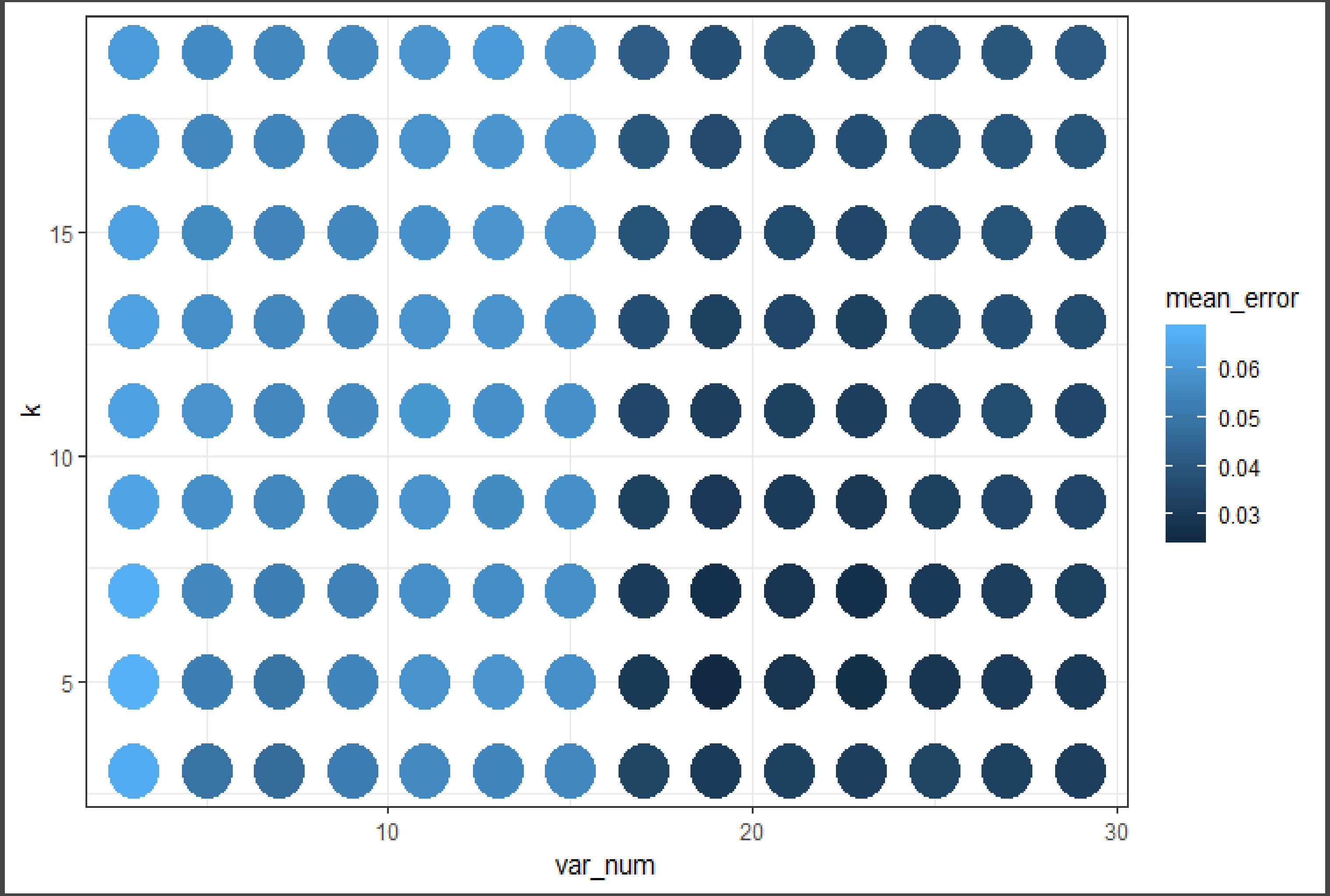
Comparison of kNN & Random Forest

Confusion Matrix

kNN	Benign	Malignant
Benign	80	3
Malignant	1	43

Random Forest	Benign	Malignant
Benign	280	6
Malignant	14	156

Post processing & Visualization



kNN performs pretty well with 3 variables and high k

kNN gets best for 19 variables and low value of k

Next Steps

- Applying the model on other Breast Cancer data sets to get more insights .

- Optimize parameter values of Random Forest to increase accuracy

- Applying stratified sampling method as sampling method for our data set.

- Increasing the accuracy of the model by using more efficient modelling techniques

Results & Summary

- It is observed that KNN is performing better for the data set with an accuracy rate of 97% compared to other model.
- With this accuracy rate we will be able to detect cancerous cells at early stage and provide preventive measures .

Thank You