VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

MACHINE LEARNING FUNDAMENTALS ECE3047 (L57+L58)

Allen Ben Philipose – 18BIS0043

LAB FAT

Problem Statement

Analyse the performance of KNN by Choosing 2 different data sets.

Train and Test KNN classifier using the cancer dataset for K=3,4,5. Calculate the result using three performance metrics.

Tools Required

Jupyter/VS Code – Python Notebook Code Editor.

Model – I, Inference – I, Evaluation – I in the notebook answers the first question and **Model – II, Inference – II, Evaluation – II** in the notebook answers the second question.

Dataset link

Cancer dataset for K=3,4,5

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Dataset for comparison of KNN performances

https://www.kaggle.com/uciml/pima-indians-diabetes-database

Header files included

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt


from sklearn import preprocessing

from sklearn.metrics import accuracy_score, log_loss,
confusion_matrix, f1_score

from sklearn.model_selection import train_test_split,
cross_val_score

from sklearn.model_selection import StratifiedKFold,
GridSearchCV, KFold

from sklearn.preprocessing import MinMaxScaler

from sklearn.metrics import classification_report


from sklearn.ensemble import RandomForestClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.naive_bayes import GaussianNB

from sklearn.neighbors import KNeighborsClassifier

from sklearn.linear_model import LogisticRegression
```

Imported all these modules for the completion of the program.

Inference – I

We can understand that the model from the Dataset 1 is giving **much** better performance values that Dataset 2.

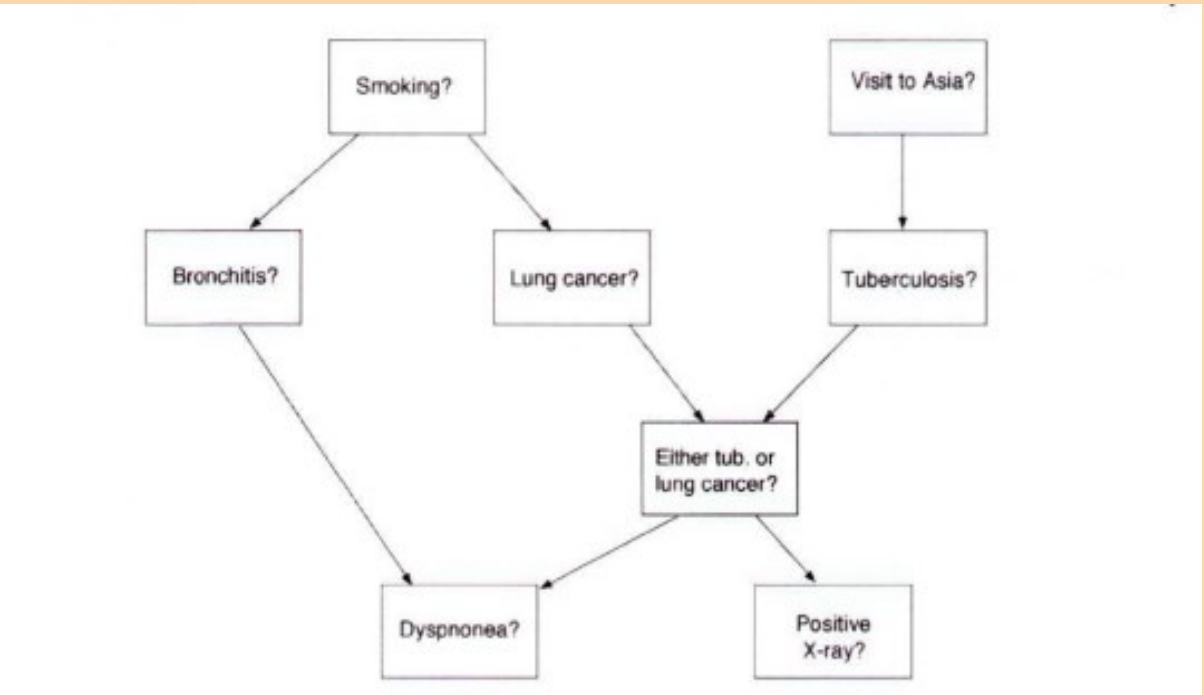
- Accuracy decreased in the second dataset because of the variance of values which causes underfitting of the model.
- F1 score, specificity and sensitivity can be derived from the confusion matrix of each, and even all those parameters show lesser value in the second dataset.

Inference – II

Hence from the experiment we have figured out that the highest performance is received when the value of K = 4 by analysing the performance metrics such as Accuracy, sensitivity which are derived from the confusion matrix and F1 score giving an overall analysed score of the model performance.

Question – 2

Considering the following Bayesian network connection various factors related to chest diseases, are Bronchitis and Tuberculosis independent when nothing is observed?



These are independent of each other because Tuberculosis is only caused if the action is “visit to Asia”. It has no relevance with the action “smoking”. This inference can be concluded by checking the action which produces the disease as an output. If nothing is observed, the probability of non-visitors of Asia having Bronchitis is 0. Hence we can say Tuberculosis and Bronchitis are independent when nothing is observed.