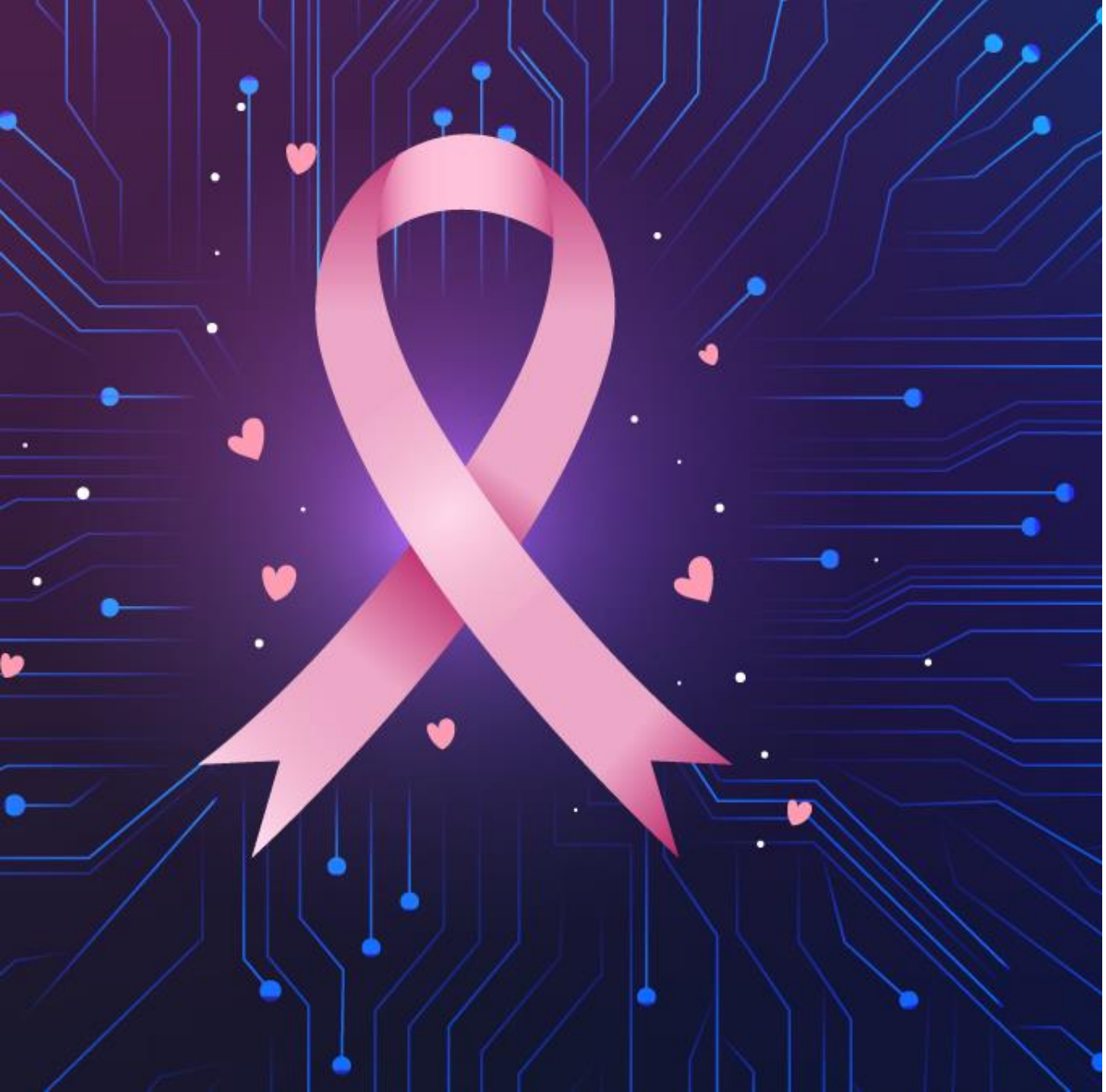
Data Science & Analytics

Breast Cancer Detection Using Machine Learning:

www.geninvo.com

North America | UK | Europe | India



Breast Cancer Detection Using Machine Learning

Business Need

- Breast cancer is the most common cancer and principal cause of death from cancer among women globally.
- Early detection is the most effective way to reduce breast cancer deaths.
- Early diagnosis requires an accurate and reliable procedure to distinguish between benign breast tumors from malignant ones
- Breast cancer detection is done by Biopsy approach which provide highly accurate results, but this process is very painful for the patients. Other techniques are mammogram and Magnetic Reasoning imaging (MRI) which is not very reliable.
- So, there is a need to construct a system to accurately differentiate between benign and malignant breast tumors.

Objective

- Given a list of features calculated from a digitized image of the FNA of a breast mass from a patient, the aim was how to diagnose (determine) whether the patient has breast cancer.
- Develop a comparative study of detecting breast cancer using different machine learning algorithms.
- Select the algorithm which has more promising results with respect to accuracy, F1 score, precision and recall and optimize that algorithm.t

Data Description

- This breast cancer databases was obtained from UCI Machine Learning Repository (Breast Cancer Wisconsin (Origional) Dataset: created by Dr. William H. Wolberg.)

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

First two columns:

- Sample ID
- Diagnosis

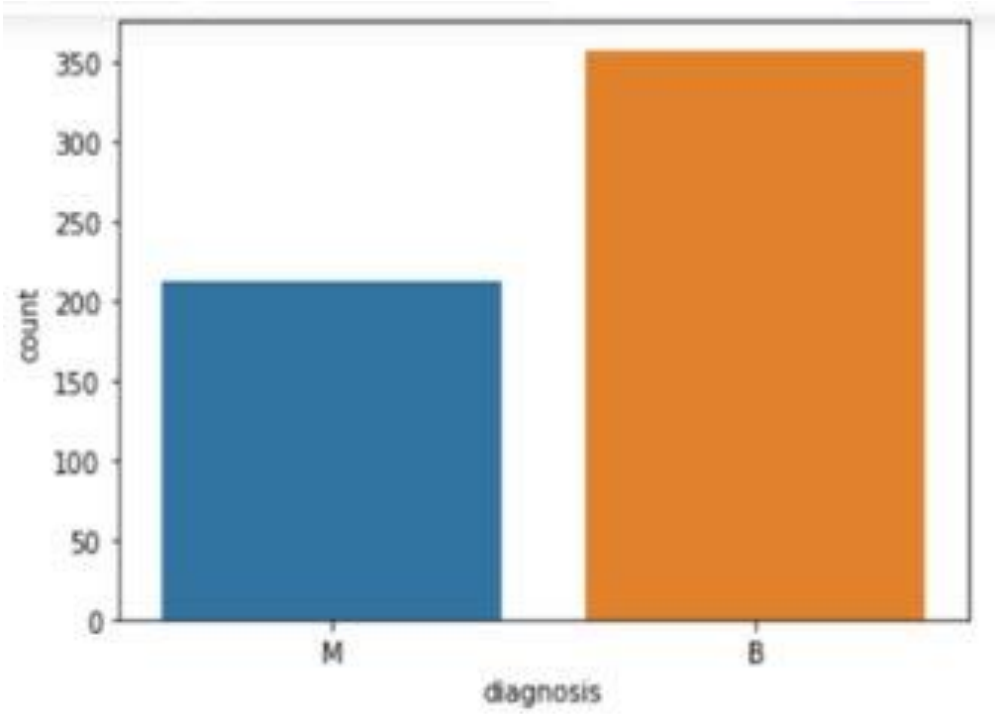
Column 3-32: For each cell nucleus, ten characteristics were measured:

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concave points
- Symmetry
- Fractal

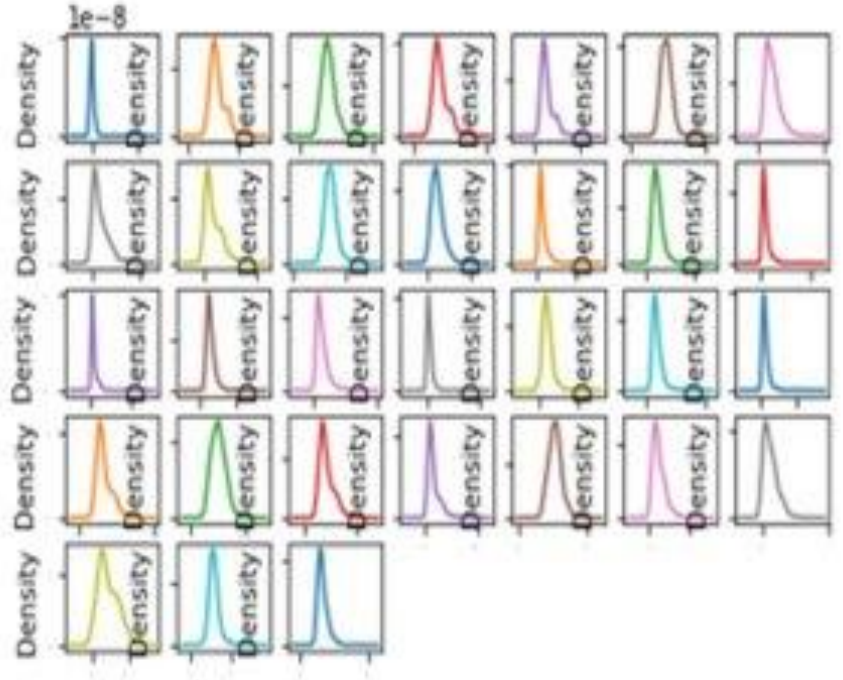
For each characteristic three measures are given:

- Mean
- Standard error
- Largest/ “worst”

Visualization:



Number of Malignant & Benign cases in dataset

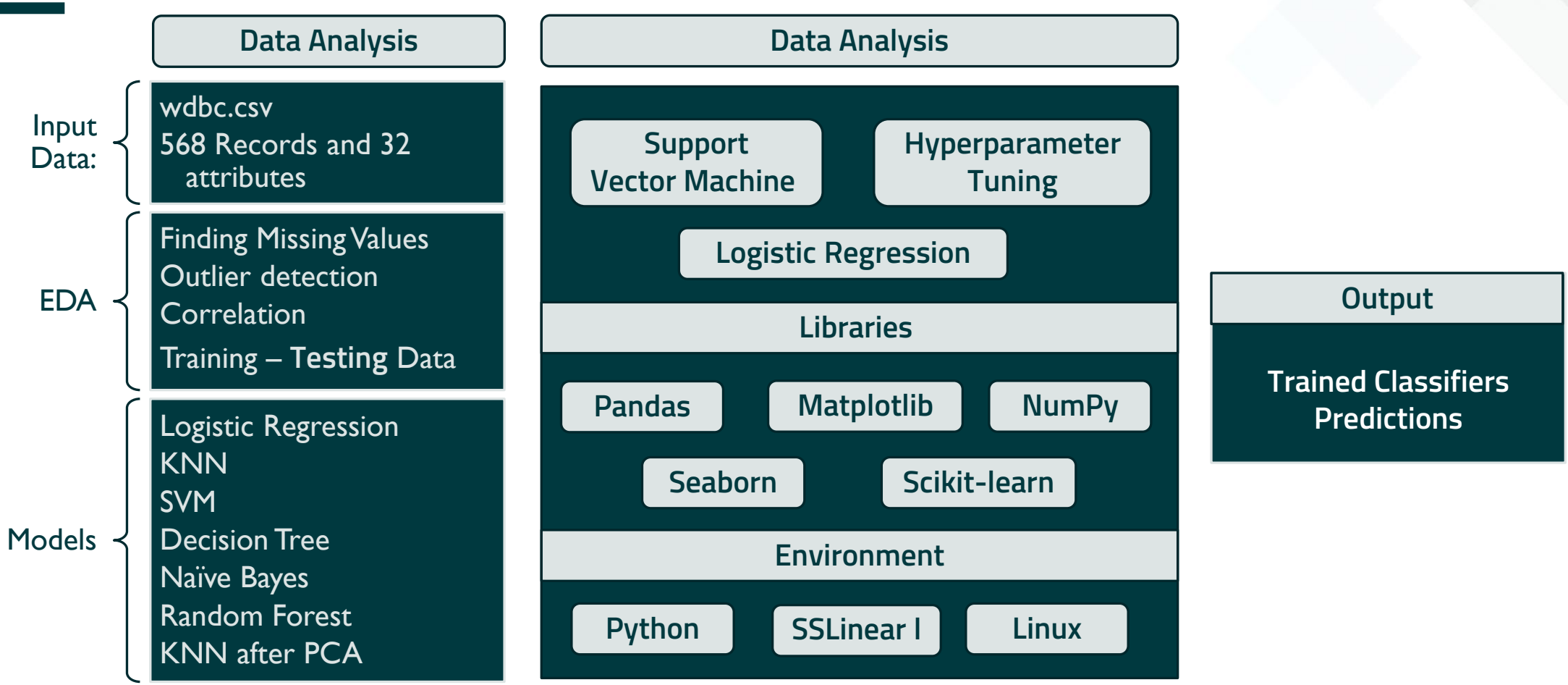


Density Curve

Solution Methodology

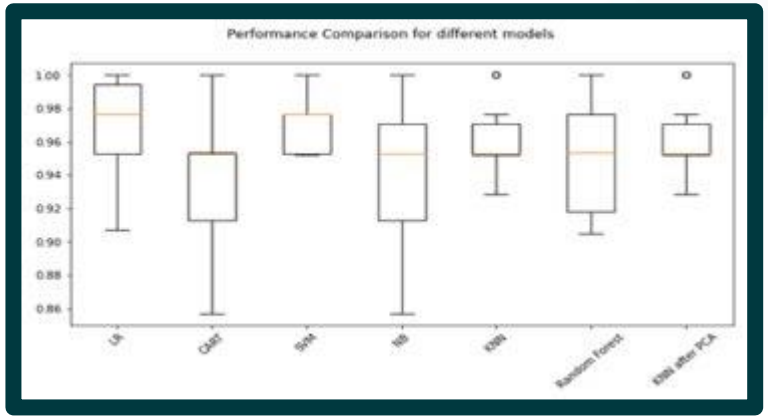
- Using the Breast Cancer Wisconsin (Diagnostic) Database and exploring different machine learning techniques we are creating a classifier that can help in predicting likelihood of breast cancer.
- Different machine learning algorithm explored are: Logistic Regression, KNN, SVC, Decision Trees, Random Forest and KNN after PCA
- Support Vector Machine with linear performed better with 97.6% accuracy. Logistic regression also has comparable results with 96.9% accuracy.
- First part of the study includes Exploratory data analysis, and the second part includes building different Machine learning models and comparison of their results. Third pard includes hyperparameter tuning for the SVM model.

Process


Data Analysis
Input Data:
wdbc.csv
568 Records and 32 attributes
EDA
Finding Missing Values
Outlier detection
Correlation
Training – Testing Data
Models
Logistic Regression
KNN
SVM
Decision Tree
Naïve Bayes
Random Forest
KNN after PCA
Data Analysis
Support Vector Machine
Hyperparameter Tuning
Logistic Regression
Libraries
Pandas
Matplotlib
NumPy
Seaborn
Scikit-learn
Environment
Python
SSLinear I
Linux
Output
Trained Classifiers
Predictions

Results

```
LR: Accuracy: 0.969491 (Standard Deviation:  0.027735) (run time: 0.393111)
CART: Accuracy: 0.936489 (Standard Deviation:  0.041308) (run time: 0.123792)
SVM: Accuracy: 0.971705 (Standard Deviation:  0.017747) (run time: 0.057464)
NB: Accuracy: 0.938760 (Standard Deviation:  0.042778) (run time: 0.023136)
KNN: Accuracy: 0.959967 (Standard Deviation:  0.023692) (run time: 0.117712)
Random Forest: Accuracy: 0.952879 (Standard Deviation:  0.035060) (run time: 0.331640)
KNN after PCA: Accuracy: 0.959967 (Standard Deviation:  0.023692) (run time: 0.101056)
```



Best Model

After Hyperparameter Tuning for SVM: `Best: 0.976467 using {'C': 0.1, 'kernel': 'linear'}`

This study demonstrates that SVM model was the best model with the highest level of accuracy. Hence, this model is recommended as a useful tool for breast cancer prediction as well as medical decision making.

Business Impact

- The ML/AI could help improve the scalability of breast cancer screening and ameliorate the shortage of mammography professionals around the world.
- 97% accurately identifies whether the cancer type is benign or malignant.
- There is no need to be an experienced physician as it provides substantial accuracy for senior or junior physicians alike.
- Helps diagnose malignant tumors quicker.
- Compared to FNA(biopsy) that produces fast, reliable, and economic evaluation of tumor lesions with a 79% accuracy there is an 18% improvement in breast cancer predictions through SVM.

References:

- Dr. WIlliam H. Wolberg (physician), Breast Cancer Wisconsin (Original) Data Set(1991), UCI Machine Learning Repository.

Ideas
Delivered
Intuitively

Thank You!

www.geninvo.com | Reach out to us info@geninvo.com