Applications of Machine Learning in the analysis of breast cancer

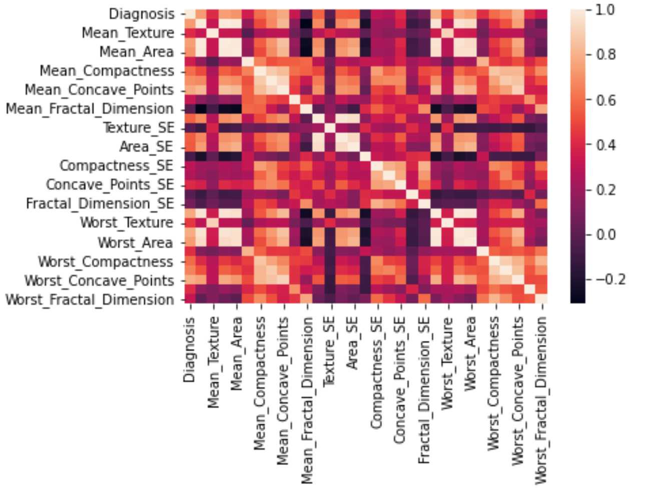
By Ultan Kearns & Liam Millar

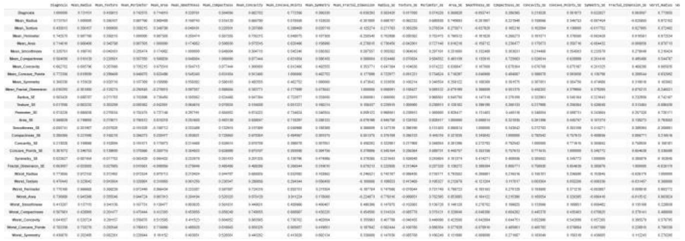
Introduction

- When starting this project we investigated many datasets finally settling on https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
- This dataset is from the University of Wisconsin – a well-known institute of higher learning in the USA
- The main objective of our project was to use Machine Learning models to accurately predict the presence / absence of breast cancer based on features in the data
- We used a combination of both Supervised and Unsupervised learning in this project to train our models

Cleaning and analyzing the data

- First thing we did when starting this project
- We noticed diagnosis had M for malignant and B for benign we decided to replace these with 1 and 0 respectively
- Did this so that we could perform numerical operations easier
- We started by showing a heatmap of our correlations as you can see on the right hand side
- Did this to analyze relations in the data and to see which features were highly correlated with our diagnosis
- We also played around with the training / test set ratios and finally settled on a 70 / 30 split





Models Used

Linear Regression – to predict and analyze the correlation of 2 features in our dataset

K Means – This is an unsupervised learning technique creating clusters of data based on a centroid

KNN – This is a clustering technique which analyzes data nearest to other data to predict a diagnosis

Naïve Bayes – This is a technique used to predict cancer by taking features of our dataset and using the same weight for each -> Assumes data has same effect on output hence Naïve

Models Used - Continued

Decision Trees – Which were used to predict diagnosis by making a tree of features in our dataset which will either lead to positive or negative diagnosis based on their values

Random Forest – Creates multiple trees then merges the best models to predict our diagnois

PCA (Principle Component Analysis) - unsupervised dimensionality-reduction which reduced features in our dataset by removing similar features which were highly correlated

Analysis of our models

- Linear Regression was used to analyze the strength of the relationship between certain features and our Diagnosis – Also used Cramérs V
- Had good results we could see which features were positively correlated with a diagnosis value and the strength of this correlation
- K Means – Performed fairly well with 86% accuracy
- KNN – Performed better than KNN with 91% accuracy
- Gaussian Naïve Bayes – without smoothing or scaling had 90% accuracy on testset

Analysis of our models – Continued

- Gaussian Naïve Bayes with scaling had around a 91% accuracy rate – not much improvement
- Decision Tree model had an accuracy of 93%
- Random forest had around the same degree of accuracy
- Entropy Forest had 90% accuracy
- PCA Gini Tree had an accuracy of 0.94%
- PCA Random forest had the best accuracy of all our models at 97%

Final Results Table

- Here we can see our final results table
- Notice which models performed correctly and which didn't
- Trial and error process – it took time finding the right ratio of the training / test sets
- Also it took time to analyze the models and determine how we could get the best performance from them

A. Accuracy of Models

Method Description	Accuracy
PCA Gini Random Forest:	0.98%
cross validation gini random forest:	0.96%
Standard trained gini random forest:	0.95%
cross validation entropy random forest:	0.95%
PCA Gini Tree:	0.94%
Standard trained entropy random forest:	0.94%
Standard trained gini decision tree:	0.94%
cross validation entropy decision tree:	0.92%
Navie Bayes - with scaling:	0.91%
Standard trained entropy decision tree:	0.91%
cross validation gini decision tree:	0.91%
K Neighbour:	0.91%
K Means:	0.86%

Conclusion

From our study we we determined PCA Random Forest had the best accuracy when predicting the diagnosis

We also learned the limitations of machine learning in healthcare – should be used as an assistant not an expert as even the best trained models can yield false predictions

We learned the importance of data cleaning and analysis

We learned which models worked on our dataset and which didn't – it was a fairly small set of only 570 values!

Questions