From Unsupervised To Supervised Machine Learning: Application In Healthcare

By Pallavi Gupta

Abstract

There is no field in today's world where machine learning applications are not being utilized. This paper talks about applying machine learning approaches to help making cost-effective and more targeted care of diseases (apart from cancer) among cancer patients. Data was collected from thyroid cancer patients about the diseases/comorbidities they have apart from cancer. It was hypothesized that patients could be clustered into groups on the basis of comorbidities they have; as there are more than one comorbidities that occur in most of the cancer patients. And having them clustered into groups would help in treating them medically in an efficient way. To cluster the patients having similar comorbidities, unsupervised clustering was applied, which involved hierarchical clustering and k means clustering. This gave information of the diseases forming clusters and there were 3 major clusters identified. The existing unlabelled data was then labelled as per the information derived from unsupervised learning. The supervised machine learning models were then created using this labelled data keeping in mind the future perspective. The supervised machine learning models were created using k-Nearest Neighbors(KNN), Random Forest Classifier, Logistic regression classifier (using ovr settings), and Sequential Neural Network. All the models gave accuracy ~99%, which is a good mark.

Introduction

Comorbidity is common among cancer patients; in fact, data from Medicare beneficiaries in the United States indicate that 40% of cancer patients have at least one or the other chronic condition recorded, and 15% have two or more – the most common chronic conditions include cardiovascular illness, obesity and metabolic diseases, mental health problems, and musculoskeletal conditions (doi: 10.1002/cncr.28509). Comorbidity is not limited to the elderly however; in a study of 485 young adults who were cancer patients, aged 15 to 39 years, 14.6% reported two or more comorbidities (doi: 10.1158/1055-9965.EPI-15-0401). The complexities of care coordination/prioritization for comorbidity create challenges for patients and providers alike (doi: 10.5402/2011/815013), especially in the absence of comorbidity-specific clinical guidelines (doi: 10.1158/1055-9965.EPI-17-0439). As medical advances improve and the United States (U.S.) population ages, identifying the most cost-effective management of comorbidity is salient, as evidenced by the Department of Health and Human Services' launch of the 2015 initiative focusing on optimizing health and quality of life for individuals with multiple chronic medical conditions. Such initiative is timely, especially with the increasing incidence rates for cancers with high survivorship, such as thyroid cancer.

Objectives and Motivation

Identifying clusters of comorbidities may be important for providing cost-effective and more targeted care. Unsupervised machine learning have been used extensively in other domains such as marketing. For example, it may be identified that individuals who purchase spaghetti may also purchase tomato sauce. Since spaghetti and sauce are found in the same 'clusters,' business analysts gain insight to better target advertisements to increase sales. Understanding the types of clusters that exist provide insight on where to invest. Similarly, this research study deals with utilizing pre-existing knowledge that a large number of cancer patients have co-occurring diseases(comorbidities), and this study is focused to identify clinically meaningful cluster of co-morbidities, based on their co-occurrence, to gain healthcare insight. Once this information is achieved, Machine Learning (ML) models can be created to classify the patients at a wider level keeping in mind the future perspective of providing cost-effective healthcare to the cancer patients.

Utilizing these methods allow for the opportunity to maximize insight about cancer care to avoid future disparities and wasted resources by understanding commonly occurring phenotype for hypothesis generation.

Methods

Data preprocessing and Feature Extraction

The data was collected via self assessment forms to get the information of the comorbidities in the cancer patients. The target patients were thyroid cancer patients. Each patient information was linked to a unique DUPERSID. First of all, the data was filtered based on comorbidities and only those instances were retained who had comorbidities in the record. It was surprising to see that only 1.7% of the instances had no comorbidities associated.

In clinical research, there are CCS codes assigned to each disease or comorbidity. CCS is a tool for clustering patient diagnoses and procedures into a manageable number of clinically meaningful categories developed at the Agency for Healthcare Research and Quality (AHRQ, formerly known as the Agency for Health Care Policy and Research).

108:Congestive heart failure, nonhypertensive 106:Cardiac dysrhythmias 096:Heart valve disorders 103:Pulmonary heart disease 105:Conduction disorders (114:Peripheral and visceral atherosclerosis 115:Aortic, peripheral, and visceral artery aneurysms) (098:Essential hypertension 099:Hypertension with complications and secondary hypertension) (082:Paralysis 081:Other hereditary and degenerative nervous system conditions 083:Epilepsy, convulsions) (127:Chronic obstructive pulmonary disease and bronchiectasis 128:Asthma) (049:Diabetes mellitus without complication 050:Diabetes mellitus with complications) 048:Thyroid disorders 158:Chronic renal failure (150:Liver disease, alcohol-related 006: Hepatitis 151:Other liver diseases) (139:Gastroduodenal ulcer (except hemorrhage)) 005:HIV infection 202:Rheumatoid arthritis and related disease 062:Coagulation and hemorrhagic disorders 058:Other nutritional, endocrine, and metabolic disorders 052:Nutritional deficiencies 055:Fluid and electrolyte disorders 066:Alcohol-related mental disorders 067:Substance-related mental disorders (070:Schizophrenia and related disorders 071:Other psychoses) 072:Anxiety, somatoform, dissociative, and personality disorders 069:Affective disorders 657:Mood disorders

Figure1: Showing the list of CCS codes for comorbidities.

So the next step was to remove the redundancy in data using this CCS codes and create binary variables for comorbidities based on CCS codes.

	DUPERSID	CCCODEX
60	40005101	053
61	40005101	024
62	40005101	204
63	40005101	232
64	40005101	141
65	40005101	204

Figure 2: Notice that item #62 and #65 are duplicates below. This is because the icd 9 codes or some other column in the dataset were different, despite being in the same CCS category; we need to ensure that there are only one unique row

Since, there could be more than one CCS codes contributing to the same condition, like 114,115 both lies in Peripheral Heart disease category; so the data was merged for every health condition and binary data was created for that. Later, everything was merged into a single table with 3325 rows, each representing one patient, and 25 columns (each representing a comorbidity). The dataset was ready at this point for the unsupervised clustering analysis.

Unsupervised Clustering:

Hierarchical Clustering(Paired distances)

Hierarchical clustering is the most popularly used unsupervised machine learning classification technique in the biological world. Since its unsupervised, the data has no predetermined labels associated with data. Based on various metrics and distance (like Euclidean, Manhattan) between data points, they are clustered into various groups. Since, the data in our case was binary, the paired distances technique was a good approach to follow. I used “paired distance method” to cluster data, along with 8 different metrics such as Braycurtis, Cosine, Euclidean, Hamming, Manhattan, L1, L2 and Minkowski. Using 8 different metrics helped in giving a confident consensus of diseases forming a cluster together. The results section displays the trees generated by using the 7 different metrics.

K-Means Clustering:

K means is another simple but quite widely used unsupervised machine learning classification technique in the biological world. It groups similar data points together and discover underlying patterns. For this paper study, elbow method was used to determine the best value for k to cluster the data. k (number of centroids) from 2 to 8 was used and plotted against SSE, to check which k fits the best for the data. The results section displays a graph on this.

Supervised Clustering:

Labelling of data

Results from the hierarchical clustering and k means clustering, helped identifying the possible clusters. Drawing this information, the data was labelled with the cluster numbers (1,2,3). This labelling was useful to create some supervised machine learning models for the data.

Splitting of data:

For splitting the data into train and test, ratio of 80:20 was used. The training set was further split into train and validations set making the final ratio of training set, validation set and test set to be 70:10:20.

Models Applied:

Since, from the unsupervised study the clusters identified were 3, supervised models picked were for non-linearly separable data except for logistics regression where the multiclass data could be used with the setting multi_class='ovr'.

Model 1: Random Forest Classifier:

Random forest classifier is an ensemble algorithm. This creates a set of decision trees from randomly selected subset of training data. It then aggregates the votes from different decision trees to decide the final class of the test object. For this study, both entropy and gini impurity criterion were used. N_estimators were set to 10 and max_depth (being a hyperparameter) was set to 15 to prevent overfitting.

Model 2: Logistic Regression Classifier (with ovr settings)

Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for binary classification. However, it can also be used for data with multiclasses, using ovr settings (in sklearn), which makes classification using one vs all. The same procedure was used for this study as well.

Model 3: K Nearest Neighbour Classifier (KNN)

KNN uses the distance criteria(like Euclidean, Manhattan) between the two data points to predict and classify them into groups. But finding the value of K is the most critical for this approach. To find the best k-value for this study, KNN was ran for a range of k values (using cross validation=10 for each run). Then the accuracy graph was obtained for a range of k values, which helped in identifying the best k value.

Model 4: Neural Networks (NNet)

Neural networks is one of the most powerful and widely used algorithm in the field of machine learning. For this study, the sequential model was used with one hot encoding. There were two types of models used as follows:

1)Single Hidden Layer NNet Model

There was one input layer, followed by one hidden layer and one output softmax layer. The activation of input layer and hidden layer was tried with both tanh and relu in separate runs.

```
25 inputs -> [10 hidden nodes]  ->   3 outputs
(Input layer)   (Hidden layer)       (softmax layer)
```

2)Two Hidden Layer NNet Model

There was one input layer, followed by two hidden layers and one output softmax layer. The activation of input layer and hidden layers was tried with both tanh and relu in separate runs.

```
25 inputs  -> [10 hidden nodes]-> [8 hidden nodes] -> 3 outputs
(Input layer) (1st Hidden layer) (2nd Hidden layer)   (softmax)
```

Both the models have 3 output values because one-hot encoding was used and since dataset had 3 types of labels/classes associated; hence the output layer must create 3 output values, one for each class. Compilation of both the models were done using categorical_crossentropy for loss, adam and rms_prop as optimizers, and finally accuracy as metrics.

There is one more important feature that keras library provides and was used in this study. There are wrapper classes to allow use of neural network models developed with Keras in scikit-learn. There is a KerasClassifier class in Keras that can be used as an Estimator in scikit-learn, the base type of model in the library. The KerasClassifier takes the name of a function as an argument. This function must return the constructed neural network model, ready for training. Last but not the least, the models were evaluated with k-fold cross validation with k=10.

Results:

Unsupervised Approach:

Two step analysis was done to confirm the unsupervised clusters. Hierarchical clustering helped in portraying a consensus, for which diseases form a cluster together. K-means helped in identifying the possible numbers of clusters without overfitting the data. This two step approach worked very well. As per k means elbow method plot, there was a sharp elbow at k=3, but a slight shallow elbow at k=4. However, k=3 gave better clusters.

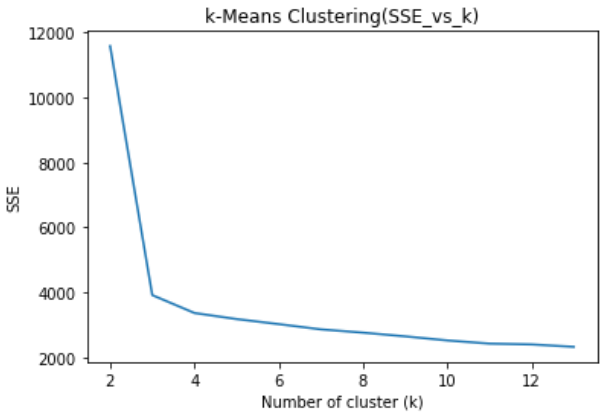

Figure 3: K means clustering Elbow method plot (SSE vs k)

As per hierarchical clustering results, there were 3 major clusters identified as follows:

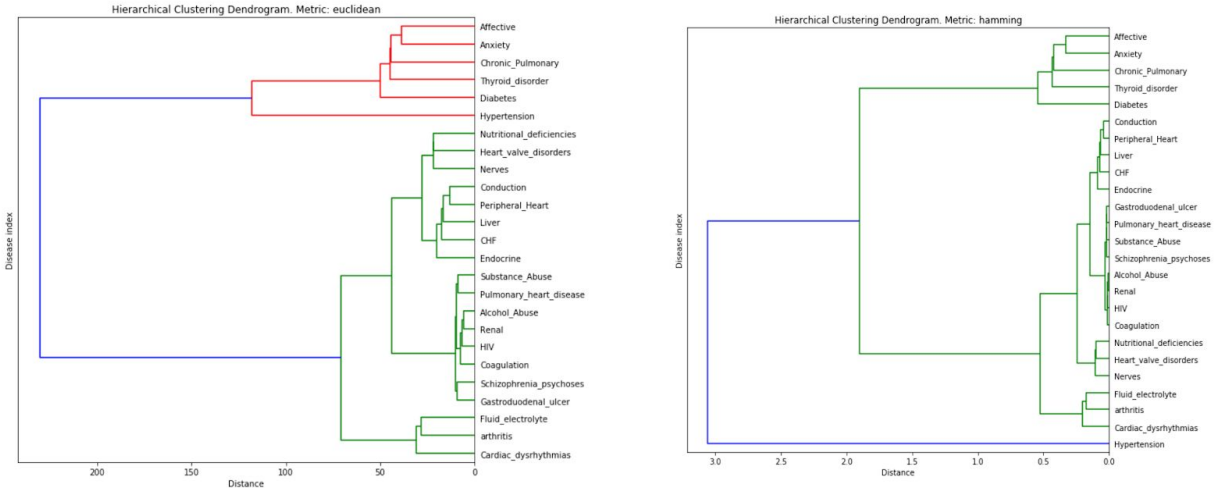

Figure 4: Dendrogram based on Euclidean distance metric (on left); based on Hamming distance metric(on right)

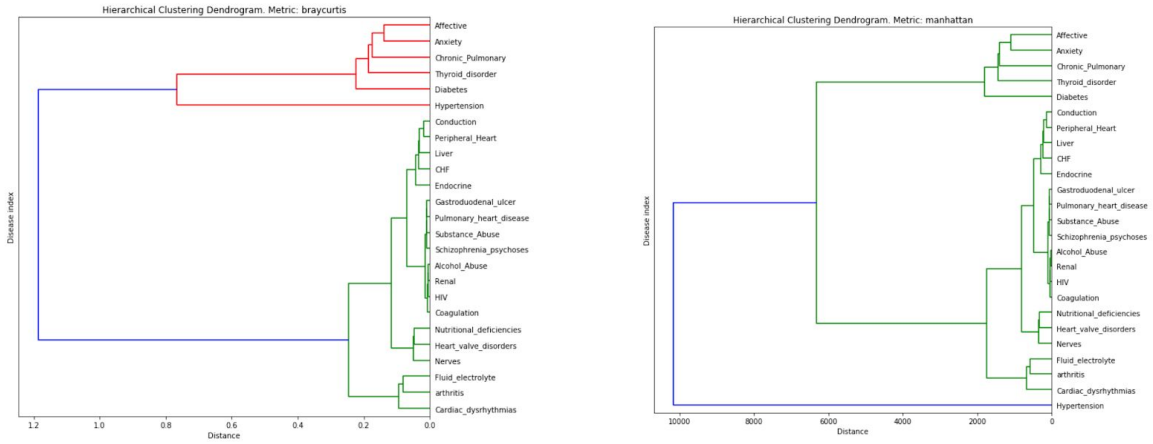

Figure 5: Dendrogram based on Bray Curtis distance metric (on left); based on Manhattan distance metric(on right)

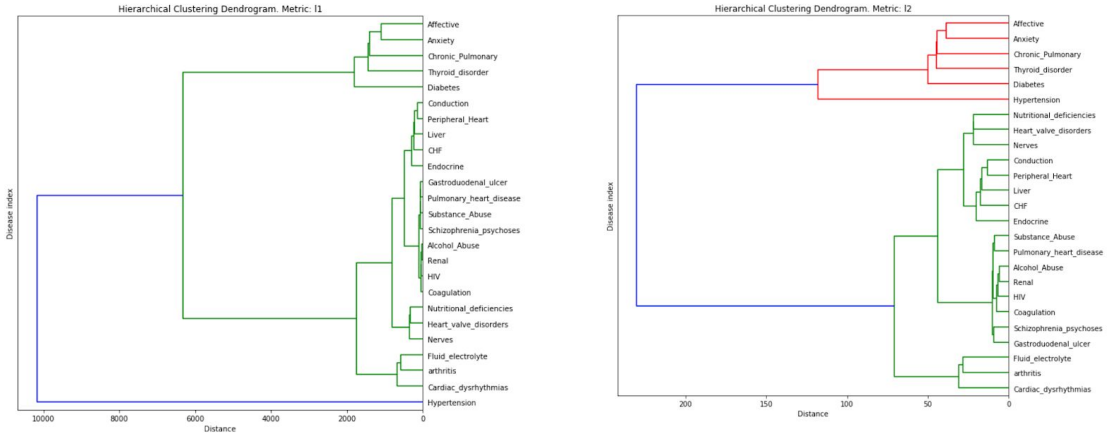

Figure 6: Dendrogram based on l1 distance metric (on left); based on l2 distance metric(on right)

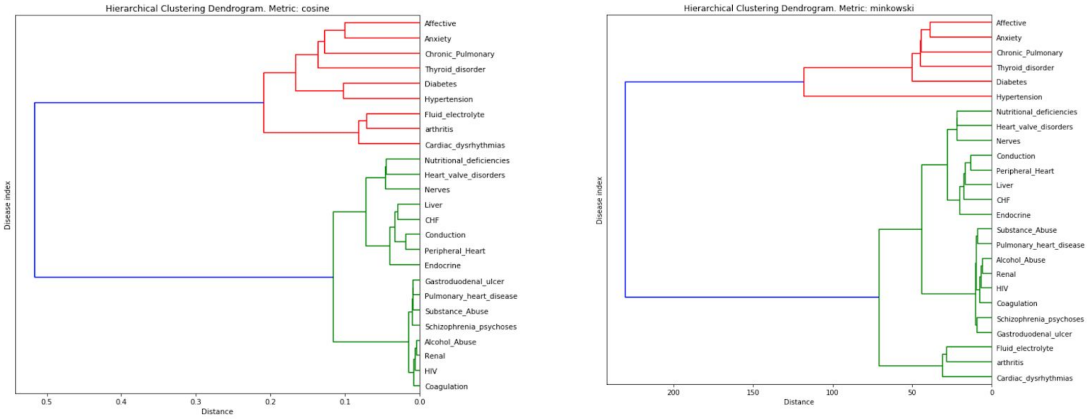

Figure 7: Dendrogram based on Cosine distance (on left); based on Minkowski distance(on right)

Cluster1: Affective, Anxiety, chronic pulmonary, thyroid_disorder, diabetes, and Hypertension

Cluster 2:

- **Sub Cluster1**: Nutritional deficiencies, Heart_valve_disorders, Nerves.
- **Sub Cluster2:** Conduction, Peripheral_Heart, Liver, CHF, Endocrine.
- **Sub Cluster3:** Substance Abuse, Pulmonary Heart Disease, Alcohol Abuse, Renal, HIV, Coagulation, Schizophrenia_psychoses, Gastrouodenal_ulcer.

Cluster3: Fluid Electrolytes, Arthritis, Cardiac_dysrhythmias.

Biological relevance to clusters

Clusters of disease obtained, also had biological relevance to it. e.g cluster 1 which clustered together diseases affective, anxiety, chronic pulmonary, thyroid_disorder, diabetes, and hypertension. Biologically all the diseases in this cluster were observed to be correlated. Anxiety is a type of mood disorder and can be taken as a types of affective disorder. Anxiety can cause dramatic, temporary spikes in your blood pressure, which if happens repeatedly, causes hypertension**.** Pulmonary hypertension is a type of high blood pressure that affects the arteries in your lungs and the right side of your heart. This increases your risk of developing diabetes, and it makes it harder to manage your blood sugar if you already have diabetes. Hyperthyroidism, which is overactive thyroid hormone, and hypothyroidism, which is underactive thyroid hormone, are both associated with mild hyperglycemia (elevated glucose levels). So it can be clearly seen how one disease can switch the cascade effect for other diseases and if one disease is triggered it is quite obvious that other diseases associated can also get triggered. So the diseases triggered together were put in one cluster. Similarly, the other diseases clustered together had a correlation with each other. This helped us in segregating diseases in 3 different clusters.

Cluster 2 had some subclusters, but they did not have much consensus in terms of data points associated with them, so majorly 3 major clusters were picked. To be more elaborative, there were total 3325 instances. Sub clustering, was distributing cluster2 into a small clusters which would have added to the skewness; hence all the subclusters were added to one major cluster cluster2.

<u>Supervised Approach (Models):</u>

Model 1: Random forest Classifier

Usage of entropy and gini impurity did not make any much different, but setting n_estimators=10 and max_depth=15 helped in achieving 99% accuracy.

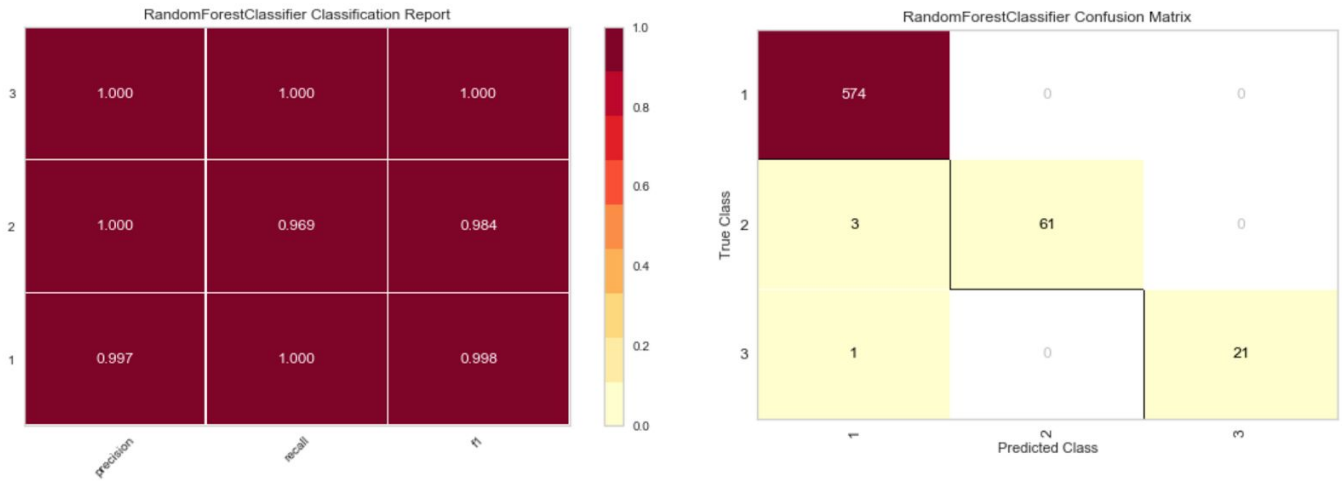

Figure 8: Random Forest Classifier Precision, Recall Plot(on left); Confusion matrix(on right)

Model 2: Logistic Regression(with ovr settings)

This model gave 98.9% accuracy.

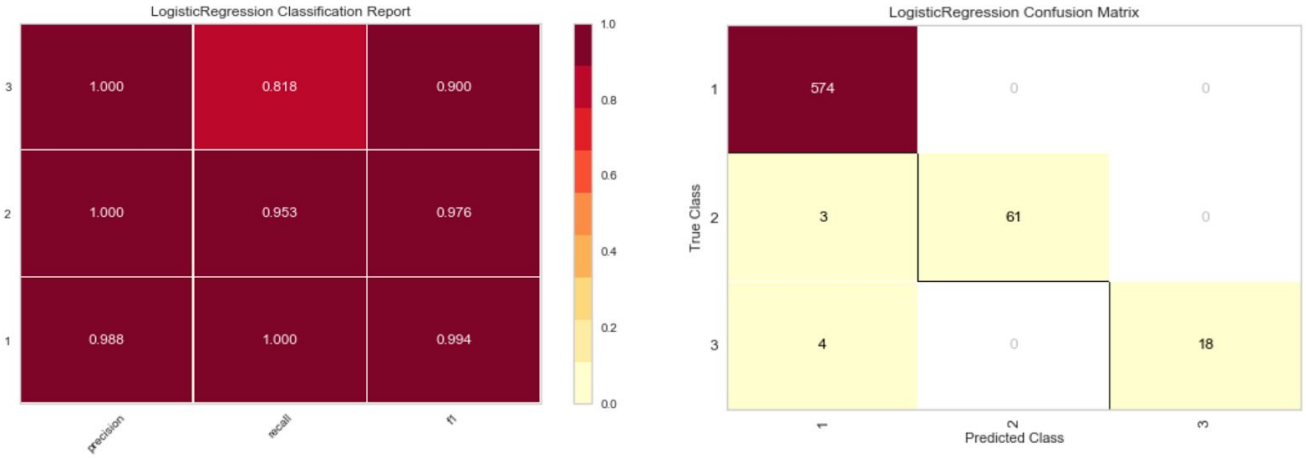

Figure 9: Logistic Regression Precision, Recall Plot(on left); Confusion matrix(on right)

Model 3: K nearest neighbors (KNN)

It is hard to find the best value for k in k nearest neighbors algorithm. Due to this constraint, scores for a varying range to k was captured in order to achieve the best value for k. Also for each value of k, cross validation of value 10 was applied. The result gave the best accuracy of 98% for k=7 and 3. But k=7 was chosen.

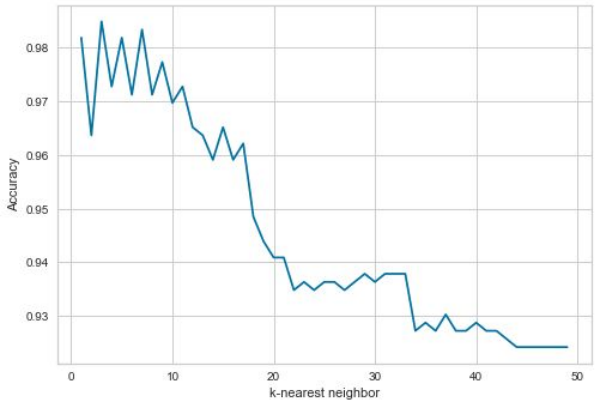

Figure 10 : Accuracy with varying values of k and cross-validation=10

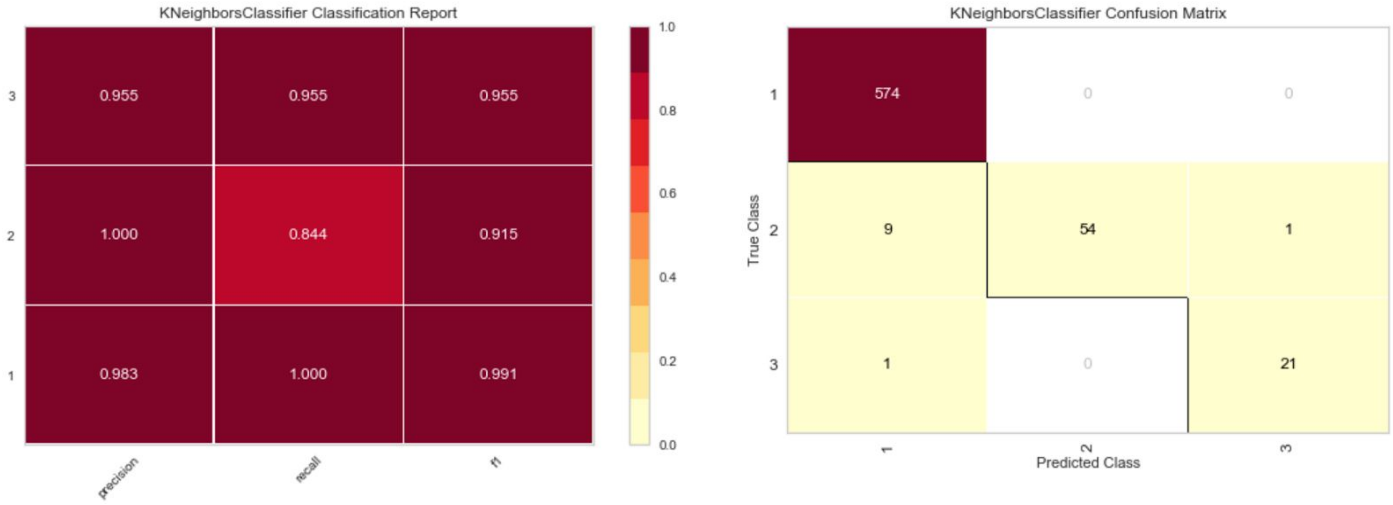

Figure 11 : KNN Precision, Recall Plot(on left); Confusion matrix(on right)

Model 4: Neural Network

The Single Hidden Layer NNet Model gave ~98% accuracy with both tanh and relu. However, adding a hidden layer and using Two Hidden Layers NNet Model gave 99.7% accuracy.

Discussion:

How much easier it would be to treat people medically, if it is predetermined that they fall in which cluster of comorbidities. Sometimes, medicine given to treat a disease, can have side effects; which could be negative and might worsen conditions for other disease, for the same person. But if the patient is segregated into such clusters, cost effective and targeted health care can be provided for the cancer patients; that was the hypothesis initially, before starting this study. After the study, looking at the results, the hypothesis is proven right to an extent. Machine learning is really a game changer in such kind of analysis. Who would have imagined that data collected by some kind of forms from patients could develop into such an informative

analysis using machine learning techniques. ML is indeed a boon to scoop out information given any data, if used judicially.

References:

1. Edwards BK, Noone AM, Mariotto AB, et al. Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer.* 2014;120(9):1290-1314.
2. Wu XC, Prasad PK, Landry I, et al. Impact of the AYA HOPE Comorbidity Index on Assessing Health Care Service Needs and Health Status among Adolescents and Young Adults with Cancer. *Cancer Epidemiol Biomarkers Prev.* Dec 2015;24(12):1844-1849.
3. Psarakis HM. Clinical Challenges in Caring for Patients With Diabetes and Cancer.*Diabetes Spectrum.* 2006;19(3):157.
4. Gold HT, Makarem N, Nicholson JM, Parekh N. Treatment and outcomes in diabetic breast cancer patients. *Breast Cancer Res Treat.* Feb 2014;143(3):551-570.
5. Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatment. *CA: a cancer journal for clinicians.* 2016.