CS205 Project Report

Shruti Sharan
Department of Computer Science
University of California, Los Angeles
shruti5596@g.ucla.edu

Abstract

This paper presents a study of building a predicitve model to predict being positive to cancer or malignancy. It describes the different stages of feature selection and feature engineering, data preprocessing, and finally predictive modelling. Different Machine learning algorithms are explored to achieve this task. A complete analysis of all the models and their performances are analyzed in the final section of the paper.

1 Introduction

In this project we work with the dataset curated by CDC NHANES for the years 2015-2016 which has the cumulative information of those years and all the years prior to it. We use this dataset to predict positive cancer or malignancy. Accurately assessing cancer risk in average and high-risk individuals and determining cancer prognosis in patients are crucial to controlling the suffering and death due to cancer. Cancer prediction models provide an important approach to assessing risk and prognosis by identifying individuals at high risk, facilitating the design and planning of clinical cancer trials, fostering the development of benefit-risk indices, and enabling estimates of the population burden and cost of cancer. Medical events are very often associated with multiple risk factors. Considering risk factors in isolation means ignoring relevant information and will typically lead to a decrease in predictive accuracy. Predictive modelling aid in the evaluation of treatments and interventions. Thus we endeavor to predict this real problem of Cancer by using Machine Learning.

2 NHANES

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States[1]. The survey is unique in that it combines interviews and physical examinations. The NHANES program began in the early 1960s and has been conducted as a series of surveys focusing on different population groups or health topics. In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the country, 15 of which are visited each year. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. The sample for the survey is selected to represent the U.S. population of all ages. To produce reliable statistics, NHANES over-samples persons 60 and older, African Americans, and Hispanics.

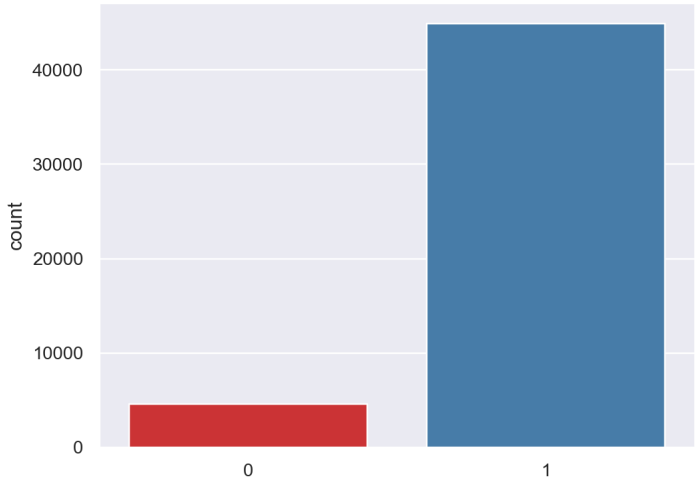

Figure 1: Data distribution of people with and without cancer.

3 Machine Learning for Cancer Prediction

Every year, Pathologists diagnose 14 million new patients with cancer around the world. That's millions of people who'll face years of uncertainty. Machine Learning (ML) is one of the core branches of Artificial Intelligence. It's a system which takes in data, finds patterns, trains itself using the data and outputs an outcome. Firstly, machines can work much faster than humans. A biopsy usually takes a Pathologist 10 days. A computer can do thousands of biopsies in a matter of seconds. With the advent of the Internet of Things technology, there is so much data out in the world that humans can't possibly go through it all. That's where machines help us. They can do work faster than us and make accurate computations and find patterns in data.
There has been a lot of ongoing research on using Machine learning models for prediction and analysis of data. This field of study is rapidly increasing with the growth and acceptance of Neural Networks and deep learning. Though its working is still a Black box, it is proving to give very good results in the medical domain and this paper aims at exploring this.

4 Dataset

The dataset is curated by CDC NHANES for the years 2015-2016 which has the cumulative information of those years and all the years prior to it. The data is divided into different categories as follows:

- Demographics Data
- Dietary Data
- Examination Data
- Laboratory Data
- Questionnaire Data

The Demographic data provides individual, family, and household-level information of the participants. It also includes related information such as Languages, interpreters used, gender, age, race/Hispanic origin, education, marital status, military service status, country of birth, citizenship, and years of U.S. residence.
The objective of the dietary interview component is to obtain detailed dietary intake information from NHANES participants. It is used to estimate the types and amounts of foods and beverages (including all types of water) consumed during the 24-hour period prior to the interview, and to

estimate intakes of energy, nutrients, and other food components from those foods and beverages. The Examination data section provides data for three consecutive blood pressure (BP) measurements and other methodological measurements to obtain an accurate BP, Heart rate or pulse, depending on age. The Laboratory Data includes biological specimens (biospecimens) for laboratory analysis to provide detailed information about participants' health and nutritional status. Eligibility for specific laboratory tests was based on the survey participants' gender and age at the time of screening by NHANES. The biological specimens included blood, urine, oral rinse and vaginal swabs. The Questionnaire data comprises of different interviews with the patients. These questions span over a variety of topics and gauge a lot of information about the patient. The topics range from alcohol use, smoking, relationship status, drug use, medical symptoms etc.

Each of these categories have subcategories providing detailed information about different features along with the feature names. One of the features in the questionnaire category is MQ220 which states "Doctor ever told you had cancer or malignancy". This feature is used as the target class. This is a very difficult problem as the dataset is extremely skewed. The number of people without cancer outnumber the people with cancer by a large extent making it a very unbalanced dataset as can be seen in Figure 1. It is a very skewed dataset, which shows how difficult our problem is going to be.

5 Feature selection

For an initial approach the technique employed was an exhaustive grid search through all the features in the given dataset. After multiple attempts this approach was discarded, as the dataset was very large and the scope of computation was beyond capabilities. As an alternative, possible features that made sense were selected manually after reading research papers and searching available resources. These features were then verified by consulting with an oncologist (Dr Indranil Ghosh, India). 110 features were selected across all the sub categories of the NHANES website. This seemed as a good starting point for our research and further statistical tools were employed for feature selection on this reduced feature set. The standard statistical measures considered were Simple Correlation, Mutual Information and Pearson's correlation constant. Since the latter only determines liner correlations, the results found were not very useful. Thus only Correlation and Mutual Information was used.

5.1 Correlation

The data was represented as a dataframe using the Pandas framework on Python to leverage all the data functions that can be applied directly to it. Using the direct correlation between each feature a matrix was computed to see how features are related to each other. After the correlation matrix

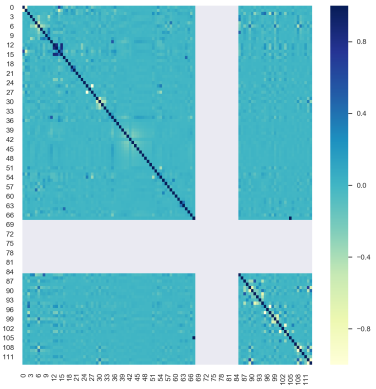

Figure 2: Correlation before removal of invalid features

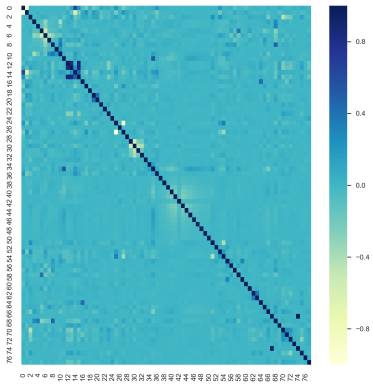

Figure 3: Correlation after removal of invalid features

was plotted, there was a huge missing band (Figure 2) that was seen. This implied that the values in

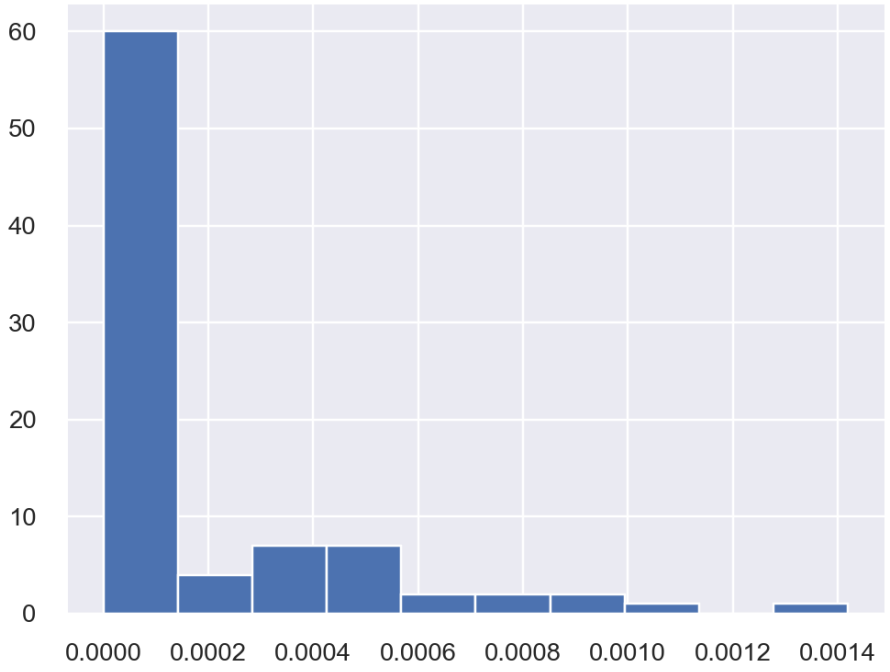

Figure 4: Distribution of Mutual Information

these features were missing completely. Fault localization was performed in a binary search method to pin point which feature was causing this error. After multiple tries, it was discovered that only one feature was causing this error. It spanned over a band, since it was one-hot encoded and had multiple values. After the feature was removed, the correlation matrix was perfect, though it was very noisy. (Complete correlation matrix included in appendix) The darker colours of the graph imply high correlation while the lighter parts imply low correlation.

5.2 Mutual Information

Mutual information is one of many quantities that measures how much one random variables tells us about another. This measure was used to compute the relation between each feature with the target feature. The features which had high mutual information were retained while the ones that were 0 or less than 10^{-15} were dropped.

As we can see, most of the features have a mutual information of 0 which implies that they have no correlation with the target. Many of the features were dropped initially, but after further analysis, it was seen that some of the features proved to be useful despite having a very low mutual information. This was probably because there were certain features that were correlated in higher dimensionality and mutual information isn't enough to capture such information.

5.3 Neural network for capturing non linear correlations

After experimenting with the selected features, it was seen that certain features resulted in better performance despite having close to 0 mutual information. Since Mutual correlation calculates the distance between two probability distributions, it may not be able to capture all the non-linear dependencies. A hypothesis was to use a simple neural network with a non linear activation function such as sigmoid, tanh etc to capture such relations. Unfortunately, after running the model it was difficult to understand any relations as extracting information from the hidden layers was challenging.

After calculating the above mentioned metrics some features were dropped and finally a feature set comprising of 56 features, was selected. (Complete list included in appendix). While adding all the features as columns into a feature dataset, it was encountered that all the SEQN numbers in the rows were not unique across all the features. This was due to a naming discrepancy at the NHANES end. To overcome this problem, an outer join was made to merge all the data into one dataframe. Since an outer join was performed, there would be a lot of repetition of data, thus it was grouped by SEQN number and the first of each kind were taken into consideration. This ensured that all the data over the years were merged together with all unique rows of patients. One of the features present int he dataset is MQ220 which represents the questionnaire feature "Ever told you had cancer or malignancy". This feature is dropped from the feature set and made the target value for the predictive models described later. This feature is a direct result of whether the patient has cancer or not.

6 Data Preprocessing

Once there was a feature set in place, the next step was to apply pre-processing functions to make the data standardized to work with. The first step was to compute basic statistics of each feature to understand the data better. After computing the mean, variance, standard deviation, inter-quartile ranges, it was noticed that many values were 0 or missing. This implied that some imputation methods needed to be employed to discard such values. Some of the common pre-processing functions used were as follows:

6.1 One hot encoding : preproc_onehot, preproc_mode_onehot

One hot encoding is a process by which categorical variables are converted into "binarized" values of the category. Categorical data is data that takes only a limited number of values. The categorical value represents the numerical value of the entry in the dataset which are converted into limited values to represent each feature, while the others are zero.

6.2 Stardard Scalar : preproc_real

Normalizing function to impute the values by mean. The mean was calculated for each feature and subtracted from each feature and then divided by the standard deviation. This scales the values in a standard form throughout and makes the data normalized.

6.3 Min Max Scalar : preproc_min_max

Similar to the above function, this function is also used for normalizing the dataset. This estimator scales and translates each feature individually such that it is in the given range on the training set, between zero and one. That is, the maximum value in the dataset is set to 1 while the lowest value is set to 0 and all the other features are normalized in this range.

6.4 Scaling with Mode : preproc_mode

This function is the same as the Standard Scalar function except that instead of imputing the values by mean, it is imputed by the mode. This function is used since in many features with discrete values, having a fraction number (mean) does not make sense.

6.5 Basic Impute : preproc_impute

In this function, all the values that are missing from the dataset are replaced by the mean of that feature. Though this is a very basic technique, it gives better results than simply discarding those values.

6.6 Iterative imputer

Python has a package called fancy impute which provides a range of different imputing mechanisms. Each feature is modelled with missing values as a function of other features, and that is used to estimate the value for imputation. Since many features had Nan values and a lot of missing data, this preprocessing tool seemed promising.

6.7 Binning : preproc_bin

For continuous real value functions, the values are categorized into fixed value bins for a standard classification. The values are assigned to bins based on the linspace function which creates equidistant intervals between the maximum and minimum values of the range.

For different selected features, different pre-processing functions are chosen depending on the kind of values that the feature has. Categorical values are one-hot encoded. Features with lots of missing values are imputed by mean or mode depending on the feature. Continuous values were categorised by the binning function and made into discrete values for reasonable comparisons.
The iterative imputer was applied to a few of the features, but unfortunately didn't add much value to the analysis. This is probably because of the distinct nature of the features from each other. Perhaps performing them on highly correlated features alone would have given better results.
It was noticed that in most of the features, the values beyond a certain range were irrelevant. For example in the feature "Ever told by doctor have sleep disorder?" (SLQ050), values beyond 2 are "missing", "refused" and "don't know". These values are not relevant and do not need to be categorized into different classes as they don't add any information to the analysis. Hence, for such features, a cutoff was specified. In this example, the cutoff was set to 2, which meant that if a value was more than 2, the value was first set to nan and then based on the nature of the pre-processing function it was assigned a new value.
A list of some of the features chosen and their corresponding pre-processing functions are described in Table 1. The features are tabulated in accordance to the subcategory of data that they belong to in the order Demographics, Examination, Laboratory, Dietary followed by Questionnaire. (All the categories had numerous features, only the interesting ones are included here.)
After the pre-processed data is obtained the values are passed into different predictive models for further analysis.

7 Predictive Model

Predictive modeling uses statistics to predict outcomes. The event to be predicted is an unknown event, and based on other factors and features the value of the unknown event is found. In this case, the unknown event to be predicted is whether a person has cancer or not. The type of cancer is not relevant to this problem. The only goal is to predict if a patient has cancer or not. An extension of this problem is to predict whether the patient is likely to get cancer in the future based on their current health data available.

The ability of Machine Learning tools to detect key features from complex datasets make them an important part of predictive modelling . A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Thus it is evident that the use of Machine Learning methods can improve our understanding of cancer progression.
Two different approaches are considered for predictive modelling with respect to the dataset considered. The first approach is using the dataset as it is by splitting it into a training set of 66% and a test set of 33%. While this gives a more realistic analysis of the data, it doesn't give good results. This is primarily because of the skewed nature of the dataset. The second approach balances the data by choosing a subsample of the data and taking 10000 from the training phase and 5000 for the testing phase. It is ensured that while sampling, equal number of positive samples and negative samples are considered. Both the approaches were experimented upon, and the models were evaluated in terms of both the balanced and unbalanced datasets.
First the model is created based on a machine learning algorithm and trained on the dataset. After the model fits the data, it is used to predict on the test feature set. The predicted values are compared with the actual target values and the accuracy is computed based on how many are correctly predicted.

Table 1: Feature Set and Pre-processing functions

Code	Feature	Preprocessing	Reason
Demographics			
RIAGENDR	Gender	preproc_mode_onehot	Demographics are statistically useful as it represents the community and have inherent features which could have a direct or indirect affect on the patients.
RIDAGEYR	Age at time of screening	preproc_real	
RIDRETH1	Race/ethnicity	preproc_mode	
INDHHINC	Annual household income	preproc_real ('cutoff':11)	
DMDEDUC3	Education level	preproc_real ('cutoff':15)	
DMDBORN4	Country of birth	preproc_mode	
INDHHINC	Annual household income	preproc_real ('cutoff':11)	
Examination			
BMXBMI	BMI	preproc_real	Standard information for medical diagnosis
BMXWAIST	Waist	preproc_real	
BMXHT	Height	preproc_real	
BMXWT	Weight	preproc_real	
OHDEXSTS	Oral Health	preproc_real	Oral health has direct relation with oral cancer[2]
Laboratory			
LBXTC	Total Cholesterol	preproc_real	Cholesterol influences the growth of stem cells in the intestines, which in turn accelerates the rate of tumor leading to cancer. [3]
LBDWFL	Fluoride	preproc_real	Potential cause of osteosarcoma (bone cancer) [4]
LBDTHGSI	Blood	preproc_bin ('n bins':50)	Used for many medical tests.
LBDBPBSI	Lead, Cadmium, Total Mercury, Selenium , Manganese	preproc_real	Directly related to causing carsegenosis [5]
URXUMA	Urine	preproc_real	Used for many medical tests.
MCQ114	Lead Poisoning	preproc_impute	Studies show direct correlation [6]
Dietary			
DBD100	salt to food	preproc_mode_onehot ('cutoff':3)	Studies show as cause of gastric cancer [7]
DR2TSODI	Sodium (mg)	preproc_real	
DBD100	plain water drank	preproc_impute	Healthy practice
Questionnaire			
ALQ101	Alcohol consumption	preproc_mode ('cutoff':2)	Strong scientific evidence [8]
PAQ605	Vigorous work activity	preproc_mode	Healthy practice
MCQ160J	Doctor told overweight	preproc_mode('cutoff':2)	Scientific studies show correlation [9]
SMQ020	Smoking	preproc_mode_onehot	Scientific studies show correlation [10].
HIQ011	Health Insurance	preproc_mode('cutoff':2)	People with Insurance have Higher probability to get checked. Potential treatment affordability.
DIQ010	Diabetes	preproc_real ('cutoff':3)	Risk factor for cancer [11]
PAQ685	Bad air quality	preproc_real ('cutoff':3)	Possible cause of Lung cancer.
SLQ060	sleep disorder	preproc_real ('cutoff':2)	Relation of melatonin and cancer[12]
DEQ038G	Sunburn	preproc_mode	Possible cause of skin cancer.
BPQ020	Blood pressure	preproc_real ('cutoff':2)	Scientific evidence of causing breast and prostate cancer[13]

Accuracy alone isn't a complete measure of the performance of a model since the data is very skewed. Thus other measures such as precision, recall and f1 score which deal with true positives and true negatives are considered.

$$Recall = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

$$Precision = \frac{True\ Positives}{True\ Positives\ +\ False\ Positives}$$

In the next section, different predictive model architectures have been explored to predict if patients have cancer or not.

7.1 Logistic regression

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). This is the simplest predictive model to start analysis with. In linear regression, the outcome (dependent variable) is continuous, that is, it can have any one of an infinite number of possible values. Since we want a categorical outcome of either 0 or 1, we use a variant of this method, logistic Regression. Logistic regression is another generalized linear model procedure using the same basic formula, but instead of the continuous variable, it regresses for the probability of a categorical outcome. The equation to represent Logistic Regression is:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

The Logistic Regression function was used from the Python Sklearn Library of linear models. This function takes in a solver and the number of iterations as input. Various solvers such as 'newton-cg', 'sag', 'lbfgs' and 'liblinear' were tried, but the best results were obtained from using the lbfgs solver. The limited-memory BFGS is an optimization algorithm in the family of quasi-Newton methods that approximates the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm using a limited amount of computer memory. It is a popular algorithm for parameter estimation in machine learning and supports L2 regularization.

7.1.1 Evaluation and Results

This model was run for 200 iterations to fit the feature dataset. The classification report is as follows:

	Balanced Dataset	Unbalanced Dataset
Accuracy	0.7536369386464263	0.906311274509804

[H]

	precision	recall	f1-score	support
0	0.44	0.64	0.52	662
1	0.89	0.78	0.83	2500
micro avg	0.75	0.75	0.75	3162
macro avg	0.67	0.71	0.68	3162
weighted avg	0.80	0.75	0.77	3162

Figure 5: Balanced

	precision	recall	f1-score	support
0	0.44	0.01	0.01	1519
1	0.91	1.00	0.95	14801
micro avg	0.91	0.91	0.91	16320
macro avg	0.68	0.50	0.48	16320
weighted avg	0.86	0.91	0.86	16320

Figure 6: Unbalanced

[!htb]

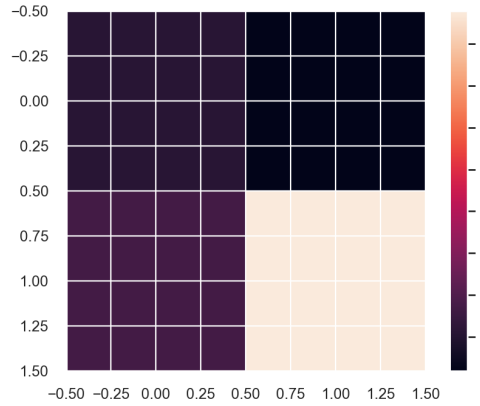

Figure 7: Balanced

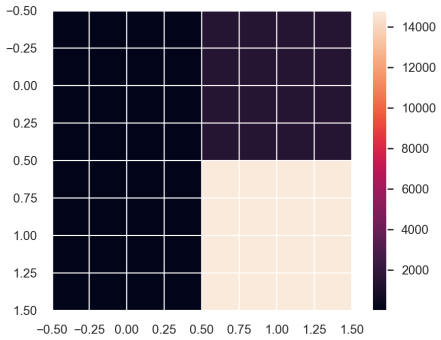

Figure 8: Unbalanced

[!htb]

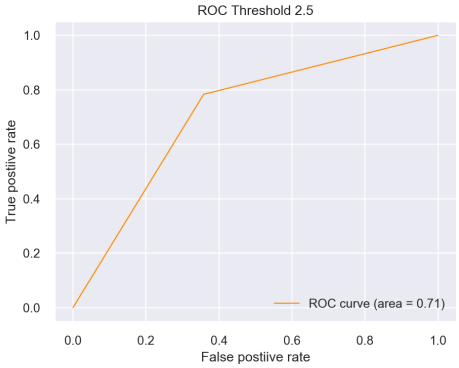

Figure 9: Balanced

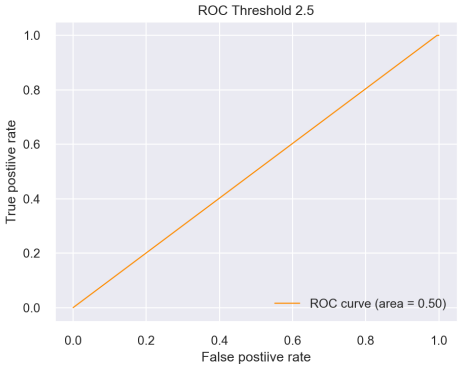

Figure 10: Unbalanced

Logistic Regression does well on the balanced dataset with high accuracy, precision and recall as it is able to classify the data into the two classes.

7.2 Recursive Feature Elimination (RFE)

Recursive Feature Elimination is based on the idea of repeatedly constructing a model and choosing either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features. This was performed after performing Logistic Regression and the values that resulted in False were omitted. This worked very well for a smaller subset of features and helped in the elimination process.

```
[ True  True  True False  True  True  True False  True  True False False
 False False  True False False False False False False False False False
 False  True  True False False False False False False False False False
 False False  True False False False False False  True False  True  True
 False False  True False False False False False False False False False
  True  True False False False  True  True False False False False False
 False False False False False False False False False False False False
 False False False False False]
[ 1  1  1 52  1  1  1  7  1  1 20 54  9 13  1  8 39 18 47 11 10 49 53 23
 64  1  1  6 46 35 51 50 63 68 67 16 44 14  1 27 19 41 38  3  1  2  1  1
 21 42  1 61 34 60 58 48 37 43 56 66  1  1 30 31  4  1  1 24 33 26  5 57
 45 28 70 29 36 40 22 17 65 15 32 59 12 69 62 55 25]
```

Figure 11: Output of Recursive Feature Elimination

7.3 Support Vector Machinesl

Support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap called the margin, that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the margin they fall into.
This was the next model of choice since it is a standard machine learning algorithm that works on the principle of maximizing the margin while optimizing on the dual form. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification as well, using kernel method, which implicitly maps the inputs into high-dimensional feature spaces. This feature is employed using various kernel methods. There are a number of kernel functions that are present, such as 'linear', 'rbf','poly' and 'sigmoid'. The rbf function has two parameters, gamma and C. The gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. The C parameter trades off correct classification of training examples against maximization of the decision function's margin. For larger values of C, a smaller margin is accepted if the decision function is better at classifying all training points correctly. A lower C encourages a larger margin, therefore a simpler decision function, at the cost of training accuracy. 'C' behaves as a regularization parameter in the SVM.

7.3.1 Evaluation and Results

The SVC function was used to implement this model using the Python Sklearn library of svm. Among all the kernel functions experimented upon, the 'rbf' kernel performed the best. The Radial basis function is defined as follows:

$$K(x, x') = exp(-\frac{(x - x')^2}{2\sigma^2})$$

This is probably because RBF kernel can choose a non-linear decision boundary which is able to classify better since the data isn't linearly separable. After evaluation, the results obtained were as follows:

	Balanced Dataset	Unbalanced Dataset
Accuracy	0.7383059418457648	0.9069240196078432

7.4 K-Means

K-means is a clustering algorithm that aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean. Though this is an unsupervised algorithm,

it was tried to see if the algorithm is able to identify patterns in the data to separate it out into different clusters. The number of clusters to form as well as the number of centroids to generate is a hyper-parameter that was experimented with.

7.4.1 Evaluation and Results

This algorithm was run only on the unbalanced dataset to see it was able to identify patterns. After trying different values for k, 3 clusters gave the best results. The accuracy was relatively low with a value of 0.67, but interesting results were seen from the confusion matrix.

7.5 K Nearest Neighbours

The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space. For k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

7.5.1 Evaluation and Results

After trying different values for the number of neighbors, the optimum value obtained was for 5 nearest neighbors.

	Balanced Dataset	Unbalanced Dataset
Accuracy	0.6698292220113852	0.8983455882352941

7.6 Random Forests

Random forests or random decision forests are an ensemble learning method for classification. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

7.6.1 Evaluation and Results

The number of trees used as estimators for a decision tree were experimented with to prevent overfitting. The best values were obtained with 10 estimators.

	Balanced Dataset	Unbalanced Dataset
Accuracy	0.7223276407337128	0.9068014705882353

7.7 Gaussian Naive Bayes

Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states that

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Though there is no prior here, we can use the information from the features to predict what the outcome may be.

7.7.1 Evaluation and Results

Though the accuracy of the classifier is not good. It does better than the others in terms of precision and recall. Most of the other classifiers have a high false negative error rate. But this one doesn't

predict people to not have cancer when they actually do. This can be directly inferred from the confusion matrix.

	Balanced Dataset	Unbalanced Dataset
Accuracy	0.692283364958887	0.6667892156862745

	precision	recall	f1-score	support
0	0.32	0.76	0.45	662
1	0.90	0.58	0.71	2500
micro avg	0.62	0.62	0.62	3162
macro avg	0.61	0.67	0.58	3162
weighted avg	0.78	0.62	0.65	3162

Figure 12: Balanced

	precision	recall	f1-score	support
0	0.17	0.69	0.28	1528
1	0.95	0.66	0.78	14792
micro avg	0.67	0.67	0.67	16320
macro avg	0.56	0.68	0.53	16320
weighted avg	0.88	0.67	0.74	16320

Figure 13: Unbalanced

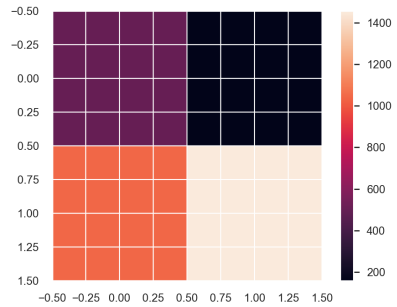

Figure 14: Balanced

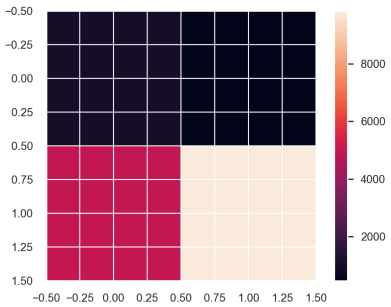

Figure 15: Unbalanced

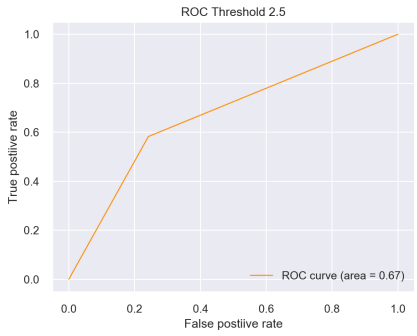

Figure 16: Balanced

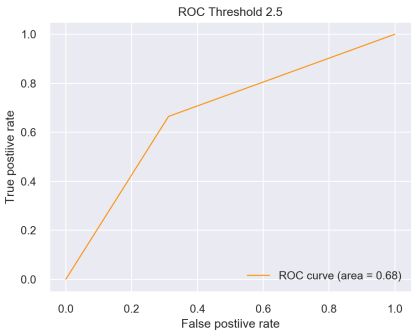

Figure 17: Unbalanced

This is the most interesting model, as it does equally well or poorly on both the balanced and unbalanced datasets. Also it has the highest recall among all the models though its precision is very poor. Thus it is always a trade-off.

7.8 Adaboost

Adaboost stands for Adaptive Boosting. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one by combining them together.Each weak classifier is associated with a corresponding weight. This algorithm is the weighted combination of weak classifiers. For any classifier with accuracy higher than 50%, the weight is positive. The more accurate the classifier,

the larger the weight. While for the classifier with less than 50% accuracy, the weight is negative. It means that we combine its prediction by flipping the sign. For example, one can turn a classifier with 40% accuracy into 60% accuracy by flipping the sign of the prediction. Thus even though the classifier performs worse than random guessing, it still contributes to the final prediction. Classifiers with exact 50% accuracy are not considered, as they don't add any information and thus contribute nothing to the final prediction.
The classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. The equation is given by:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{1}$$

where T is the number of classifiers.

7.8.1 Evaluation and Results

Different classifiers were experimented upon as base estimator and the algorithm is run for 50 estimators after which boosting is terminated. The default classifier is the Decision tree classifier which was seen to outperform the other estimators such as Logistic Regression and SVM. The classification report is as follows:

	Balanced Dataset	Unbalanced Dataset
Accuracy	0.7647058823529411	0.9047794117647059

	precision	recall	f1-score	support
0	0.46	0.65	0.54	662
1	0.90	0.80	0.84	2500
micro avg	0.76	0.76	0.76	3162
macro avg	0.68	0.72	0.69	3162
weighted avg	0.80	0.76	0.78	3162

Figure 18: Balanced

	precision	recall	f1-score	support
0	0.33	0.02	0.03	1528
1	0.91	1.00	0.95	14792
micro avg	0.90	0.90	0.90	16320
macro avg	0.62	0.51	0.49	16320
weighted avg	0.85	0.90	0.86	16320

Figure 19: Unbalanced

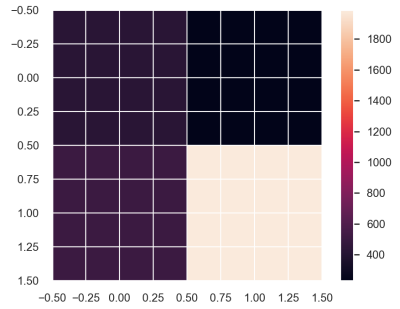

Figure 20: Balanced

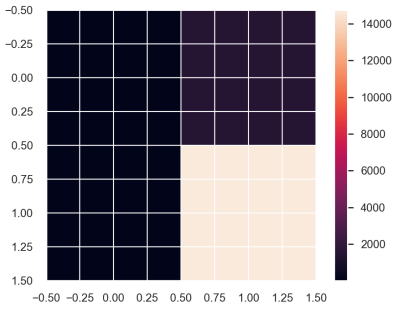

Figure 21: Unbalanced

7.9 Neural Networks

After experimenting upon all the classical machine learning algorithms neural networks were tried. Neural Networks are inspired by the biological neural networks that constitute animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. A NN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in

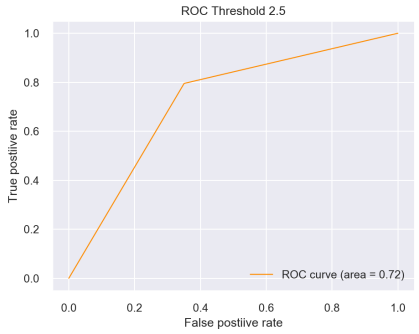

Figure 22: Balanced

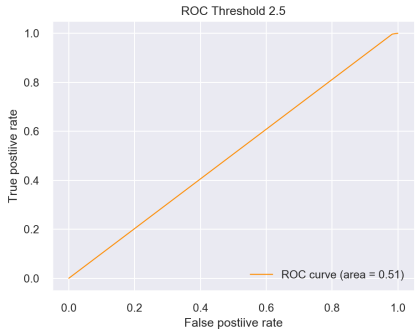

Figure 23: Unbalanced

a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. Different nodes learn different features of the data and the model is known to perform very well, given a vast amount of data.
The MLPClassifier function from the sklearn library in Python, was used to compute this model. The model architecture comprised of four hidden layers with 500,200,200 and 500 neurons respectively. The activation function used was relu. This is a linear function called Rectilinear Unit which is known to perform very well for neural networks. The model computed the binary cross entropy loss since the problem being tackled is binary classification. The adam optimizer was used to optimize the loss function. This optimizer is a combination on the adagrad and momentum optimzing functions.

7.9.1 Evaluation and Results

All the hyperparameters of the neural network were hypertuned and the best results were obtained by having four hidden layer. The learning rate used was 5e-7 and it was run for 100 iterations. The Neural network seemed to do pretty well as compared to the classical models.

	Balanced Dataset	Unbalanced Dataset
Accuracy	0.74193548387096771	0.926924019607843

	precision	recall	f1-score	support
0	0.39	0.42	0.41	662
1	0.84	0.83	0.84	2500
micro avg	0.74	0.74	0.74	3162
macro avg	0.62	0.62	0.62	3162
weighted avg	0.75	0.74	0.75	3162

Figure 24: Balanced

	precision	recall	f1-score	support
0	0.44	0.01	0.01	1519
1	0.91	1.00	0.95	14801
micro avg	0.91	0.91	0.91	16320
macro avg	0.68	0.50	0.48	16320
weighted avg	0.86	0.91	0.86	16320

Figure 25: Unbalanced

Another Neural network architecture was also tried, which was written in Keras with Tensorflow back-end. This model was built using multiple dense layers with the relu activation function. The loss function used was binary cross entropy and the adam optimizer was used to optimize the loss function. This model was run for 100 epochs. The network architecture was as follows:

Layer (type)	Output Shape	Param #
dense_28 (Dense)	(None, 64)	5760
dense_29 (Dense)	(None, 32)	2080
dense_30 (Dense)	(None, 16)	528
dense_31 (Dense)	(None, 1)	17

Total params: 8,385
Trainable params: 8,385
Non-trainable params: 0

Figure 30: Neural Network Architecture.

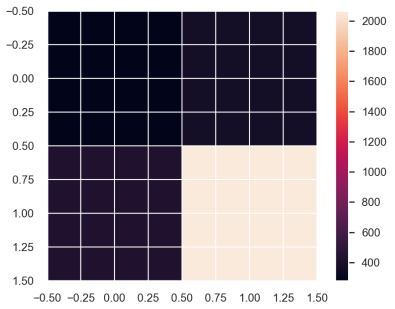

Figure 26: Balanced

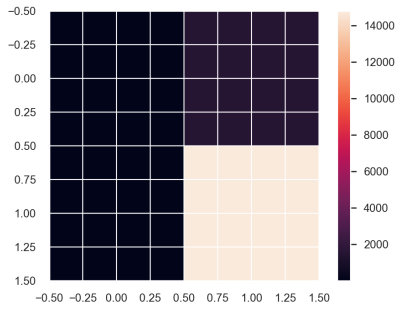

Figure 27: Unbalanced

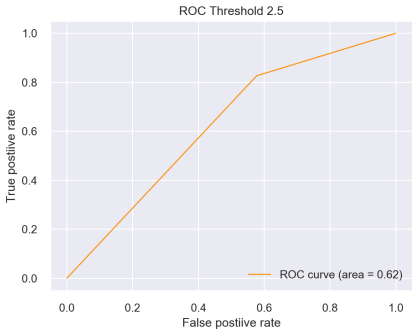

Figure 28: Balanced

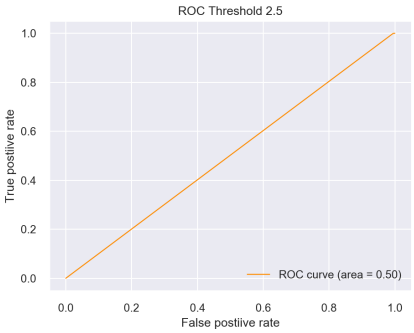

Figure 29: Unbalanced

The metrics of this model was very poor. At each step, the loss as well as the accuracy remained constant. This implies that the model was not learning anything and would eventually result in complete prediction of one class only. This is because the data is so skewed that the model is only able to learn the features of the negative class which outnumber the positive class by a huge margin. Thus even if it predicts all the values as the negative class, it gets an accuracy of close to 99%. But the extremely poor precision and recall values show that the model is only able to learn one class as there isn't enough information about both.

8 Analysis of the Models

After running all the predictive models, it is unfortunate that no model emerges as an absolute winner. This is not surprising, since the nature of the dataset was such, that the positive and negative samples were very unbalanced. there is a trade-off between accuracy, precision and recall.
On the balanced dataset, which comprises of equal number of positive and negative samples, all the models have an average accuracy value lying in the 67-78% range. The precision and recall are all average. It tends to have a higher value for class 1 (people without cancer). The precion and recall values are seen to be around 70-80% for class 1 and around 30-70% for class 0 (people with cancer). The Adaboost classifier with Decision Trees as estimators perform the best with the highest accuracy of 76%. This method creates an ensemble of different trees which learn different features and combines their values for a better prediction. Since the dataset is balanced, it learns various different features from both classes and is able to outperform the other classifiers.

For the unbalanced dataset, which is the true representation of the actual dataset most of the classical algorithms perform very similarly. They give high accuracies but have very low precision and recall values. The trade-off is evident since the classifier ends up learning the features relevant to only one class and predicts everything to be of the same class. This results in the high accuracy. Most of them have very high false negative and false positive rates, as is evident from the confusion matrix. The

K-nearest neighbours classifier does relatively better in terms of recall than the other classifiers. The Gaussian Naive Bayes classifier gives a poor accuracy, but has the best precision and recall. The commendable aspect is that the false positive rate is much lower. That is the people who have cancer but are not predicted to have it. This is a very dangerous scenario. False negatives are still better as they are only a false alarm, and patients can be screened regardless. The Gaussian NB has a lower false positive rate which shows that it is probably a better algorithm of choice.

9 Conclusion

Prediction is ubiquitous in oncology: innumerable decisions by patients, family members, oncologists and other care providers depend on assessing the likelihood of future events. This can be seen across the spectrum of cancer care, from screening to hospice. Screening is recommended for those at elevated risk of cancer, either because of age or other risk factors, such as lung imaging in those with a significant smoking history. Recent years have seen many attempts to formalize risk prediction in cancer. In place of informal and implicit prediction algorithms, such as cancer stage, many statistical prediction tools have been developed that provide a quantitative estimate of the probability of a specific event for an individual patient. In this project, a similar such analysis was attempted. Prediction modeling in cancer is a huge research area, and the enormity of the problem and lack of enough evidence makes this a very interesting as well as relevant one. While this paper has attempted to show an overview of different kinds of algorithms and their performances on the NHANES dataset, research continues in this field. Understanding the nuances of feature selection and data preprocessing for every feature can still be improved upon. Designing Neural Networks with more complex architectures that will be able to learn different kinds of features and hypertuning them to fit the data given remains as a future task.

10 Acknowledgements

I'd sincerely like to thank Professor Majid Sarrafzadeh for giving me the opportunity to work on this project. Working with a real dataset to predict such an important disease is really hard, and with the help of the Centre for Disease control and Prevention (CDC) and National Health and Nutrition Examination Survey (NHANES) we were able to study this problem. Big thank you to Orpaz Orenstein for all his help with the project. I would also like to acknowledge Dr Indranil Ghosh from Apollo, India, a leading oncologist and a close relative, who helped me understand the different features better validated my feature selection in the initial stages of the project.[**?**]

11 References

References

[1] National Health and Nutrition Examination Survey. `https://www.cdc.gov/nchs/nhanes/index.htm`.

[2] `https://www.webmd.com/oral-health/news/20130821/poor-oral-hygiene-tied-to-cancer-linked-virus-study-finds#1`, 2017.

[3] `https://www.medicalnewstoday.com/articles/320740.php`, Jan 2018.

[4] The American Cancer Society medical and editorial content team. `https://www.cancer.org/cancer/cancer-causes/water-fluoridation-and-cancer-risk.html`, July 2015.

[5] George Kazantzis. Role of cobalt, iron, lead, manganese, mercury, platinum, selenium, and titanium in carcinogenesis. *Environmental Health Perspectives*, 40:143–161, 1981.

[6] Herbert Needleman. Lead poisoning. *Annu. Rev. Med.*, 55:209–222, 2004.

[7] `https://doi.org/10.1111/j.1349-7006.2005.00006.x`, Jan 2005.

[8] https://www.cancer.gov/about-cancer/causes-prevention/risk/alcohol/alcohol-fact-sheet, Sep 2018.

[9] Eugenia E Calle, Carmen Rodriguez, Kimberly Walker-Thurmond, and Michael J Thun. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of us adults. *New England Journal of Medicine*, 348(17):1625–1638, 2003.

[10] Sara Gandini, Edoardo Botteri, Simona Iodice, Mathieu Boniol, Albert B Lowenfels, Patrick Maisonneuve, and Peter Boyle. Tobacco smoking and cancer: A meta-analysis. *International journal of cancer*, 122(1):155–164, 2008.

[11] Paolo Vigneri, Francesco Frasca, Laura Sciacca, Giuseppe Pandini, and Riccardo Vigneri. Diabetes and cancer. *Endocrine-related cancer*, 16(4):1103–1123, 2009.

[12] https://www.nature.com/articles/6604425.

[13] Linda E Carlson, Michael Speca, Peter Faris, and Kamala D Patel. One year pre–post intervention follow-up of psychological, immune, endocrine and blood pressure outcomes of mindfulness-based stress reduction (mbsr) in breast and prostate cancer outpatients. *Brain, behavior, and immunity*, 21(8):1038–1049, 2007.