December 2021

An Interactive Health Data Science Platform for Exploratory Analysis of Health Outcomes – a Case Study with Colon Cancer

Hemanth Kumar Alapati
University of Wisconsin-Milwaukee

AN INTERACTIVE HEALTH DATA SCIENCE PLATFORM FOR EXPLORATORY ANALYSIS OF HEALTH OUTCOMES – A CASE STUDY WITH COLON CANCER

by

Hemanth Kumar Alapati

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Computer Science

at

The University of Wisconsin-Milwaukee

December 2021

ABSTRACT

AN INTERACTIVE HEALTH DATA SCIENCE PLATFORM FOR EXPLORATORY ANALYSIS OF HEALTH OUTCOMES – A CASE STUDY WITH COLON CANCER

by
Hemanth Kumar Alapati

The University of Wisconsin-Milwaukee, 2021
Under the Supervision of Professor Jake Luo and Professor Mukul Goyal

Disease prediction is an important aspect of early disease detection and preventive care with wide range of applications in healthcare domain. Previous studies used image processing techniques, statistical and machine learning models to predict diseases. Prediction accuracies vary with data type and the target. Often the data is processed through models under different data conditions to identify what works best for a scenario. This results in tweaking the code, running multiple iterations making these methods usable only for people with technical skills. An interactive platform is developed that hides the technicalities and allows the users to change options like target disease for prognosis, feature selection method, sample size, ML algorithm. With this, multiple approaches can be tried and compared to find a combination of the options for an efficient outcome. Colon cancer is used to perform a case study to test this platform. 2 selection algorithms and 3 ML models are used. Although both selection methods identified identical features as significant for colon cancer prediction, the order of the features based on the scores is different. Hence, the machine learning algorithms performed similarly with both the selection methods. Random Forest, Logistic Regression, and Decision Tree had accuracies 87%, 86%, and 83% respectively.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction

Cancer is a broad term that refers to uncontrolled aberrant growth of human cells. Our body consists of trillions of cells. These cells normally expand and multiply to generate new cells as needed. This ordered process can sometimes break down, resulting in abnormal or damaged cells growth which can result in cancer. Colon cancer is a form of cancer where such uncontrolled cell growth is observed in the large intestine (colon) which is the final part of the digestive tract [1]. Colon cancer is also referred as Colorectal cancer which includes rectal cancer as well which starts in the rectum.

Colorectal cancer is the 3rd most diagnosed cancer in America excluding skin cancers [2]. **Figure 1** shows the number of new colorectal cancer cases detected in America during the years 1999 to 2018 [3].

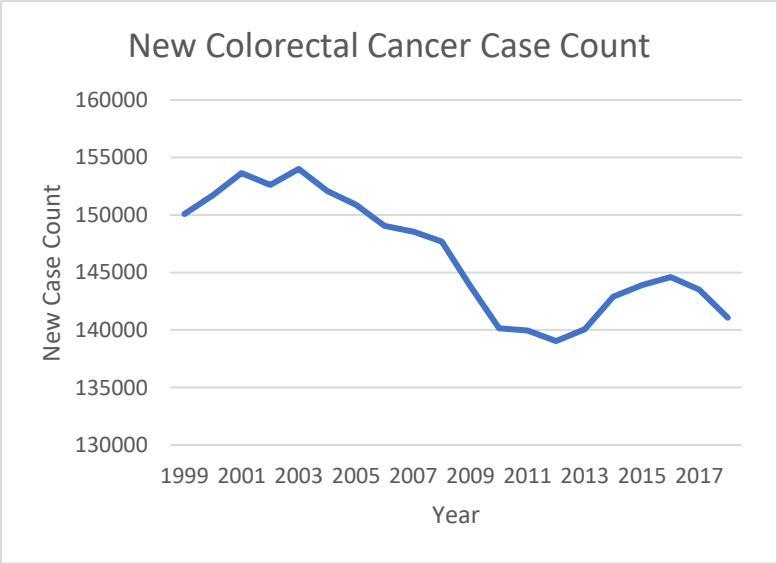

Figure 1: Annual number of new colorectal cancer cases, USA, 1999-2018 [3].

Some of the symptoms for colon cancer include a change in bowel habits, such as diarrhea or constipation, rectal bleeding, consistent stomach pain, such as cramps, gas, or bloating, a feeling as if the bowels don't empty completely, weakness or exhaustion, unexplained

weight loss. The symptoms vary widely among patients and only a few patients experience these symptoms at an early stage [1]. Colon cancer treatment uses surgery, chemotherapy, radiation therapy, targeted drug therapy, immunotherapy, supportive (palliative) care or a combination of these [4].

This dissertation uses healthcare data obtained from Nationwide Inpatient Sample (NIS) Dataset [5], and applies Machine Learning models like Decision Tree, Logistic Regression, and Random Forest to predict colon cancer. It also identifies top features that affect the prediction using the feature selection methods like Select K Best and Feature Importance methods. A generic disease prediction framework has been built with a UI component for user interaction making it accessible to broad category of users. The UI allows users to choose the amount of data that needs to be analyzed, percentage of this data used for training and testing purposes, number of features that need to be identified, type of feature selection algorithm that need to be applied to identify the features, machine learning algorithm that need to be applied on the selected data and features for the prediction.

Study Proposal

Create a platform that helps with Exploratory Data Analysis of health data. The platform needs to be interactive without exposing the technicalities so that it is available for wide range of users. It should be flexible to be able to try and test different variations of data, algorithms without changing the code. It should make the comparative analysis easy by giving the flexibility to switch between algorithms, data. It should be generic so that it can be used for prognosis of multiple diseases.

Below are the flexible features:

- Select diseases
- Number of positive cases used for analysis
- Number of negative cases used for analysis
- Number of features that need to be identified
- Select feature selection algorithms
- Select the Machine Learning model
- Percentage of the selected data used for training the model

Based on the selections made, the platform should process the data, analyze it with selected algorithms and give the feature and prediction accuracy information.

Output of the platform should have the below details:

- Features that are highly correlated with the selected disease
- Segregate the features that have positive or inverse correlation with the disease.

Use this platform to study Colon cancer prognosis. Identify the features that have positive and inverse correlation with Colon cancer. Study the research done on the correlation of these features with colon cancer. Do a comparative study of the selection algorithms, and Machine Learning algorithms. Identify which model works better for Colon cancer prediction.

Related work

Disease diagnosis

Early usage of machine learning in health care domain started with disease diagnosis. Some studies talked about the advancements around Machine Learning and showed how a variety of these models can be used in disease diagnosis [14]. Techniques like Image Processing, Artificial Neural Networks, Bayesian Networks, and Machine Learning were used in these studies. Most of these studies analyzed the specimen from patients using machine learning techniques, identified patterns for diagnosis of various diseases. These have helped in early identification of serious diseases; there by enabling early treatment of these diseases and improving the chances of cure.

Disease prognosis

As health care evolved, more patient data is digitized. Health data, storage & computing power availability and advancement in Machine Learning has enable studies on prediction of diseases by analyzing patient health data using Machine Learning techniques [15]. Most of these studies did comparative analysis on the data by applying various methods and identifying the efficient methods [16, 17]. These studies have helped in identifying groups of people that are at more risk of getting diseases. This segregation helped in giving preventive health services for individuals who are at risk for a disease. This also led to early identification of diseases based on the at-risk segregation and being proactive.

Colon cancer prognosis

Colon cancer is an area of interest for researchers as it is one of the top cancer types with significant number of cases. Machine Learning algorithms have been used to predict the stages

of colon cancer based on pathological test results [18]. There were studies about building the prognosis predictor using the gene [19], serum [20] samples from colon cancer patients.

Methodology

Data sources and data description

Healthcare Cost and Utilization Project (HCUP) developed one of the largest publicly available databases called National Inpatient Sample (NIS) database to store admission level healthcare information. Sponsored by Agency for Healthcare Research and Quality (AHRQ), NIS database stores about 7 million patient history every year since 1988. The 2016 dataset consists of three ASCII files: 'Core File', 'Hospital Weights File' and 'Severity Measures File' and has a total size of 15GB. In this dissertation, the 'Core File' data is used. 'File Specification' explains how the data elements are organized in the 'Core File'. **Figure 1** and **Figure 2** show the 'File Specification' file. It includes information such as database name, discharge year of data, file name, data element number, data element name, starting and ending column of data element, data element type, data element label, etc.

The 2016 NIS core data file has 7,135,090 records. For every record, there are 98 data elements which can be split into two categories: non-clinical and clinical. Non-clinical data includes demographic information of the patient (age, sex, race), date of admission, total cost, zip code, hospital ID, length of stay, etc. Treatment types, procedures, diagnosis categories, diagnosis codes, etc. are some of the clinical data. Each entry lists a maximum of 30 diagnosis codes that represent disease conditions the patient has history of, which are one of the most impactful data elements. The 2016 database represents these disease codes is ICD-10 (International Classification of Diseases, 10th revision) format. The WHO (World Health

Organization) designed these codes so that every disease has a unique code with a view to help healthcare personnel, insurance companies and concerned parties to specify health conditions in a uniformed manner. The 2016 dataset may include up to 69,823 diagnosis codes (ICD-10-CM).

Figure 2 and Figure 3 show the print of NIS 2016 core file layout:

FileSpecifications_NIS_2016_Core - Notepad

File Edit Format View Help

```
Data Set Name: NIS_2016_CORE
Number of Observations: 7135090
Total Record Length: 497
Total Number of Data Elements: 98


Columns   Description
=======   ===========
 1-  3    Database name
 5-  8    Discharge year of data
10- 35    File name
37- 39    Data element number
41- 69    Data element name
71- 73    Starting column of data element in ASCII file
75- 77    Ending column of data element in ASCII file
79- 79    Non-zero number of digits after decimal point for numeric data element
81- 84    Data element type (Num=numeric; Char=character)
86-185    Data element label


NIS 2016 NIS_2016_Core          1 AGE                        1   3   Num  Age in years at admission
NIS 2016 NIS_2016_Core          2 AGE_NEONATE                4   5   Num  Neonatal age (first 28 days after birth) indicator
NIS 2016 NIS_2016_Core          3 AMONTH                     6   7   Num  Admission month
NIS 2016 NIS_2016_Core          4 AWEEKEND                   8   9   Num  Admission day is a weekend
NIS 2016 NIS_2016_Core          5 DIED                      10  11   Num  Died during hospitalization
NIS 2016 NIS_2016_Core          6 DISCWT                    12  22 7 Num  NIS discharge weight
NIS 2016 NIS_2016_Core          7 DISPUNIFORM               23  24   Num  Disposition of patient (uniform)
NIS 2016 NIS_2016_Core          8 DQTR                      25  26   Num  Discharge quarter
NIS 2016 NIS_2016_Core          9 DRG                       27  29   Num  DRG in effect on discharge date
NIS 2016 NIS_2016_Core         10 DRGVER                    30  31   Num  DRG grouper version used on discharge date
NIS 2016 NIS_2016_Core         11 DRG_NoPOA                 32  34   Num  DRG in use on discharge date, calculated without POA
NIS 2016 NIS_2016_Core         12 DXVER                     35  36   Num  Diagnosis Version
NIS 2016 NIS_2016_Core         13 ELECTIVE                  37  38   Num  Elective versus non-elective admission
NIS 2016 NIS_2016_Core         14 FEMALE                    39  40   Num  Indicator of sex
NIS 2016 NIS_2016_Core         15 HCUP_ED                   41  43   Num  HCUP Emergency Department service indicator
NIS 2016 NIS_2016_Core         16 HOSP_DIVISION             44  45   Num  Census Division of hospital
NIS 2016 NIS_2016_Core         17 HOSP_NIS                  46  50   Num  NIS hospital number
NIS 2016 NIS_2016_Core         18 I10_DX1                   51  57   Char ICD-10-CM Diagnosis 1
NIS 2016 NIS_2016_Core         19 I10_DX2                   58  64   Char ICD-10-CM Diagnosis 2
NIS 2016 NIS_2016_Core         20 I10_DX3                   65  71   Char ICD-10-CM Diagnosis 3
NIS 2016 NIS_2016_Core         21 I10_DX4                   72  78   Char ICD-10-CM Diagnosis 4
NIS 2016 NIS_2016_Core         22 I10_DX5                   79  85   Char ICD-10-CM Diagnosis 5
NIS 2016 NIS_2016_Core         23 I10_DX6                   86  92   Char ICD-10-CM Diagnosis 6
NIS 2016 NIS_2016_Core         24 I10_DX7                   93  99   Char ICD-10-CM Diagnosis 7
NIS 2016 NIS_2016_Core         25 I10_DX8                  100 106   Char ICD-10-CM Diagnosis 8
NIS 2016 NIS_2016_Core         26 I10_DX9                  107 113   Char ICD-10-CM Diagnosis 9
NIS 2016 NIS_2016_Core         27 I10_DX10                 114 120   Char ICD-10-CM Diagnosis 10
NIS 2016 NIS_2016_Core         28 I10_DX11                 121 127   Char ICD-10-CM Diagnosis 11
NIS 2016 NIS_2016_Core         29 I10_DX12                 128 134   Char ICD-10-CM Diagnosis 12
NIS 2016 NIS_2016_Core         30 I10_DX13                 135 141   Char ICD-10-CM Diagnosis 13
NIS 2016 NIS_2016_Core         31 I10_DX14                 142 148   Char ICD-10-CM Diagnosis 14
NIS 2016 NIS_2016_Core         32 I10_DX15                 149 155   Char ICD-10-CM Diagnosis 15
NIS 2016 NIS_2016_Core         33 I10_DX16                 156 162   Char ICD-10-CM Diagnosis 16
NIS 2016 NIS_2016_Core         34 I10_DX17                 163 169   Char ICD-10-CM Diagnosis 17
NIS 2016 NIS_2016_Core         35 I10_DX18                 170 176   Char ICD-10-CM Diagnosis 18
NIS 2016 NIS_2016_Core         36 I10_DX19                 177 183   Char ICD-10-CM Diagnosis 19
NIS 2016 NIS_2016_Core         37 I10_DX20                 184 190   Char ICD-10-CM Diagnosis 20
NIS 2016 NIS_2016_Core         38 I10_DX21                 191 197   Char ICD-10-CM Diagnosis 21
NIS 2016 NIS_2016_Core         39 I10_DX22                 198 204   Char ICD-10-CM Diagnosis 22
NIS 2016 NIS_2016_Core         40 I10_DX23                 205 211   Char ICD-10-CM Diagnosis 23
```

Figure 2: File specification of NIS_2016_Core data (part 1)

NIS 2016	NIS_2016_Core	41	I10_DX24	212 218	Char	ICD-10-CM Diagnosis 24
NIS 2016	NIS_2016_Core	42	I10_DX25	219 225	Char	ICD-10-CM Diagnosis 25
NIS 2016	NIS_2016_Core	43	I10_DX26	226 232	Char	ICD-10-CM Diagnosis 26
NIS 2016	NIS_2016_Core	44	I10_DX27	233 239	Char	ICD-10-CM Diagnosis 27
NIS 2016	NIS_2016_Core	45	I10_DX28	240 246	Char	ICD-10-CM Diagnosis 28
NIS 2016	NIS_2016_Core	46	I10_DX29	247 253	Char	ICD-10-CM Diagnosis 29
NIS 2016	NIS_2016_Core	47	I10_DX30	254 260	Char	ICD-10-CM Diagnosis 30
NIS 2016	NIS_2016_Core	48	I10_ECAUSE1	261 267	Char	ICD-10-CM External cause 1
NIS 2016	NIS_2016_Core	49	I10_ECAUSE2	268 274	Char	ICD-10-CM External cause 2
NIS 2016	NIS_2016_Core	50	I10_ECAUSE3	275 281	Char	ICD-10-CM External cause 3
NIS 2016	NIS_2016_Core	51	I10_ECAUSE4	282 288	Char	ICD-10-CM External cause 4
NIS 2016	NIS_2016_Core	52	I10_NDX	289 290	Num	ICD-10-CM Number of diagnoses on this record
NIS 2016	NIS_2016_Core	53	I10_NECAUSE	291 293	Num	ICD-10-CM Number of External cause codes on this record
NIS 2016	NIS_2016_Core	54	I10_NPR	294 295	Num	ICD-10-PCS Number of procedures on this record
NIS 2016	NIS_2016_Core	55	I10_PR1	296 302	Char	ICD-10-PCS Procedure 1
NIS 2016	NIS_2016_Core	56	I10_PR2	303 309	Char	ICD-10-PCS Procedure 2
NIS 2016	NIS_2016_Core	57	I10_PR3	310 316	Char	ICD-10-PCS Procedure 3
NIS 2016	NIS_2016_Core	58	I10_PR4	317 323	Char	ICD-10-PCS Procedure 4
NIS 2016	NIS_2016_Core	59	I10_PR5	324 330	Char	ICD-10-PCS Procedure 5
NIS 2016	NIS_2016_Core	60	I10_PR6	331 337	Char	ICD-10-PCS Procedure 6
NIS 2016	NIS_2016_Core	61	I10_PR7	338 344	Char	ICD-10-PCS Procedure 7
NIS 2016	NIS_2016_Core	62	I10_PR8	345 351	Char	ICD-10-PCS Procedure 8
NIS 2016	NIS_2016_Core	63	I10_PR9	352 358	Char	ICD-10-PCS Procedure 9
NIS 2016	NIS_2016_Core	64	I10_PR10	359 365	Char	ICD-10-PCS Procedure 10
NIS 2016	NIS_2016_Core	65	I10_PR11	366 372	Char	ICD-10-PCS Procedure 11
NIS 2016	NIS_2016_Core	66	I10_PR12	373 379	Char	ICD-10-PCS Procedure 12
NIS 2016	NIS_2016_Core	67	I10_PR13	380 386	Char	ICD-10-PCS Procedure 13
NIS 2016	NIS_2016_Core	68	I10_PR14	387 393	Char	ICD-10-PCS Procedure 14
NIS 2016	NIS_2016_Core	69	I10_PR15	394 400	Char	ICD-10-PCS Procedure 15
NIS 2016	NIS_2016_Core	70	KEY_NIS	401 410	Num	NIS record number
NIS 2016	NIS_2016_Core	71	LOS	411 415	Num	Length of stay (cleaned)
NIS 2016	NIS_2016_Core	72	MDC	416 417	Num	MDC in effect on discharge date
NIS 2016	NIS_2016_Core	73	MDC_NoPOA	418 419	Num	MDC in use on discharge date, calculated without POA
NIS 2016	NIS_2016_Core	74	NIS_STRATUM	420 423	Num	NIS hospital stratum
NIS 2016	NIS_2016_Core	75	PAY1	424 425	Num	Primary expected payer (uniform)
NIS 2016	NIS_2016_Core	76	PL_NCHS	426 428	Num	Patient Location: NCHS Urban-Rural Code
NIS 2016	NIS_2016_Core	77	PRDAY1	429 431	Num	Number of days from admission to I10_PR1
NIS 2016	NIS_2016_Core	78	PRDAY2	432 434	Num	Number of days from admission to I10_PR2
NIS 2016	NIS_2016_Core	79	PRDAY3	435 437	Num	Number of days from admission to I10_PR3
NIS 2016	NIS_2016_Core	80	PRDAY4	438 440	Num	Number of days from admission to I10_PR4
NIS 2016	NIS_2016_Core	81	PRDAY5	441 443	Num	Number of days from admission to I10_PR5
NIS 2016	NIS_2016_Core	82	PRDAY6	444 446	Num	Number of days from admission to I10_PR6
NIS 2016	NIS_2016_Core	83	PRDAY7	447 449	Num	Number of days from admission to I10_PR7
NIS 2016	NIS_2016_Core	84	PRDAY8	450 452	Num	Number of days from admission to I10_PR8
NIS 2016	NIS_2016_Core	85	PRDAY9	453 455	Num	Number of days from admission to I10_PR9
NIS 2016	NIS_2016_Core	86	PRDAY10	456 458	Num	Number of days from admission to I10_PR10
NIS 2016	NIS_2016_Core	87	PRDAY11	459 461	Num	Number of days from admission to I10_PR11
NIS 2016	NIS_2016_Core	88	PRDAY12	462 464	Num	Number of days from admission to I10_PR12
NIS 2016	NIS_2016_Core	89	PRDAY13	465 467	Num	Number of days from admission to I10_PR13
NIS 2016	NIS_2016_Core	90	PRDAY14	468 470	Num	Number of days from admission to I10_PR14
NIS 2016	NIS_2016_Core	91	PRDAY15	471 473	Num	Number of days from admission to I10_PR15
NIS 2016	NIS_2016_Core	92	PRVER	474 475	Num	Procedure Version
NIS 2016	NIS_2016_Core	93	RACE	476 477	Num	Race (uniform)
NIS 2016	NIS_2016_Core	94	TOTCHG	478 487	Num	Total charges (cleaned)
NIS 2016	NIS_2016_Core	95	TRAN_IN	488 489	Num	Transfer in indicator
NIS 2016	NIS_2016_Core	96	TRAN_OUT	490 491	Num	Transfer out indicator
NIS 2016	NIS_2016_Core	97	YEAR	492 495	Num	Calendar year
NIS 2016	NIS_2016_Core	98	ZIPINC_QRTL	496 497	Num	Median household income national quartile for patient ZIP Code

Figure 3: File specification of NIS_2016_Core data (part 2)

An ICD10CM code – description file has been used to translate ICD10 codes to their short descriptions. **Figure 4** represents the sample file and **Table 1** has the file layout information.

icd10cm_order_2018 - Notepad

File Edit Format View Help

```
00001 A00     0 Cholera                                                Cholera
00002 A000    1 Cholera due to Vibrio cholerae 01, biovar cholerae     Cholera due to Vibrio cholerae 01, biovar cholerae
00003 A001    1 Cholera due to Vibrio cholerae 01, biovar eltor        Cholera due to Vibrio cholerae 01, biovar eltor
00004 A009    1 Cholera, unspecified                                   Cholera, unspecified
00005 A01     0 Typhoid and paratyphoid fevers                         Typhoid and paratyphoid fevers
00006 A010    0 Typhoid fever                                          Typhoid fever
00007 A0100   1 Typhoid fever, unspecified                             Typhoid fever, unspecified
00008 A0101   1 Typhoid meningitis                                     Typhoid meningitis
00009 A0102   1 Typhoid fever with heart involvement                   Typhoid fever with heart involvement
00010 A0103   1 Typhoid pneumonia                                      Typhoid pneumonia
00011 A0104   1 Typhoid arthritis                                      Typhoid arthritis
00012 A0105   1 Typhoid osteomyelitis                                  Typhoid osteomyelitis
00013 A0109   1 Typhoid fever with other complications                 Typhoid fever with other complications
00014 A011    1 Paratyphoid fever A                                    Paratyphoid fever A
00015 A012    1 Paratyphoid fever B                                    Paratyphoid fever B
00016 A013    1 Paratyphoid fever C                                    Paratyphoid fever C
00017 A014    1 Paratyphoid fever, unspecified                         Paratyphoid fever, unspecified
00018 A02     0 Other salmonella infections                            Other salmonella infections
00019 A020    1 Salmonella enteritis                                   Salmonella enteritis
00020 A021    1 Salmonella sepsis                                      Salmonella sepsis
00021 A022    0 Localized salmonella infections                        Localized salmonella infections
00022 A0220   1 Localized salmonella infection, unspecified            Localized salmonella infection, unspecified
00023 A0221   1 Salmonella meningitis                                  Salmonella meningitis
00024 A0222   1 Salmonella pneumonia                                   Salmonella pneumonia
00025 A0223   1 Salmonella arthritis                                   Salmonella arthritis
00026 A0224   1 Salmonella osteomyelitis                               Salmonella osteomyelitis
00027 A0225   1 Salmonella pyelonephritis                              Salmonella pyelonephritis
00028 A0229   1 Salmonella with other localized infection              Salmonella with other localized infection
00029 A028    1 Other specified salmonella infections                  Other specified salmonella infections
```

Figure 4: ICD10CM ORDER 2018 file

Position	Length	Contents
1	5	Order number, right justified, zero filled.
6	1	Blank
7	7	ICD-10-CM or ICD-10-PCS code. Dots are not included.
14	1	Blank
15	1	0 if the code is a “header” –not valid for HIPAA-covered transactions. 1 if the code is valid for submission for HIPAA-covered transactions.
16	1	Blank
17	60	Short description
77	1	Blank
78	To end	Long description

Table 1: ICD10CM ORDER 2018 file layout

Data storage and management

A database has been created using PostgreSQL open-source database software. The NIS 2016 core file and ICD10CM Order 2018 files have been imported to the database using the file layout specification information.

Data extraction, transformation, and analysis

The data stored in the database is extracted using python programming language on a Jupyter Notebook. Jupyter Notebook is a web based open-source interactive computing platform. Code snippets can be written in cells. A Jupyter Notebook cell state is saved even after completion of the code execution. As a changed code cell doesn't require a complete program rerun. Running the program right form the cell that is modified to the last cell in the notebook is sufficient. **Figure 5** shows the solution architecture.

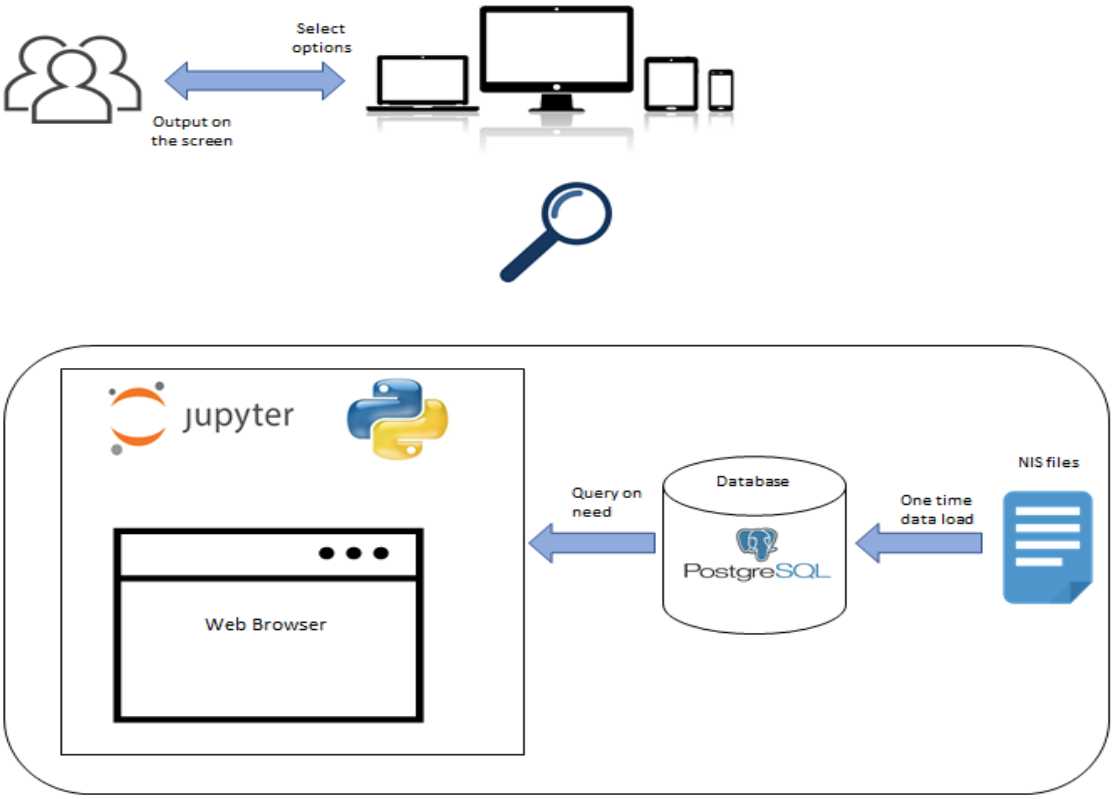

Figure 5: Solution architecture

The extracted data is then verified and cleansed. A unique integer is given to each ICD code found in the extracted data and the mapping is maintained in a dictionary. The data extract, unique integer code information is passed to a custom transformer. The transformer converts the

data extract to LIBSVM format by replacing all the ICD codes with their equivalent integer values in the ICD dictionary. The non-ICD columns are also replaced with unique integers. Below is an example of a LIBSVM record. First value in the record is the target value. It is followed by key value pairs. The key is the integer equivalent of the ICD codes and non-ICD columns. The integers are followed by a colon ':' and a value. The value for all the ICD code keys is defaulted to 1 and the values for the non-ICD codes are the actual values the columns have in the data extract. Below is a sample LIBSVM record:

'0 114:1 525:1 588:1 629:1 920:1 923:1 2051:1 3942:1 3949:1 7123:48 7124:1 7125:2 7126:1 '

The LIBSVN format records are then converted into a sparse matrix. Each key value in the LIBSVM record becomes a column in the sparse matrix and the corresponding value is now passed to the appropriate cell in the sparse matrix. The cell that corresponds to a column that doesn't have a key entry in the LIBSVM record is filled with a zero. The sparse matrix is then given as input to a selection method. The selection method then selects the columns that are significantly contributing towards the colon cancer prediction. These columns are then filtered out from the sparse matrix and a sub dataset is formed. This sub dataset is then passed to the pipeline for training using the selected machine learning algorithm. The remaining data is then passed through the trained model for prediction. The predicted outcomes are then compared with the actual values for accuracy. **Figure 6** shows the process in a flow chart.

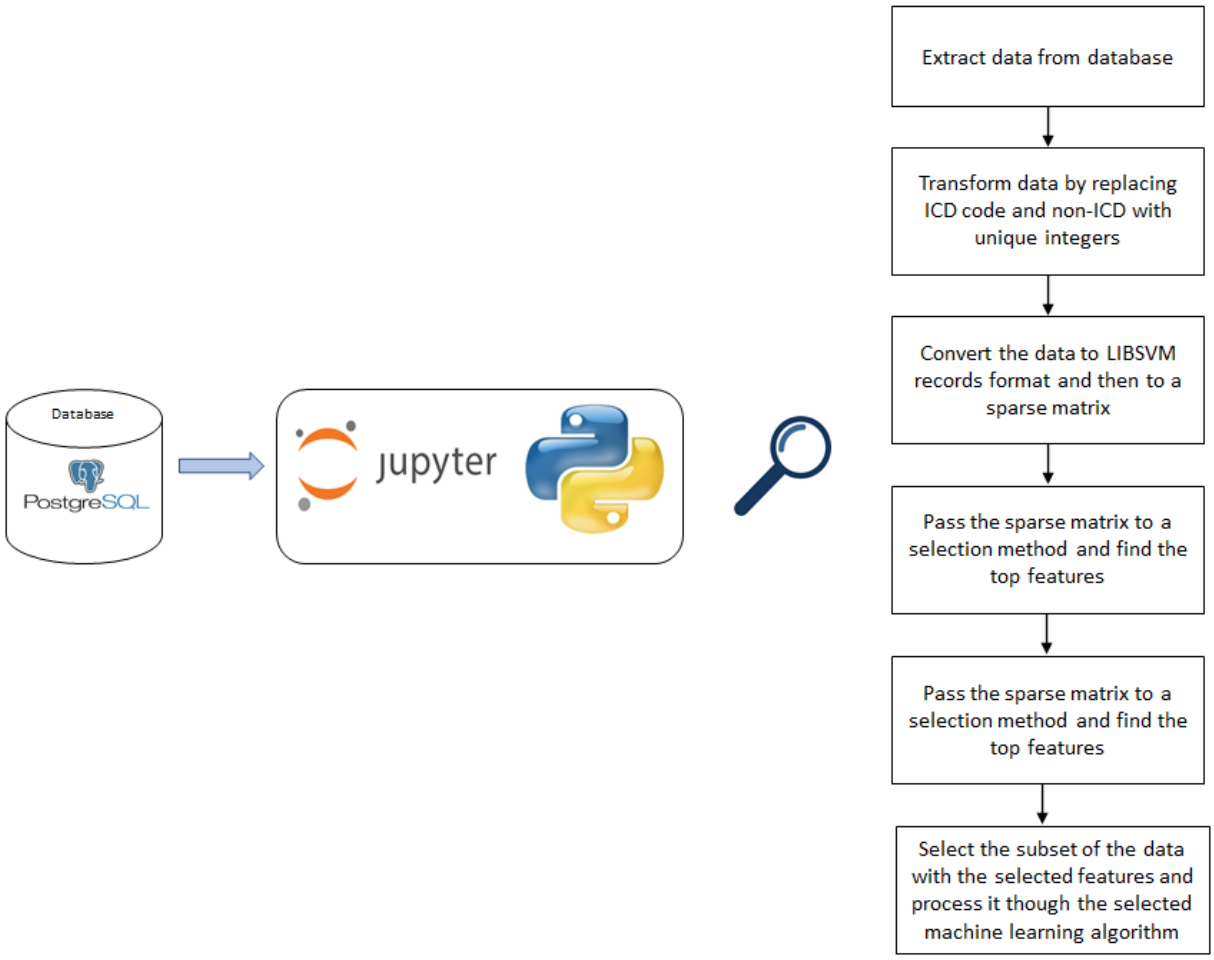

Figure 6: Data processing

Feature Selection Algorithms

Feature selection algorithms help in filtering the number of columns in the data. Below are the few advantages of using feature selection algorithms:

- Identifies the features that are important for the outcome.
- Helps eliminate non-significant feature and thereby avoids over fitting of the model.
- Model performs well for new samples as it retains the generality.
- Since the number of columns are a subset of the original data, the volume of data that is processed goes down and the model takes less time.

In this dissertation 2 feature selection algorithms are used.

Select K Best with chi square

In the chi square method, chi square value is calculated to identify the dependency between the features and the target. The higher the chi square value, the higher the dependency. The features that have higher chi square values in association with the target are identified as significant features for the target prediction.

Feature Importance using Extra Trees Classifier

The extremely randomized trees classifier is an ensemble learning technique that has multiple decision trees forming a forest like the Random Forest classifier. However, it is different from Random Forest Classifier in the way the decision trees are constructed in the forest. A random number of features are allocated to each decision tree and each decision tree selects the best feature to split the data based on a mathematical criterion. This mathematical criterion is used for feature selection.

Machine Learning Models

Decision Tree

Decision Trees are a supervised Machine Learning model in which data is constantly divided based on a certain parameter. The tree can be explained using two entities: decision nodes and leaves. **Figure 7** shows how the leaves symbolize the decisions or outcomes. Decision Trees works with both categorical and continuous input and output data [6].

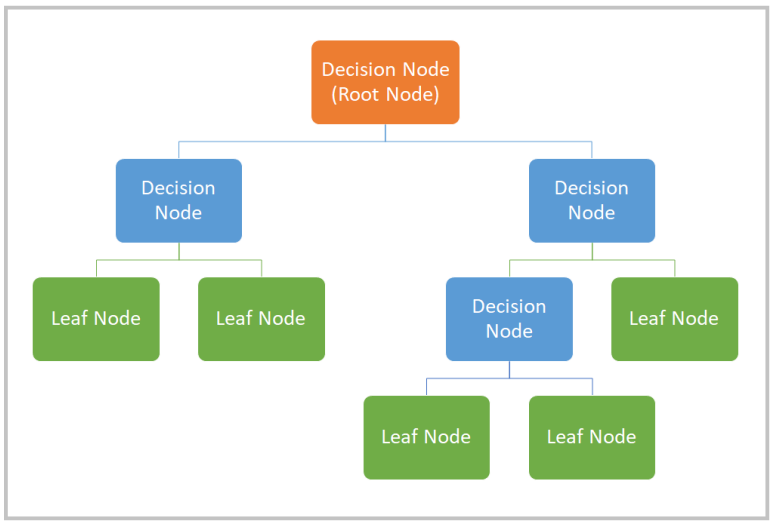

Figure 7: Decision Tree Model [7].

Logistic Regression

Logistic Regression is another Supervised Learning model that uses a set of independent variables to estimate unique binary values (true/false, 0/1, yes/no). As seen in **Figure 8**, it fits data to a logit function to estimate the probability of an event. Logistic Regression is sometimes known as Logit Regression because of this. Because it forecasts probability [6], its output falls within the range of 0 to 1.

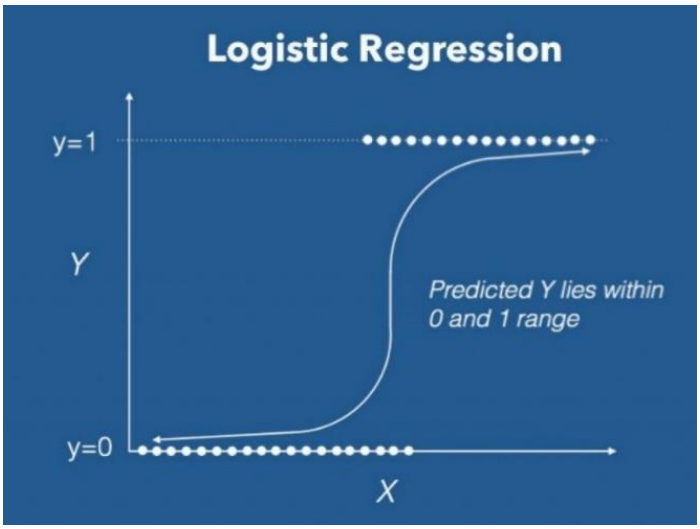

Figure 8: Logistic Regression [8].

Random Forest

Random Forest, also known as Random Decision Forest, is a Supervised Learning algorithm that constructs a "forest" out of a collection of Decision Trees. It's trained using the "bagging" method which is built upon the idea that combining many learning models enhances total output. A Random Forest combines several Decision Trees to provide a more accurate and consistent prediction [9]. This model can be used for regression as well as classification. A Random Forest model architecture is displayed in **Figure 9**.

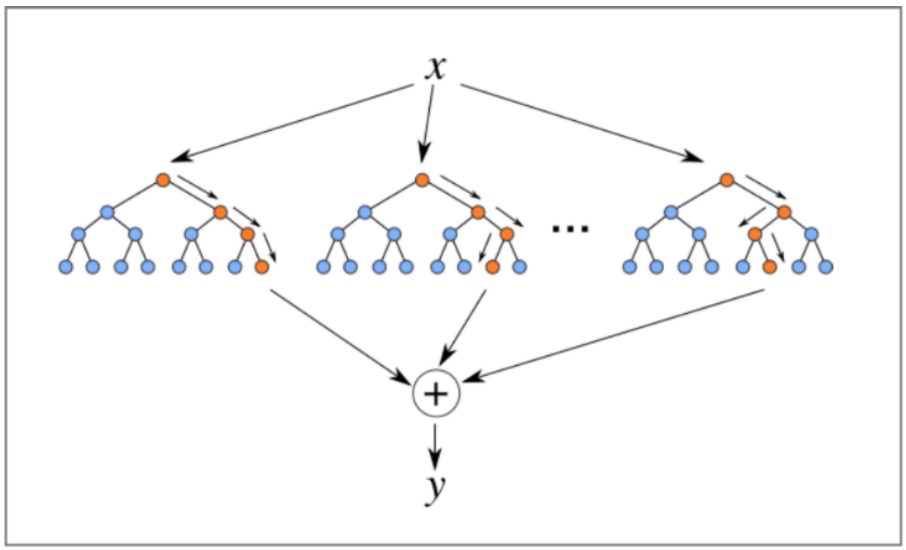

Figure 9: Random Forest architecture [10].

Pipelines

A Machine Learning Pipeline is a way for automating the procedures involved in generating a machine learning model. Different steps such as data extraction, preprocessing, model training, model testing, and deployment are all handled by ML pipelines [11]. Each pipeline stage's behavior can be generalized, and each step can be created as a reusable component. It is possible to set the order in which the components are executed, as well as how

inputs and outputs flow through the pipeline [11]. The pipeline allows the code to work with a variety of selectors, machine learning models, and estimators. The movement of an ML pipeline is seen in **Figure 10.**

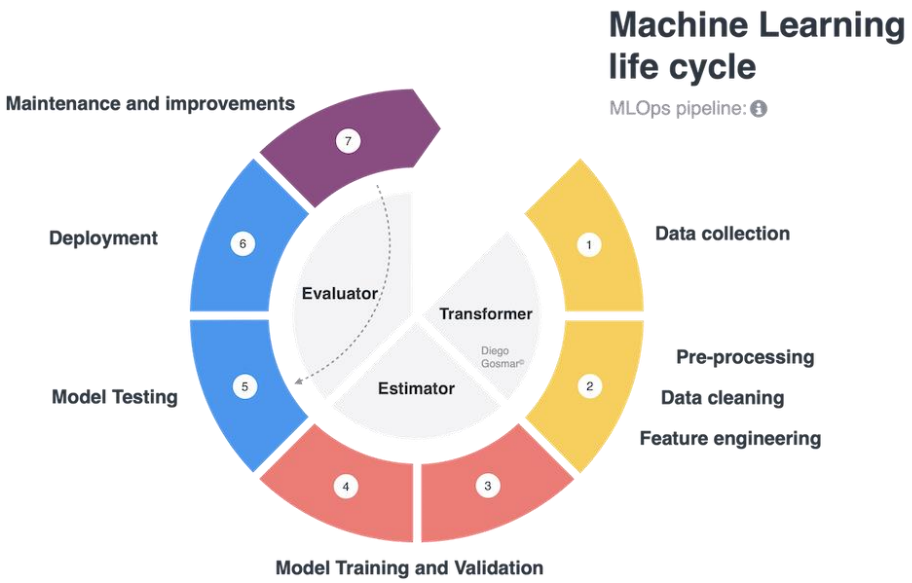

Figure 10: Machine Learning Pipeline [12].

The solution has been developed using Python 3 language because of the rich data science related libraries. Pandas library is used for data manipulation. Scikit learn library has been used to import machine learning, pipeline, feature selection, metrics methods. Browser based notebook programming platform Jupyter Notebook has been used as an IDE. ipywidgets library has been used to build the UI. Jupyter Notebook's Appmode feature has been used to hide the technicalities, for better user experience and interaction.

User Interface

One of the goals of this study is to make the solution available for people with no coding skills. A user interface with options that enables the users is doing a comparative study of various techniques under difference circumstances has been created.

ipywidgets

ipywidgets is an open-source python library that offers interactive HTML widgets for Jupyter notebooks. These widgets are light weight and easy to use with minimal code. They are apt for data science like project that don't need extensive UI capabilities. They can help in creating simple UI options for user input and output. The library offers simple widgets like a TextBox, slider bar, dropdown to complex Asynchronous widgets. shows the widgets we created for this study.

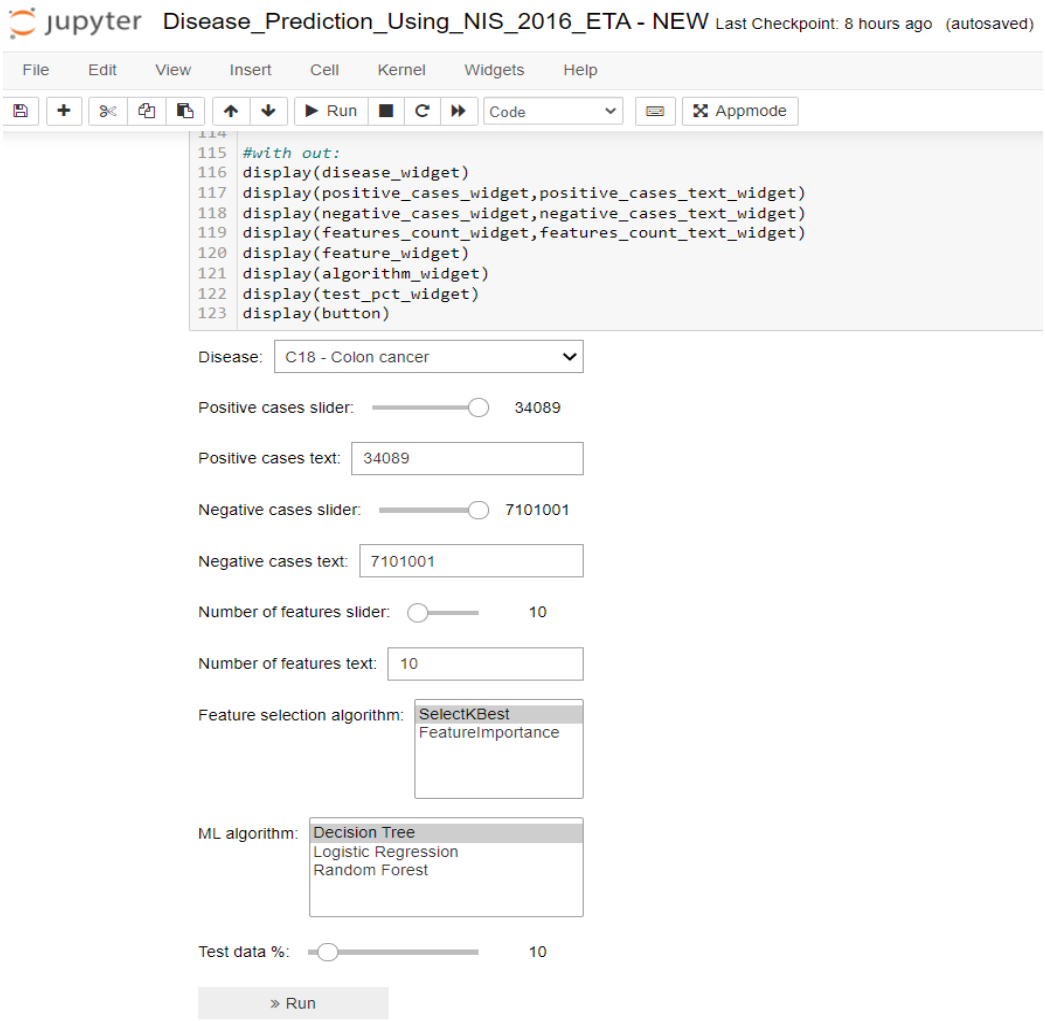

Figure 11: ipywidgets displayed on a jupyter notebook

This is a good option. However, the users still have access to the code as these widgets appear under the code cell in the Jupyter notebook.

Appmode

To hide the code completely, an extension to Jupyter Notebook can be used. Once this extension is installed, Appmode button appears on the Jupyter Notebook tool bar. It is highlighted in **Figure 12**.

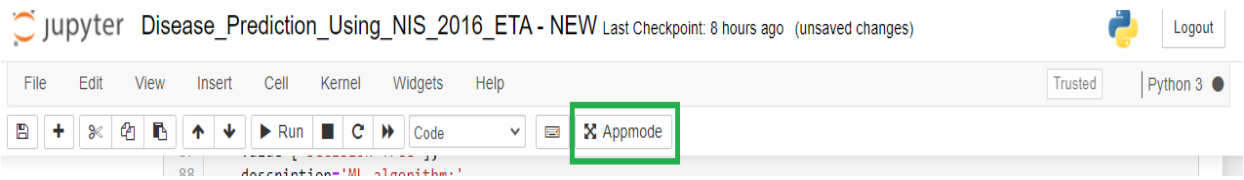

Figure 12: Appmode Jupyter Notebook extension

On clicking on the Appmode extension button, it will take us to a UI only interface that hides the code completely. Although we can go back to the code by selecting the “Edit App” button, it can be disabled and only the UI part can be shared with the users. **Figure 13** shows the Appmode view.

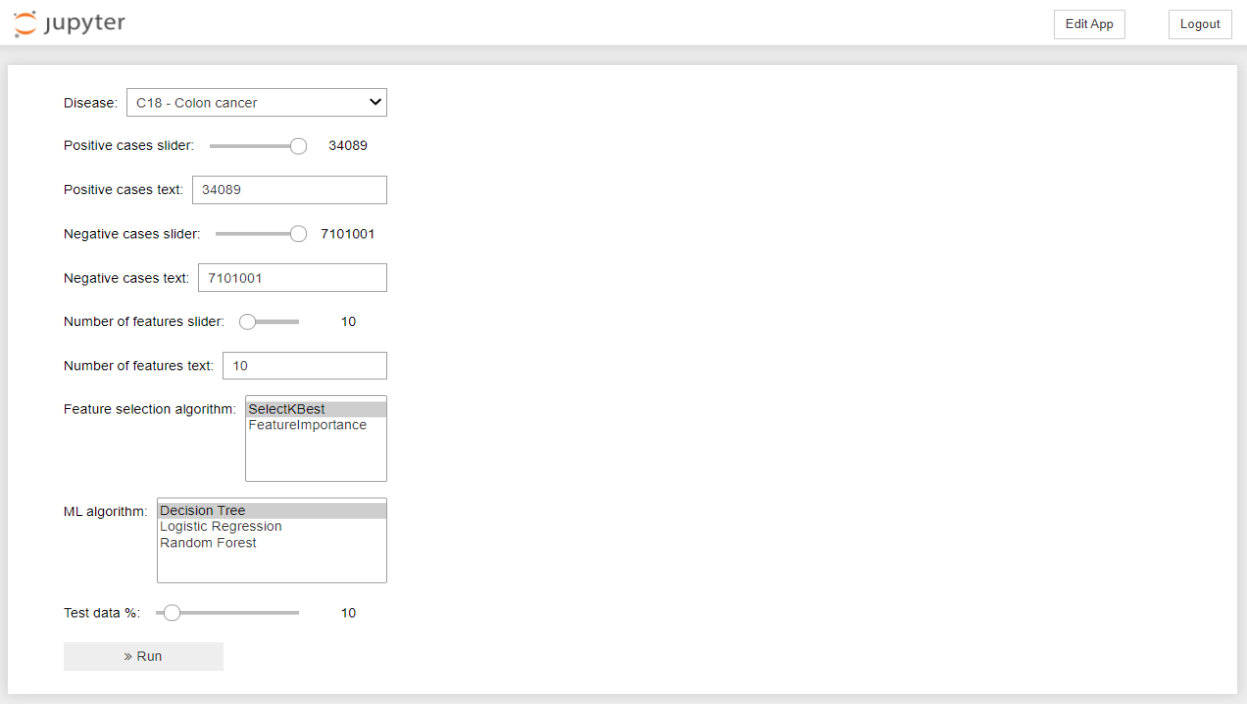

Figure 13: Appmode view on Jupyter Notebook

Features

Table 2 has the options available for the users. Using a Ctrl or Shift button users have the option to select multiple Feature selection or Machine Learning algorithms. This will run multiple iterations of the analysis by using all combinations of the selection and machine learning algorithms.

Feature	Description	Default Value
Disease	Disease code for prognosis study	Colon cancer
Positive cases slider	Number of records with the selected disease code that need to be pulled from the database	Maximum cases for the selected disease
Positive cases text	Number of records with the selected disease code that need to be pulled from the database	Maximum cases for the selected disease
Negative cases slider	Number of records without the selected disease code that need to be pulled from the database	Maximum cases for the selected disease
Negative cases text	Number of records without the selected disease code that need to be pulled from the database	Maximum cases for the selected disease
Number of features	Number of features that need to be selected by the selection algorithm	10

Feature Selection Algorithm	Feature method selection that needs to be used for feature identification	SelectKBest
ML algorithm	Machine Learning algorithm that needs to be used for disease prediction	DecisionTree
Test data %	Percentage of the data pulled from that databased used for training purpose	10

Table 2: UI user input options

Output

Output is displayed on the Jupyter Notebook right below the "Run" button. **Figure 14** shows the user selected options, selected features, Machine learning algorithm related information. **Figure 15** shows selected feature scores and their representation is a bar chart. **Figure 16** shows the selected features and target value heatmap.

```
Feature selection method: SelectKBest
Machine Learning Algorithm: Random Forest
No of features selected: 20
Selected features:
race
C787 - Secondary malig neoplasm of liver and intrahepatic bile duct
Z370 - Single live birth
Z23 - Encounter for immunization
Z3800 - Single liveborn infant, delivered vaginally
C786 - Secondary malignant neoplasm of retroperiton and peritoneum
C7800 - Secondary malignant neoplasm of unspecified lung
Z9049 - Acquired absence of other specified parts of digestive tract
C772 - Secondary and unsp malignant neoplasm of intra-abd nodes
Z3A39 - 39 weeks gestation of pregnancy
Z515 - Encounter for palliative care
Z3801 - Single liveborn infant, delivered by cesarean
K5660 - Unspecified intestinal obstruction
D509 - Iron deficiency anemia, unspecified
Z9221 - Personal history of antineoplastic chemotherapy
D630 - Anemia in neoplastic disease
K913 - Postprocedural intestinal obstruction
K660 - Peritoneal adhesions (postprocedural) (postinfection)
Z933 - Colostomy status
K5669 - Other intestinal obstruction
Training time: 4.85 seconds
Accuracy: 80.10%
Confusion matrix: [[789 188]
 [210 813]]
```

Figure 14: User selected options, feature list, ML outcome

```
Features and Scores:
                                          Features        Scores
1                                             race  32079.486233
2   C787 - Secondary malig neoplasm of liver and i...   2316.411674
3                            Z370 - Single live birth   1292.012308
4                     Z23 - Encounter for immunization    919.528387
5   Z3800 - Single liveborn infant, delivered vagi...    837.000000
6   C786 - Secondary malignant neoplasm of retrope...    815.302243
7   C7800 - Secondary malignant neoplasm of unspec...    681.984281
8   Z9049 - Acquired absence of other specified pa...    649.074405
9   C772 - Secondary and unsp malignant neoplasm o...    589.006745
10            Z3A39 - 39 weeks gestation of pregnancy    564.000000
11               Z515 - Encounter for palliative care    536.737271
12  Z3801 - Single liveborn infant, delivered by c...    440.000000
13         K5660 - Unspecified intestinal obstruction    432.894410
14         D509 - Iron deficiency anemia, unspecified    390.480591
15  Z9221 - Personal history of antineoplastic che...    377.758294
16                D630 - Anemia in neoplastic disease    353.241449
17       K913 - Postprocedural intestinal obstruction    347.115385
18  K660 - Peritoneal adhesions (postprocedural) (...    346.529175
19                            Z933 - Colostomy status    342.007905
20               K5669 - Other intestinal obstruction    340.984810
```

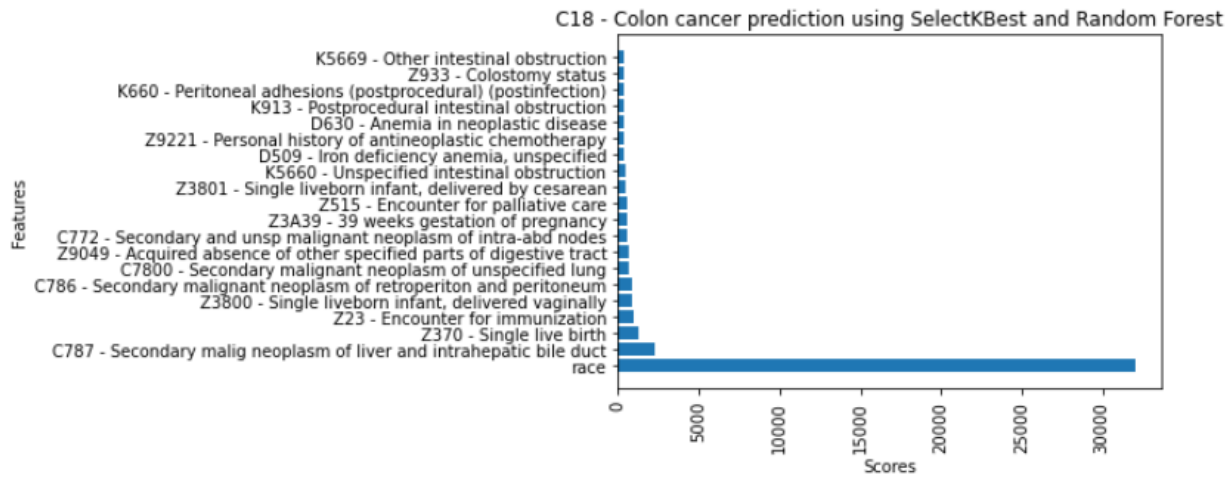

Figure 15: Selected feature scores

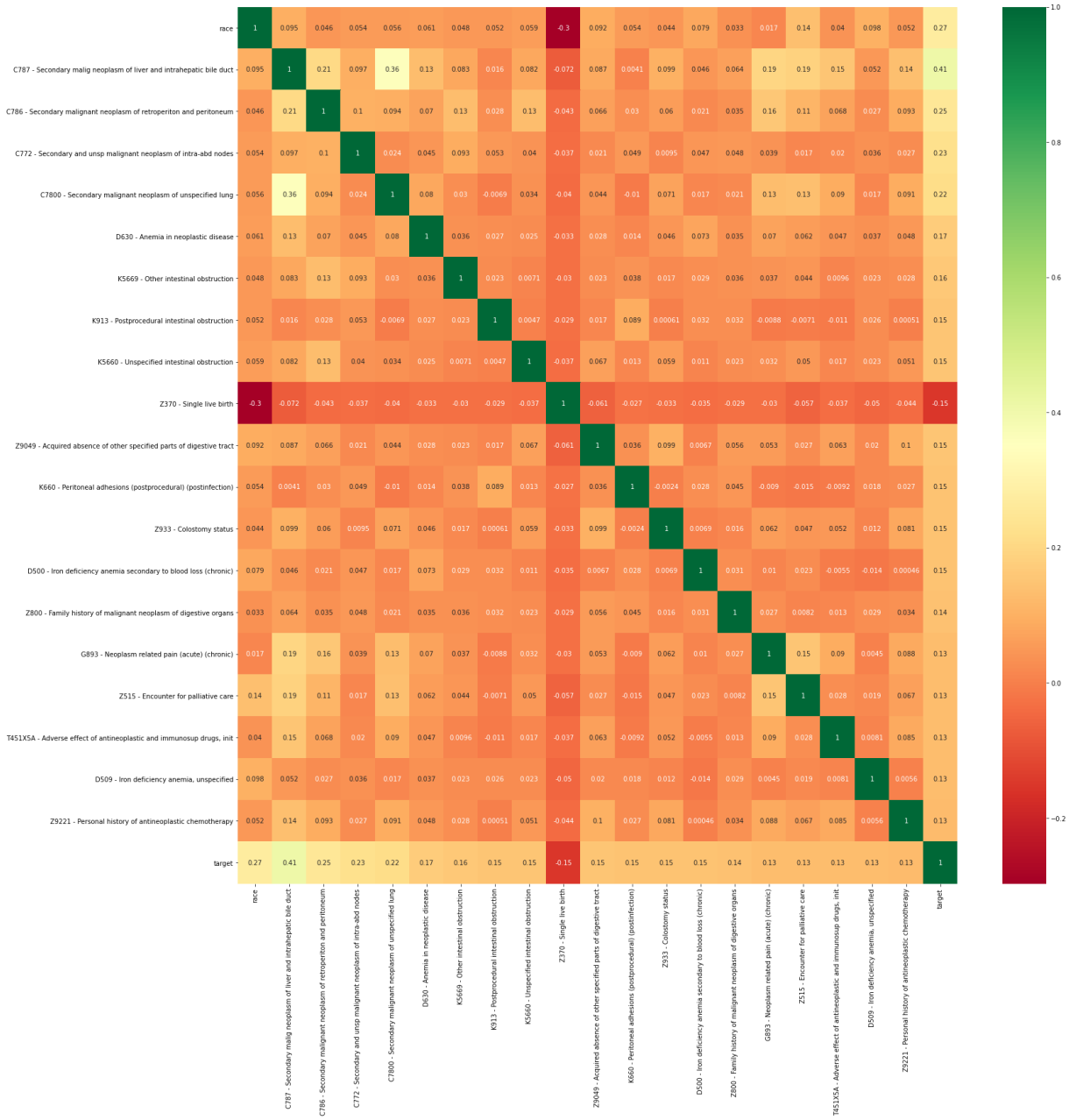

Figure 16: Selected features and target heatmap

The output is thoroughly discussed in the Results section.

Case study

The pipeline was tested using ‘Colon Cancer’ as the disease condition. 130,000 patient data was randomly selected which consisted of 30,000 positive cases and 100,000 negative cases. The top 20 features were selected using the selectors ‘FeatureImportance’ and ‘SelectKBest’. Predictions are made with the use of machine learning models such as Decision Tree, Logistic Regression and Random Forest. The data was split into 2 parts where 10% of the data was used for testing and the rest was used for training.

There can be 6 test cases since multiple feature selection techniques and machine learning models are used. To keep the results comparable the different algorithms are used on the same dataset. In the results section, the identified top features are analyzed, heat maps are drawn to find correlation among these features. Finally, the prognosis obtained from the 6 test cases are thoroughly compared to find the best performing combination of feature selector and model.

Results

Selector 1: SelectKBest

The top 20 features identified by the SelectKBest selector are listed in **Table 3** along with their feature scores.

No	Feature	Score
1	race	110969.986336
2	C787 - Secondary malig neoplasm of liver and intrahepatic bile duct	20481.787019
3	C786 - Secondary malignant neoplasm of retroperiton and peritoneum	7986.015385
4	C772 - Secondary and unsp malignant neoplasm of intra-abd nodes	6525.202963
5	C7800 - Secondary malignant neoplasm of unspecified lung	6025.506305
6	D630 - Anemia in neoplastic disease	3688.020718
7	K5669 - Other intestinal obstruction	3428.428736
8	K913 - Postprocedural intestinal obstruction	3042.813102
9	K5660 - Unspecified intestinal obstruction	2879.670609
10	Z370 - Single live birth	2854.102364
11	Z9049 - Acquired absence of other specified parts of digestive tract	2849.583594
12	K660 - Peritoneal adhesions (postprocedural) (postinfection)	2782.868982
13	Z933 - Colostomy status	2756.720466
14	D500 - Iron deficiency anemia secondary to blood loss (chronic)	2697.030743
15	Z800 - Family history of malignant neoplasm of digestive organs	2628.255751
16	G893 - Neoplasm related pain (acute) (chronic)	2296.386377
17	Z515 - Encounter for palliative care	2165.093978
18	T451X5A - Adverse effect of antineoplastic and immunosup drugs, init	2144.598389
19	D509 - Iron deficiency anemia, unspecified	2104.356476
20	Z9221 - Personal history of antineoplastic chemotherapy	2011.939061

Table 3: Top 20 features selected by SelectKBest selector.

Figure 17 represents the heat map indicating the correlation between the target and each of the features, where the correlation score ranges from 1 (deep green) to -.05 (deep red). A positive correlation score means the 2 features move in the same direction whereas a negative correlation indicates the features moving in the opposite direction. Assuming features A and B are negatively correlated, the value of A will increase if the value of B decreases and vice versa [13].

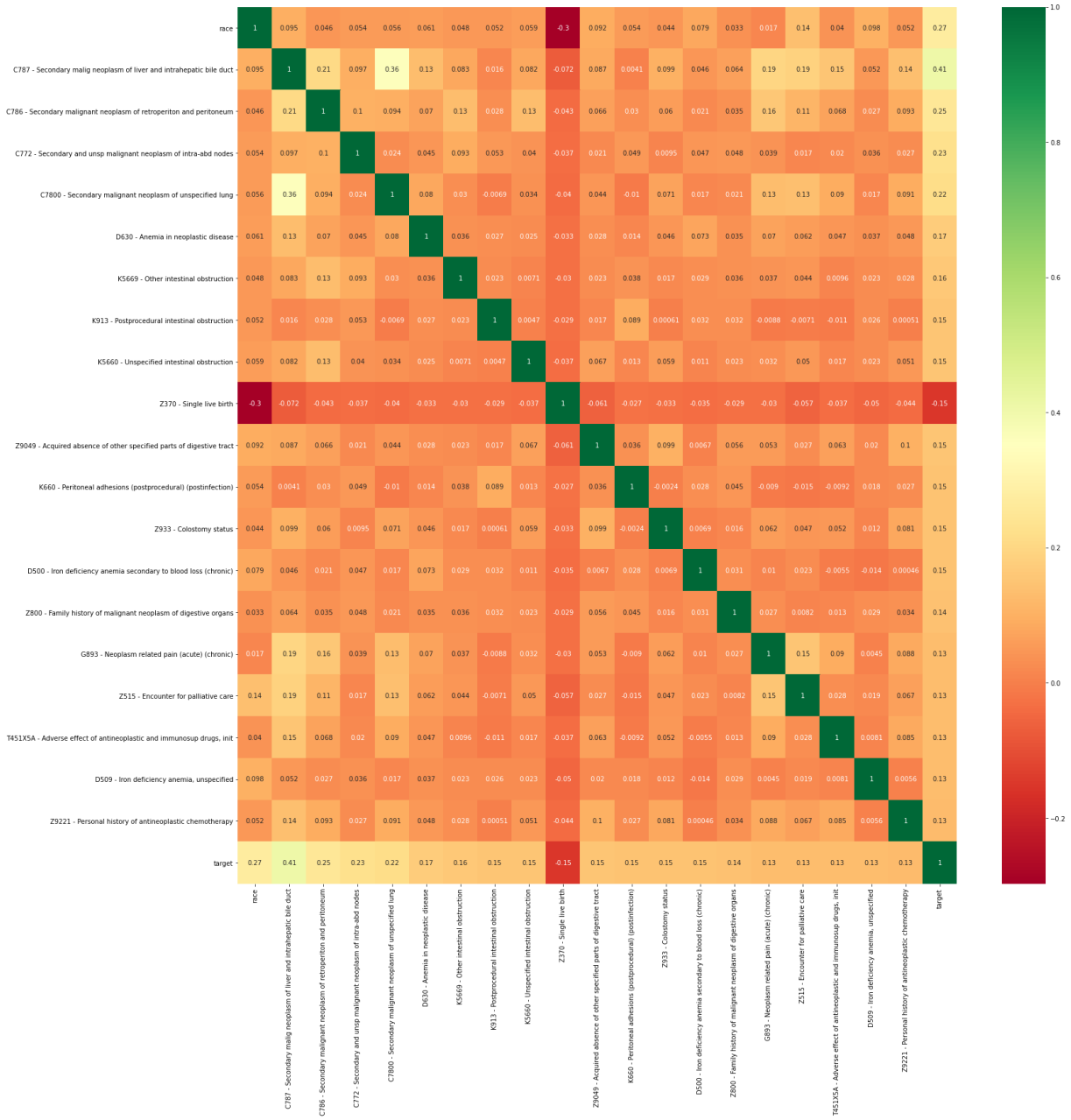

Figure 17: Heat map for correlation between feature and target (SelectKBest selector).

After identifying the top features, they are passed through the pipeline and fitted into the machine learning models. The performance of the 3 models is compared in **Figure 18**. Random Forest performs best with the SelectKBest selector and achieves an accuracy of 87.33% followed closely by Logistic Regression being 86.81% accurate and lastly Decision Tree is the least accurate having an accuracy of 83.15%.

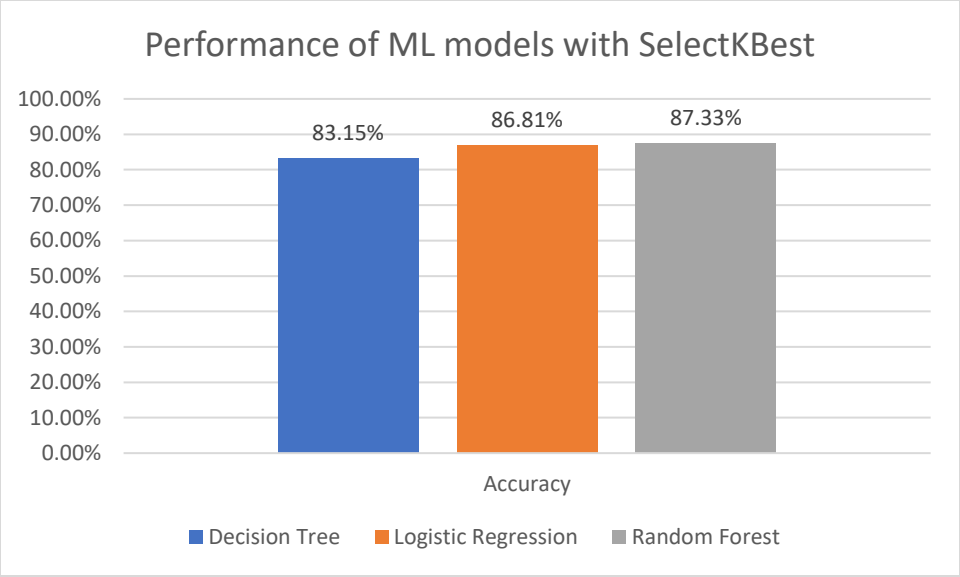

Figure 18: Accuracy comparison for different models using SelectKBest selector.

Selector 2: Feature Importance

The same data is now passed through the Feature Importance selector to identify the 20 features. The sample data and the feature numbers are kept identical to keep the results comparable. **Table 4** represents the selected features by the Feature Importance selector and their feature scores.

No	Feature	Score
1	C787 - Secondary malig neoplasm of liver and intrahepatic bile duct	0.067829
2	race	0.033172
3	C772 - Secondary and unsp malignant neoplasm of intra-abd nodes	0.022952
4	C786 - Secondary malignant neoplasm of retroperiton and peritoneum	0.020745
5	C7800 - Secondary malignant neoplasm of unspecified lung	0.014536
6	female	0.012065
7	K660 - Peritoneal adhesions (postprocedural) (postinfection)	0.010818
8	K913 - Postprocedural intestinal obstruction	0.009963
9	D630 - Anemia in neoplastic disease	0.009318
10	K5669 - Other intestinal obstruction	0.008566
11	Z9049 - Acquired absence of other specified parts of digestive tract	0.008354
12	Z933 - Colostomy status	0.008087
13	Z800 - Family history of malignant neoplasm of digestive organs	0.007681
14	D500 - Iron deficiency anemia secondary to blood loss (chronic)	0.007513
15	K5660 - Unspecified intestinal obstruction	0.007428
16	D509 - Iron deficiency anemia, unspecified	0.007115
17	age	0.006702
18	K567 - Ileus, unspecified	0.006696
19	I10 - Essential (primary) hypertension	0.006410
20	Z370 - Single live birth	0.005775

Table 4: Top 20 features selected by Feature Importance selector.

The correlation between the target and each of the features shown in **Figure 19** using a heat map. The correlation score ranges from 1 (deep green) to -.05 (deep red) like **Figure 17**.

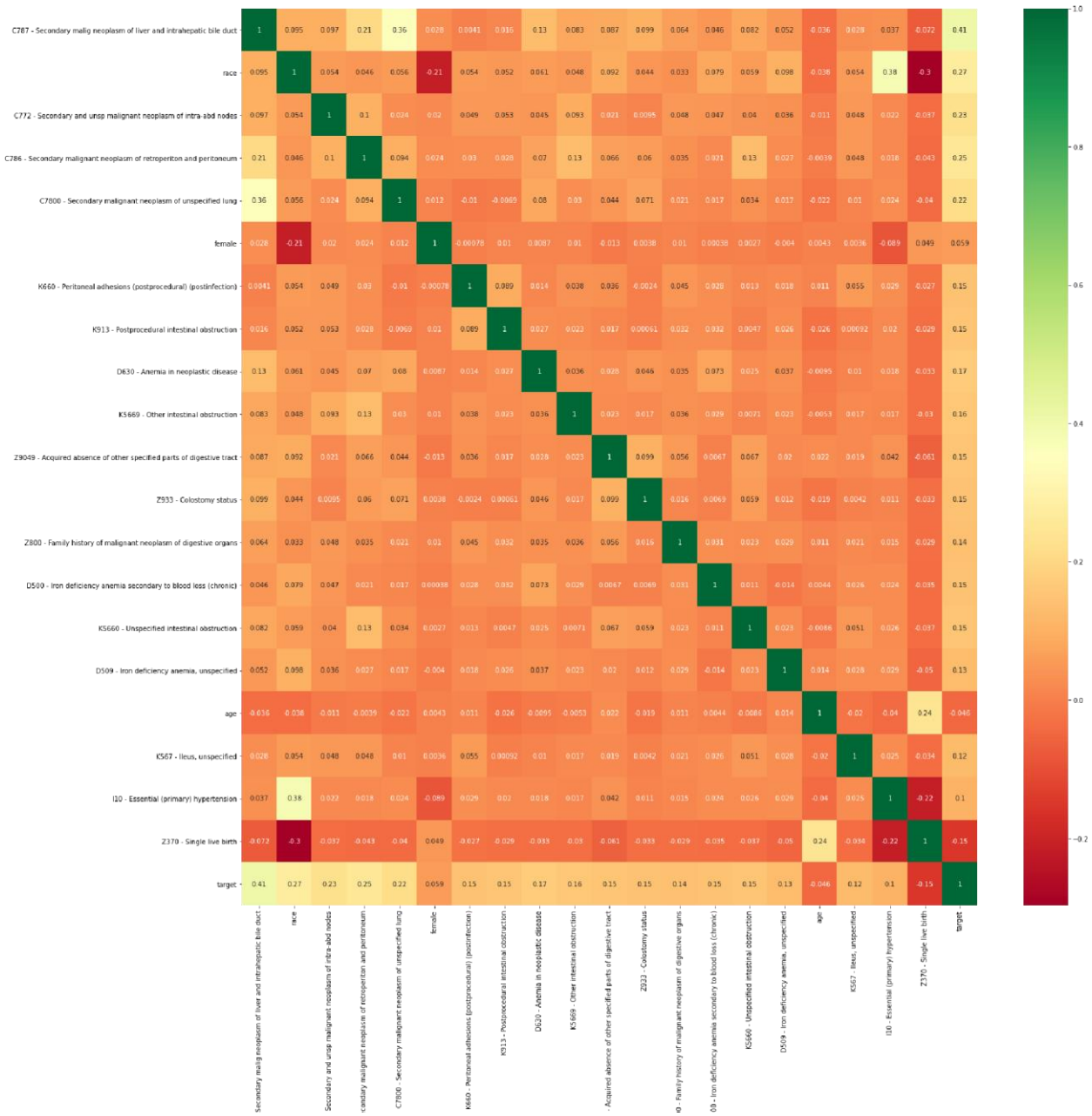

Figure 19: Heat map for correlation between feature and target (SelectKBest selector).

The 3 machine learning models perform very similarly to the previous results under the Feature Importance selector. Their performance is compared in **Figure 20**. Random Forest still performs best with a slightly higher accuracy achieved with SelectKBest selector. The Random Forest model and Feature Importance Selector combination provided the highest accuracy of 87.61%. Logistic Regression slightly decreases, and Decision Tree performs the same. The

machine learning models produce almost identical results using the 2 different feature selector models since both the selectors identify similar features.

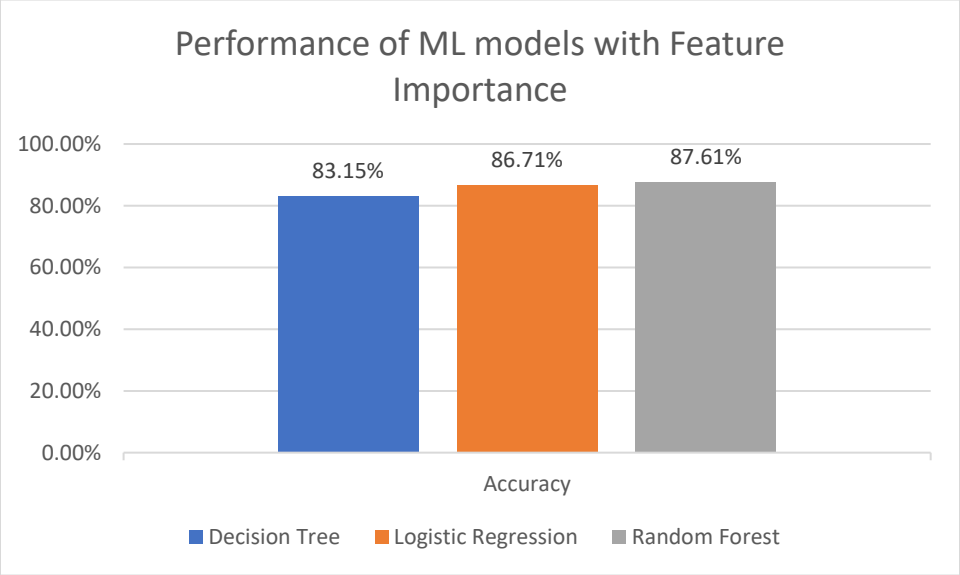

Figure 20: Accuracy comparison for different models using Feature Importance selector.

Conclusion

Early diagnosis of diseases is extremely crucial for the effective treatment for many diseases. The goal of this dissertation is to predict diseases and help healthcare professionals to make more accurate diagnosis. An interactive platform is created which allows the user to pick among multiple feature selection methods and machine learning models to generate predictions. The user also has the option of tweaking disease condition to predict, sample data size, test to train ratio, etc.

With a view to test the platform a case study is performed using colon cancer. Using 2 selection algorithms (Feature Importance and SelectKBest) and 3 machine learning algorithms (Decision Tree, Logistic Regression, Random Forest) promising results are observed. Some of the most important features that contributes to the prognosis include Secondary malig neoplasm of liver and intrahepatic bile duct, Secondary malignant neoplasm of retroperiton and peritoneum, Secondary and unsp malignant neoplasm of intra-abd nodes. The three Machine Learning algorithms performed consistently. Random Forest performed slightly well than Decision Tree and Logistic Regression.

Application & future work

Applications:

This dissertation can be widely used in the healthcare domain to predict diseases which in turn will help the healthcare professionals to provide more accurate diagnosis and better treatment to patients. It's also expected to have significant impact in preventive care as it helps early diagnosis. Insurance companies can use this platform to predict client's possibility of suffering from a disease and finalize insurance premium accordingly.

Future work:

Though the result for this dissertation is promising, there are a few improvements that can be incorporated in the future. Instead of using only a single year data, multiple year data can be used to improve the performance of the platform. This would include building a database of disease conditions which is compatible with multiple disease classifications since different year data uses different disease classifications. As the data format for multiple year is slightly different, data should be processed into a general format before using multiple year data.

Deep learning models can be used to increase accuracy as well. Neural networks are extremely powerful to make accurate predictions. Due to resource constraint deep learning could not be explored.

References

1. "Colon cancer - Symptoms and causes - Mayo Clinic". [Online]. Available: https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669. [Accessed: 09-Dec.-2021].

2. "Key Statistics for Colorectal Cancer". [Online]. Available: https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html. [Accessed: 09-Dec.-2021].

3. "USCS Data Visualizations - CDC". [Online]. Available: https://gis.cdc.gov/Cancer/USCS/. [Accessed: 09-Dec.-2021].

4. "Colon cancer - Diagnosis and treatment - Mayo Clinic". [Online]. Available: https://www.mayoclinic.org/diseases-conditions/colon-cancer/diagnosis-treatment/drc-20353674. [Accessed: 09-Dec.-2021].

5. "Introduction to the HCUP National Inpatient Sample (NIS), 2016,". [Online]. Available: https://www.hcup-us.ahrq.gov/db/nation/nis/NIS_Introduction_2016.jsp. [Accessed: 09-Dec.-2021].

6. "Commonly Used Machine Learning Algorithms | Data Science". [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/. [Accessed: 09-Dec.-2021].

7. "Machine Learning Decision Tree Classification Algorithm - Javatpoint". [Online]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm. [Accessed: 09-Dec.-2021].

8. "Logistic Regression - A Complete Tutorial with Examples in R". [Online]. Available: https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/. [Accessed: 09-Dec.-2021].

9. "Random Forest Algorithms: A Complete Guide | Built In". [Online]. Available: https://builtin.com/data-science/random-forest-algorithm. [Accessed: 09-Dec.-2021].

10. "Random Forest Regression". [Online]. Available: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84. [Accessed: 09-Dec.-2021].

11. "What is a Machine Learning Pipeline?". [Online]. Available: https://valohai.com/machine-learning-pipeline/. [Accessed: 09-Dec.-2021].

12. "MLOps scalability - Machine Learning". [Online]. Available: https://www.gosmar.eu/machinelearning/2021/01/02/mlops-scalability/. [Accessed: 09-Dec.-2021].c

13. “Correlation Definitions, Examples & Interpretation | Simply Psychology”. [Online]. Available: https://www.simplypsychology.org/correlation.html. [Accessed: 09-Dec.-2021].

14. “Machine Learning for Detection and Diagnosis of Disease | Annual ...”. [Online]. Available: https://www.annualreviews.org/doi/abs/10.1146/annurev.bioeng.8.061505.095802. [Accessed: 09-Dec.-2021].

15. “Machine learning for improved diagnosis and prognosis in ...”. [Online]. Available: http://ieeexplore.ieee.org/document/7943950/. [Accessed: 09-Dec.-2021].

16. “Comparing different supervised machine learning algorithms for ...”. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8. [Accessed: 09-Dec.-2021].

17. “Disease Prediction by Machine Learning Over Big Data From ...”. [Online]. Available: https://ieeexplore.ieee.org/document/7912315. [Accessed: 09-Dec.-2021].

18. “Stage II Colon Cancer Prognosis Prediction by Tumor Gene ...”. [Online]. Available: https://ascopubs.org/doi/abs/10.1200/jco.2005.05.0229. [Accessed: 09-Dec.-2021].

19. “Colon cancer prognosis prediction by gene expression profiling ...”. [Online]. Available: https://www.nature.com/articles/1208984. [Accessed: 09-Dec.-2021].

20. “Serum-based microRNA signatures in early diagnosis and ...”. [Online]. Available: https://academic.oup.com/carcin/article/37/10/941/2196570. [Accessed: 09-Dec.-2021].