

HAP 780: Final Project

LENGTH OF STAY IN HOSPITAL AFTER GETTING DISCHARGED FROM ICU WARD

Professor: Janusz Wojtusiak

**Sai Shashank Vinnakota
G01221348
12/13/2020**

Table of Content

Project Description	1
Introduction	2
Dataset	2
SW/HW	3
Objective	3
Data Pre-Processing	3
1. Admissions_Table	4
2. Patients_Table	5
3. Diagnosis_Table	6
4. Procedures_Table	7
5. CPT Events Table	7
6. ICUSTAYS Table	8
7. Service Table	8
8. Joining all temporary tables	10
9. Final Dataset	11
WEKA	11
Machine Learning Models	13
Naïve Bayes Classifier 70 – 30 Split	13
Bayes Net Classifier 70 – 30 Split	13
Decision Tree 70 – 30 Split	14
Additive Logistic Regression (Logit Boost) 70 – 30 Split.	15
Naïve Bayes Classifier 10 - Fold Cross-Validation	15
Bayes Net Classifier 10 - Fold Cross-Validation	16
Decision Tree Classifier 10 - Fold Cross-Validation	16
Additive Logistic Regression Classifier 10 - Fold Cross-Validation	17
RESULTS	17
70 - 30 Split	17
10 – FOLD CROSS VALIDATION	18
Conclusion	18
Future Work	18
References	19

Project Description

In this project, I have used various machine learning models in weka for predicting the length of stay of patient's in the hospital who are in the ICU and after careful study, I have chosen the best model that has the best accuracy when compared to other models and then even checked their ROC, Precision and recall values and then came to a final conclusion. I have seen that the average stay after leaving ICU is 7 days and the median is around 4-5 days. So, I have decided to consider that if the length of stay after leaving ICU > 5 days then it would be that the patient is serious and if it is < 5 it shows that the patient had an issue but was cured easily.

Introduction

Forecasting Length of Stay hospitals to distinguish patients who might be in danger for a lengthy stay, and afterward change a patient's treatment plan from the purpose of admission to lessen the time in the clinic. The length of stay (LOS) in medical clinics is regularly utilized as a marker of productivity of care and clinic execution. Anticipating Length of Stay likewise engages clinics to improve patient treatment by meeting assumptions about the hospital stay that are set during confirmation. At last, it improves hospitals' ability in allocating resources and predicting the number of beds that are required in the ICU ward if the count of patients keeps increasing. Hospitals generally cannot predict that tomorrow they would have these many patients but once the patient is admitted in the hospital, we can check if the patient is going to be admitted in the ICU ward or the general ward. If the patient is admitted to the ICU ward, then it becomes critical in order to understand how many days he would stay in this ward and after he leaves the ICU ward. This can be affected due to several other reasons but in general, the length of stay would help allocate the beds accordingly and reduce the variance of the required beds and helps in efficient utilization of staff in the ICU. [1]

There are many factors that affect the length of stay of a patient. The patient might be wrongly diagnosed in the beginning and then later the way he responds to the medication might lead to the actual understanding of the problem the patient is going through. Even the lack of intensive care of the nurses, when required due to the wrong prediction of the length of stay in the hospital, might lead to such issues. All these factors play a major role in the prolonged length of stay in hospitals. But after keeping in mind all such factors as well I have considered proceeding with the dataset. In the United States of America alone the total amount that is spent on health services is 36% of the total budget. The average length of stay in the USA is around 4-5 days. The amount that is being spent on an average per stay is around \$12,000 and the average ages were between 48-85 an older year. [2]

When a patient is admitted to the hospital, he/she might have multiple diseases with him or might have a single disease or even might be admitted due to an accident. So, in order to study the overall length of stay of any patient who is admitted in ICU and after leaving it to another ward, I have chosen this project.

Dataset

For the project I have used the MIMIC – III data set in order to predict the length of stay after data analysis and using machine learning models. Mimic – III data is one of the databases that comprise a relational database that contains data of patients who have stayed after leaving ICU. The tables are interlinked with each other and have two types of ID's that are HADM_ID and SUBJECT_ID that refer to unique patients. This is a free database containing insights on Laboratory measurements, Medications ordered, Hospital assigned diagnosis and Patient procedures. [3]

I have considered the following datasets from the mimic – III data for understanding the length of stay of the patients in ICU.

1. ADMISSIONS - this data set contains the data regarding the patient's admission to the hospital. Every unique hospitalization for each patient in the database (HADM_ID).
2. CPTEVENTS – procedures are recorded as CPT codes. This data can be helpful in determining which specific procedures have been performed. For every such procedure, there is a CPT code.
3. DIAGNOSES-ICD – this data set consists of data that helps in understanding the diagnosis that is undergone by each patient and the codes are present for the same.
4. ICU_STAYS – This data set helps us understand the information of each patient's stay in ICU hospital, this gives us the ICUSTAY_ID.
5. PATIENTS – This dataset contains data of all patients, but this is hospital – independent information and consists of attributes such as gender and DOB.
6. PROCEDURES_ID – In this data set we find all the procedures that each patient has gone through with ICD standards.
7. SERVICES – In this data set, we can understand what services the patient has taken like Ortho, GYN, etc.

In this data set, I have observed that every icustaysid is linked to a single hadm_id and subject_id. A single-subject can be subjected to many hadm and icustaysid. That means we can observe that the length of stay of each patient can be in one hospital or multiple hospitals or even both cases can be possible.

SW/HW

Software used: SQL Server Management Studio and Weka.

Hardware used: Windows 10 Good RAM and ROM.

Objective

The main objective of this project is to understand the length of stay of patients in the hospital after leaving the ICU ward. The main motto is to understand that if the length of stay is >5 days after leaving the ICU then that means such patients have a high probability of getting back to the ICU ward. So, in order to avoid this, the results of this project would help the hospital administrators to understand the care with which such patients have to be treated. [4]

Data Pre-Processing

I have used the Microsoft SQL server management tool in order to understand the data and convert it into the format that supports data processing in Weka. Here I have taken the admissions table in the first place which has the following factors that have to be converted to meet the requirements of weka for data mining.

1. Admissions_Table

- Admission_Location
 - Admission_Type
 - Marital_Status
 - Insurance
 - Religion
 - Ethnicity

HAP_PROJECT_SQL_XPRTStuvena (62) - Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Help

APPI

Object Explorer

master

LAPTOP-4TTRK9PE\master (SQL Server 15.0.2070.41 - LAPTOP-4TTRK9PE\vinna (62)) - Microsoft SQL Server Management Studio

Quick Launch (Ctrl+C)

LAPTOP-4TTRK9PE (15.0 RTM) LAPTOP-4TTRK9PE\vinna - master | 00:00:00 | 9 rows

use project

select * from ADMISSIONS

ADMISSIONS_TYPE

select ad.SUBJECT_ID as subject_id, ad.HADM_ID as adm_id,

case

when ad.ADMISSION_LOCATION = 'MDN REFERRAL/ICK' then 0

when ad.ADMISSION_LOCATION = 'CLINIC REFERRAL/PREEMATURE' then 1

when ad.ADMISSION_LOCATION = 'EMERGENCY ROOM ADMIT' then 2

when ad.ADMISSION_LOCATION = 'TRANSFER FROM HOSP/EXTANT' then 3

when ad.ADMISSION_LOCATION = 'TRANSFER FROM OTHER HOSP' then 4

when ad.ADMISSION_LOCATION = 'TRANSFER FROM OTHER HEALTH' then 5

when ad.ADMISSION_LOCATION = '** INFO NOT AVAILABLE **' then 6

when ad.ADMISSION_LOCATION = 'PHYS REFERRAL/NORMAL DELI' then 7

when ad.ADMISSION_LOCATION = 'TRANSFER FROM SKILLED NUR' then 8

end admission_location

case

when ad.ADMISSION_TYPE = 'ELECTIVE' then 1

when ad.ADMISSION_TYPE = 'EMERGENCY' then 2

when ad.ADMISSION_TYPE = 'NEONATAL' then 3

when ad.ADMISSION_TYPE = 'URGENT' then 4

end admission_type

,

case

when ad.MARITAL_STATUS = 'SINGLE' then 1

when ad.MARITAL_STATUS = 'WIDOWED' then 2

when ad.MARITAL_STATUS = 'SEPARATED' then 3

when ad.MARITAL_STATUS = 'LIFE PARTNER' then 4

when ad.MARITAL_STATUS = 'MARRIED' then 5

when ad.MARITAL_STATUS = 'UNMARRIED (OFFICIAL)' then 6

when ad.MARITAL_STATUS = 'DIVORCED' then 7

else 8

end marital_status

case

when ad.INSURANCE = 'Private' then 1

when ad.INSURANCE = 'Medicaid' then 2

when ad.INSURANCE = 'Self Pay' then 3

when ad.INSURANCE = 'Government' then 4

when ad.INSURANCE = 'Medicare' then 5

end insurance

,

case

when ad.RELIGION = 'PROTESTANT' then 1

when ad.RELIGION = 'CATHOLIC' then 2

when ad.RELIGION = 'JEWISH' then 3

when ad.RELIGION = 'BAPTIST' then 4

when ad.RELIGION = 'METHODIST' then 5

when ad.RELIGION = 'ORTHODOX' then 6

when ad.RELIGION = 'PROTESTANT' then 7

when ad.RELIGION = 'CATHOLIC' then 8

when ad.RELIGION = 'JEWISH' then 9

when ad.RELIGION = 'BAPTIST' then 10

when ad.RELIGION = 'METHODIST' then 11

else 12

end religion

Connected (1/1)

Le 31 Cal 53 Ch 53 Int

100% 12/11/2020

```

CREATE TABLE #ADMISSIONS_TABLE_1 (
    subject_id INT,
    adm_id INT,
    admission_location INT,
    admission_type INT,
    marital_status INT,
    insurance INT,
    religion INT,
    ethnicity NVARCHAR(50),
    admit_time DATETIME,
    discharge_time DATETIME
)

-- Insertion logic
INSERT INTO #ADMISSIONS_TABLE_1
SELECT * FROM ADMISSIONS

```

The screenshot shows the SQL query being run in SSMS. The results pane displays the first 15 rows of the temporary table, which are identical to the original ADMISSIONS table data.

In the above 3 steps (screenshots) we can see that I have converted all the data in the admissions table to numerical values that help in working in WEKA. The final output table is Admission_table_1. I have made this table a temporary table as it becomes easy to run the code even if the connection is lost.

2. Patients_Table

Now I have taken into consideration the patient's table where we can find the gender column. Gender is one of the important factors that contribute to the length of stay. I have converted Female 'F' as 1 and Male as '2', in the code.

```

CREATE TABLE #PT1 (
    subject_id INT,
    gender NVARCHAR(1),
    dob DATETIME
)

-- Insertion logic
INSERT INTO #PT1
SELECT p.SUBJECT_ID AS subject_id,
       CASE WHEN p.GENDER = 'F' THEN 1 ELSE 2 END AS gender,
       p.DOB
FROM [dbo].[PATIENTS] p

```

The screenshot shows the SQL query being run in SSMS. The results pane displays the first 15 rows of the temporary table, showing the converted gender values (1 for Female, 2 for Male).

3. Diagnosis_Table

Here I have taken into consideration the diagnosis table with ICD codes here I have removed the codes that start with A and R because ICD9 codes that start with 'A' or 'R' do not help in my criteria as they do not consist of data that deals with diseases or external injuries.

In the below figure we can see the table that consists of codes that do not include the codes that are starting from A and R. This table is put into a temporary table named #icd9_table. We can see three columns that have been put into temporary table #icd9_table Subject_id, Adm_id, icd9_code.

In the above figure I have joined the table based on subject_id and admin_id with seq_num hence, this additional column has been added into #dt1 from the #td table. The following are the columns in the table: Subject_id, Adm_id,icd9_codes and Seq_num.

4. Procedures_Table

```

--select * from #td
--drop table #td

--select * from #proceduretab1
--drop table #proceduretab1

```

subject_id	adm_id	proc_num
1	108700	2
2	16537	3
3	23031	184185
4	88291	113932
5	6914	134765
6	22507	139627
7	22264	159150
8	10343	160056
9	10463	161731
10	76005	179180
11	10393	190118
12	69776	137906
13	22207	150107
14	31140	163444

In the above figure, we can see that I have calculated the total number of procedures that every patient has gone through. This even lets us guess that the total number of procedures would be one of the leading factors that lead to the length of stay in the hospital after they are discharged from ICU.

5. CPT Events Table

```

--select * from cptevents
--select cast(cpt.CPT_NUMBER as int) cpt_num
--FROM [dbo].[CPTEVENTS] cpt;

```

subject_id	adm_id	cpt_num
1	108700	2
2	16537	3
3	23031	184185
4	88291	113932
5	6914	134765
6	22507	139627
7	22264	159150
8	10343	160056
9	10463	161731
10	76005	179180
11	10393	190118
12	69776	137906
13	22207	150107
14	31140	163444

In the above figure, we can see that I have used cpt events table in order to calculate the Mean value and then for the cost center we can see that there are only 2 values, so for ICU I have assigned 1 and for the other value I have assigned 2 (RESP), and then I have put them into a temporary table #cpt_table. The column names in the table are: subject_id,adm_id,cpt_num and cost_center.

6. ICUSTAYS Table

Here in the above figure, I have taken into consideration about firstcareunit, lastcareunit, length of stay into consideration and have converted the nominal data into numerical data and then transferred them into a temporary table #icu_table.

7. Service Table

In the two figures below I have converted all the services that a patient goes through in his length of stay. If a person has multiple issues, then he has to go through different services and if the patient has only one service then there would be only one associated with the patient. For example, the services are as follows: GYN, ORTHO, NB, SURG, etc. I have converted them into numbers as seen in the figure and then transferred them into a temporary table #service_table.

File Edit View Query Project Tools Window Help

Object Explorer

LAPTOP-47TKRPSE (SQL Server 15.0.2070.41 - LAPTOP-47TKRPSE\vinna (S1)) - Microsoft SQL Server Management Studio

```

drop table #icu_table

select s.SUBJECT_ID subject_id, s.HADM_ID adm_id,
       case
           when s.PREV_SERVICE = 'TRAUM' then 1
           when s.PREV_SERVICE = 'NSURG' then 2
           when s.PREV_SERVICE = 'NED' then 3
           when s.PREV_SERVICE = 'MED' then 4
           when s.PREV_SERVICE = 'PSVCH' then 5
           when s.PREV_SERVICE = 'NB' then 6
           when s.PREV_SERVICE = 'GU' then 7
           when s.PREV_SERVICE = 'OBS' then 8
           when s.PREV_SERVICE = 'ORTHO' then 9
           when s.PREV_SERVICE = 'ENT' then 10
           when s.PREV_SERVICE = 'VSURG' then 11
           when s.PREV_SERVICE = 'DENT' then 12
           when s.PREV_SERVICE = 'MED' then 13
           when s.PREV_SERVICE = 'PSURG' then 14
           when s.PREV_SERVICE = 'NWED' then 15
           when s.PREV_SERVICE = 'TSURG' then 16
           when s.PREV_SERVICE = 'CEND' then 17
           when s.PREV_SERVICE = 'CRED' then 18
           when s.PREV_SERVICE = 'CSUNG' then 19
           when s.PREV_SERVICE = 'SURG' then 20
           else 21
       end prev_service
      , case
           when s.CURR_SERVICE = 'TRAUM' then 1
           when s.CURR_SERVICE = 'NSURG' then 2

```

Results Messages

subject_id	adm_id	first_wardid	last_wardid	los	first_care	last_care
1	3127	104718	23	23	2.0232	4
2	3127	144842	57	33	2.9432	6
3	3127	167832	23	23	1.1362	4
4	3127	157193	14	14	1.0426	5
5	3128	173717	56	56	2741	3
6	3129	157279	14	23	13.0139	6
7	3129	157279	52	52	1.0426	4
8	3132	158940	52	52	43.9522	4
9	3133	135744	15	57	19.1905	1
10	3133	154403	23	23	4.24	6
11	3133	154403	23	23	31.6057	6

Query executed successfully.

File Edit View Query Project Tools Window Help

Object Explorer

LAPTOP-47TKRPSE (SQL Server 15.0.2070.41 - LAPTOP-47TKRPSE\vinna (S1)) - Microsoft SQL Server Management Studio

```

drop table #icu_table

select s.SUBJECT_ID subject_id, s.HADM_ID adm_id,
       case
           when s.PREV_SERVICE = 'TRAUM' then 1
           when s.PREV_SERVICE = 'NSURG' then 2
           when s.PREV_SERVICE = 'NED' then 3
           when s.PREV_SERVICE = 'MED' then 4
           when s.PREV_SERVICE = 'PSVCH' then 5
           when s.PREV_SERVICE = 'NB' then 6
           when s.PREV_SERVICE = 'GU' then 7
           when s.PREV_SERVICE = 'OBS' then 8
           when s.PREV_SERVICE = 'ORTHO' then 9
           when s.PREV_SERVICE = 'ENT' then 10
           when s.PREV_SERVICE = 'VSURG' then 11
           when s.PREV_SERVICE = 'DENT' then 12
           when s.PREV_SERVICE = 'CRED' then 13
           when s.PREV_SERVICE = 'PSURG' then 14
           when s.PREV_SERVICE = 'NWED' then 15
           when s.PREV_SERVICE = 'TSURG' then 16
           when s.PREV_SERVICE = 'CEND' then 17
           when s.PREV_SERVICE = 'CRED' then 18
           when s.PREV_SERVICE = 'CSUNG' then 19
           when s.PREV_SERVICE = 'SURG' then 20
           else 21
       end prev_service
      , case
           when s.CURR_SERVICE = 'TRAUM' then 1
           when s.CURR_SERVICE = 'NSURG' then 2

```

Results Messages

subject_id	adm_id	prev_service	curr_service
1	3042	159630	21
2	5043	159630	21
3	5044	159631	4
4	9044	134722	21
5	9045	178512	21
6	9046	196105	21
7	9047	169379	21
8	9049	141096	21
9	9051	113986	21
10	9052	152269	21
11	9052	152269	13

Results Messages

subject_id	adm_id	prev_service	curr_service
1	3042	159630	21
2	5043	159630	21
3	5044	159631	4
4	9044	134722	21
5	9045	178512	21
6	9046	196105	21
7	9047	169379	21
8	9049	141096	21
9	9051	113986	21
10	9052	152269	21
11	9052	152269	13

Query executed successfully.

8. Joining all temporary tables

In the below 2 figures we can clearly see that all the temporary tables that I have created have been based on 2 factors and have them as common. hence I have joined them based on subject_id and adm_id.

LAPTOP-47TKRPS5 (SQL Server 15.0.2070.0) - LAPTOP-47TKRPS5\venna (S1) - Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Help

project

Connect to...

project

New Query

Execute

td

Object Explorer

Connect to...

LAPTOP-47TKRPS5 (SQL Server 15.0.2070.0) - LAPTOP-47TKRPS5\venna (S1)

Databases

System Databases

Database Snapshots

Logins

Projects

Tables

System Tables

FileTables

External Tables

Graph

Admissions

dbo.ADMISIONS

dbo.ADMISIONS_TAB_1

dbo.ADMISIONS_TABLE_1

dbo.classify_table_final

dbo.cpt_table

dbo.EVENTS

dbo.DIAGNOSES_ICD

dbo.drt

dbo.final_1_table_1

dbo.final_1_length_of_stay

dbo.final_1_LOS

dbo.final_1_table

dbo.icd9_table

dbo.icu_table

dbo.ICUSTAYS

dbo.PATIENTS

dbo.PROCEDURES_ICD

dbo.readmission1

dbo.service_table

dbo.SERVICES

dbo.td

dbo.temp_table_1

dbo.temp_table_2

dbo.temp_table_3

dbo.temp_table_4

dbo.temp_table_5

Views

External Resources

Synonyms

Premodifiability

Service Broker

Storage

Security

project

109 %

Results Messages

Query executed successfully.

```
select a.* , d.seq_num, d.lidc9_code
into #temp_table_1
from #ADMISSIONS_TABLE_1 a
inner join #td t on a.subject_id = d.subject_id and a.adm_id = d.adm_id;

select * from #temp_table_1

drop table #temp_table_1

-----

select distinct tl.* , s.prev_service, s.curr_service
into #temp_table_2
from #service_table s
inner join temp_table_1 tl on tl.subject_id = s.subject_id and tl.adm_id = s.adm_id;

select * from #temp_table_2

drop table #temp_table_2

-----

select * from #cpt_table;

select t2.* , c.cost_center, c.cpt_num
into #temp_table_3
from #temp_table_2 t2
inner join #cpt_table c on t2.subject_id = c.subject_id and t2.adm_id = c.adm_id;
select * from #temp_table_3;

drop table #temp_table_3

-----

select * from #icu_table;
select t3.* , i.first_wardid, i.last_wardid, i.first_care, i.last_care, i.los
into #temp_table_4
from #temp_table_3 t3
inner join #icu_table i on t3.subject_id = i.subject_id and t3.adm_id = i.adm_id;

select * from #temp_table_4
```

HAP_PROJECT_SQL_QUERY.sql - LAPTOP-47TKRP3E.project (LAPTOP-47TKRP3E\vinu) - Microsoft SQL Server Management Studio

File Edit View Project Tools Window Help

Object Explorer

project

Connect ▾

LAPTOP-47TKRP3E (SQL Server 10.0.2070.41 - LAPTOP-47TKRP3E\vinu)

Connect to database

Databases

System Databases

Database Snapshots

tempdb

project

Database Diagrams

Tables

System Tables

FileTables

External Tables

Graph Tables

dbo.ADMISIONS

dbo.ADMISIONS_TAB_1

dbo.ADMISONS_TABLE_1

dbo.classify_table_1_final

dbo.cpt_table

dbo.CPTEVENTS

dbo.DIAGNOSES_ICD

dbo.dbo_final_1_table_1

dbo.final_Length_of_stay

dbo.final_LOS

dbo.final_table

dbo.ICUSTAYS

dbo.PATIENTS

dbo.PROCEDURES_ICD

dbo.proceduretab1

dbo.procedure_table

dbo.SERVICES

dbo.td

dbo.temp_table_1

dbo.temp_table_2

dbo.temp_table_3

dbo.temp_table_4

dbo.temp_table_5

Videos

External Resources

Synonyms

Programmability

Service Broker

Storage

Security

HAP_PROJECT_SQL_XPSPviews(DT) = x

```
select * from #icu_table
select t3..i.first_wardid, i.last_wardid, i.first_care, i.last_care, i.los
into #temp_table_4
from #temp_table_3 t3
inner join #icu_table i on t3.subject_id = i.subject_id and t3.adm_id = i.adm_id;

select * from #temp_table_4

drop table #temp_table_4

select * from #proceduretab1

select t4.. p.proc_num
into #temp_table_5
from #temp_table_4 t4
inner join #proceduretab1 p on t4.subject_id = p.subject_id and t4.adm_id = p.adm_id;

select * from #temp_table_5

drop table #temp_table_5

select pt.. p.gender, p.dob
into #final_table_1
from #temp_table_5 pt
inner join #pt1 on pt.subject_id = p.subject_id

select * from #final_table_1

drop table final_1_table_1
```

109 %

Results Messages

subject_id	adm_id	admission_location	admission_type	marital_status	insurance	religion	ethnicity	admit_time	discharge_time	seq_num	icd9_code	prev_service	curr_service	cost_center	cpt_num	first_wardid	last_wardid	first_care	last_care	los	
1	28460	13236	2	2	1	5	6	48575	48580	9	830	4	10	2	94002	52	52	4	4	2.1817	
2	28460	13236	2	2	1	6	33	48575	48580	9	830	4	10	2	94002	52	52	4	4	2.1817	
3	28460	13236	2	2	1	6	33	48575	48580	9	830	4	10	1	87944	52	52	4	4	1.7311	
4	28460	13236	2	2	1	5	6	33	48575	48580	9	830	4	10	1	87944	52	52	4	4	2.1817
5	28460	13236	2	2	1	5	6	33	48575	48580	9	830	10	4	2	94002	52	52	4	4	1.7311
6	28460	13236	2	2	1	5	6	33	48575	48580	9	830	10	4	2	94002	52	52	4	4	2.1817

Query executed successfully.

LAPTOP-47TKRP3E (10.0 RTM) LAPTOP-47TKRP3E\vinu... project 00:00:00 63,998 rows

Ready

Time here to search

Quick Launch (Ctrl+F)

Ln 347 Col 1 Ch 1 INS

100% 9:32 PM

9. Final Dataset

The screenshot shows the Microsoft SQL Server Management Studio interface. A query window is open with the following SQL code:

```

drop table final_1_table_1

select f.admit_time - f.dob age, (f.discharge_time - f.admit_time) as full_LOS
into final_1_table_1
from #final_table f where (f.discharge_time - f.admit_time) > 0;
drop table #final_table

select count(*) as full_LOS from final_1_table_1

select f.case when f.full_LOS > 5 then 1 else 0 end LOS_B
into final_LOS
from final_1_table_1 f;

select * from final_LOS
----- Final table generated

```

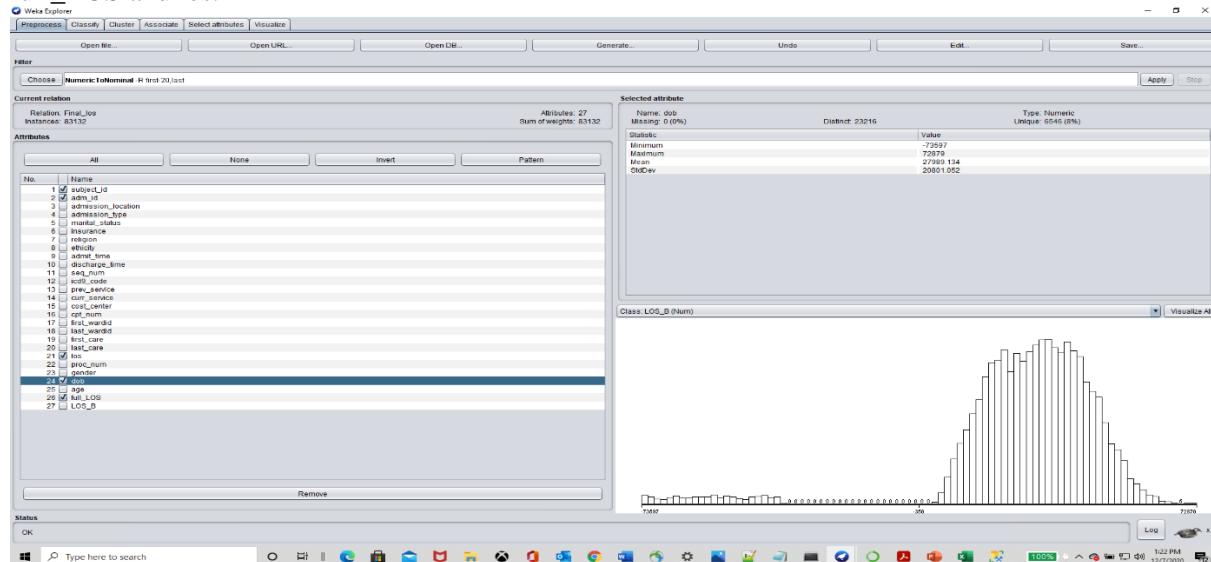
The results pane shows the output of the last select statement, displaying 83,132 rows of data.

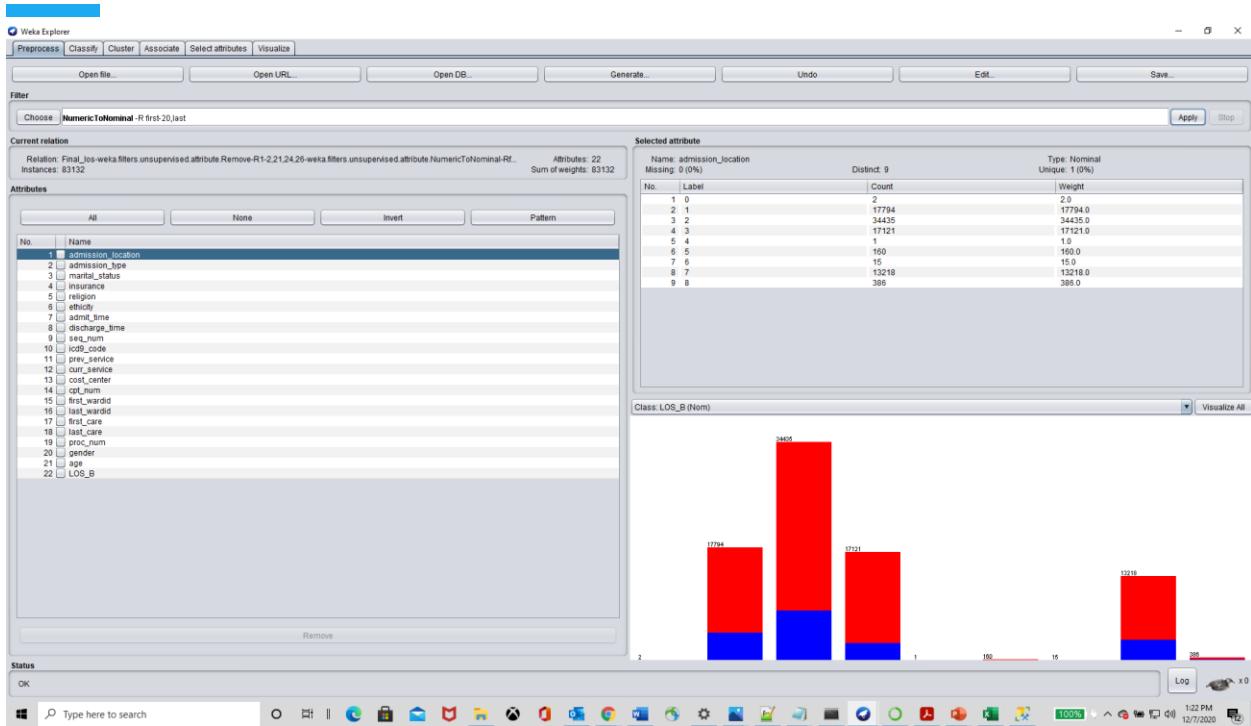
In the final data set, I have considered the length of stay of patients > 5 as 1 and ≤ 5 as 0. For getting to this conclusion, I have subtracted the total length of stay of a patient in the hospital and the length of stay in ICU which gave me the length of stay of patients after they have left the ICU ward.

WEKA

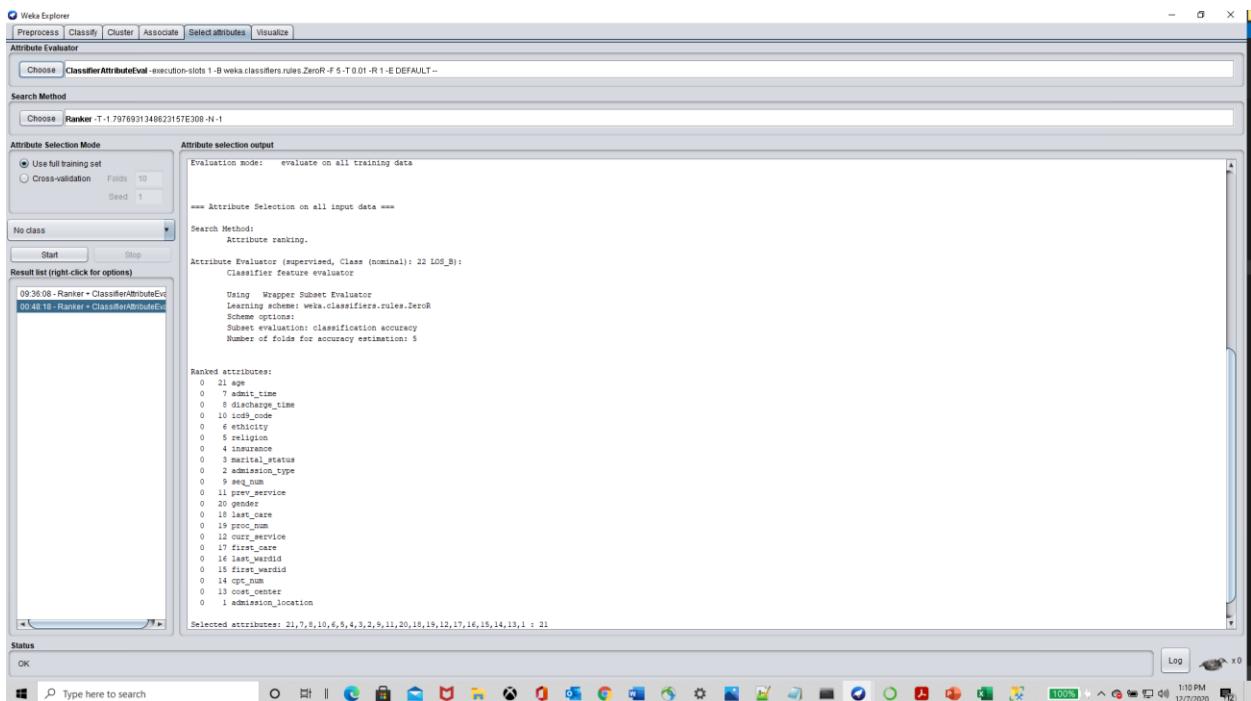
After getting the final dataset from SQL I have exported that data into CSV from SQL and saved in CSV format. Then I have imported that CSV file into weka.

In the above figure, you can see that I have selected a few variables that I feel do not contribute to the prediction of the target variable (LOS). I have removed the following variables: Subject_Id, adm_id, dob, full_LOS and los.





In the above figure, we can see that I have converted the required variables to nominal data that is from the 1st variable to 20 and the last variable. LOS_B is my final target variable that consists of 1 and 0.



In the above figure, we can see that I have used the ranked function in order to rank the attributes according to their contribution towards LOS_B. I thought of considering the top 10 variables and run the models but then that would lead to overfitting the model and hence I have considered all the data that I have selected to predict the LOS_B. There are a total of 21 variables that contribute to the prediction of length of stay.

Machine Learning Models

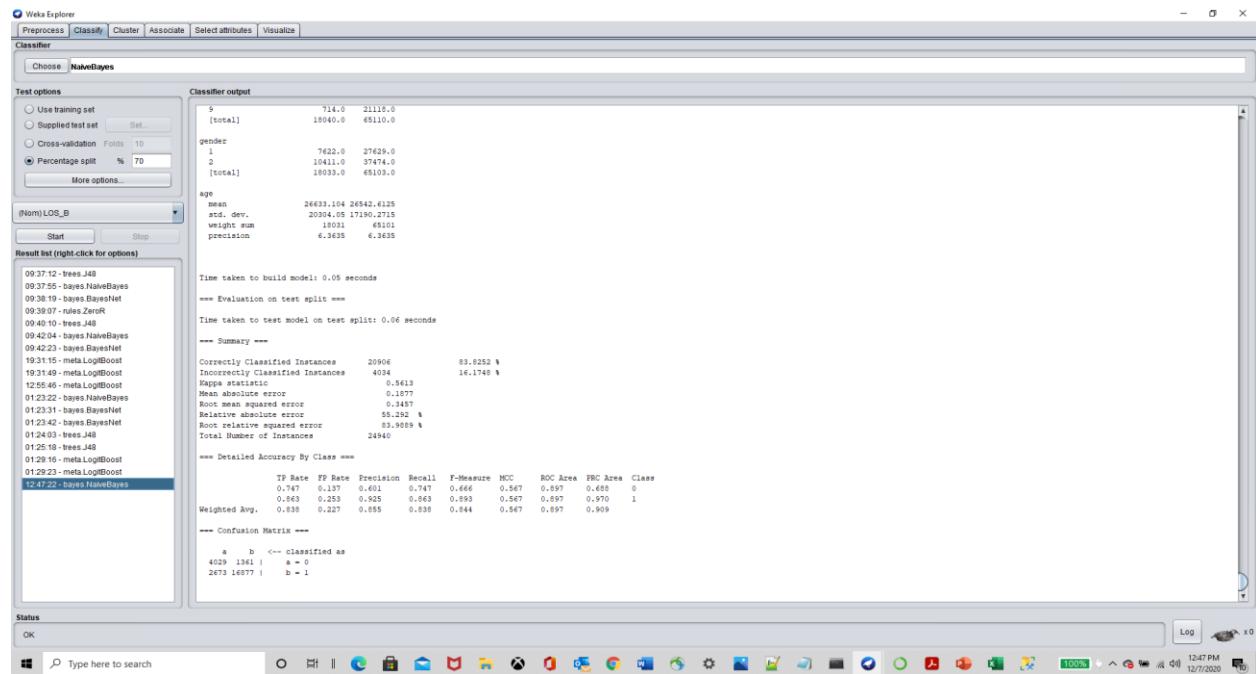
I have used 4 machine learning models for the prediction of LOS.

- Decision Tree Classifier
- Naïve Bayes Classifier
- Bayes Network Classifier
- Additive Logistic Regression

I have used 2 methods to check which method performs better.

- 70 – 30 Split
- 10 – Fold Cross-Validation

Naïve Bayes Classifier 70 – 30 Split



The above figure shows the results of Naïve Bayes using a 70-30 split.

Bayes Net Classifier 70 – 30 Split

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'NaiveBayes'. The 'Test options' section shows a 'Percentage split' of 70%. The 'Classifier output' pane displays the following log:

```

opt_num(1459): LOS_B
first_wardid(0): LOS_B
last_wardid(0): LOS_B
first_care(6): LOS_B
last_care(6): LOS_B
proc_num(1): LOS_B
process(1): LOS_B
age(8): LOS_B
LOS_B(1):
LogScore Bayes: -4841546.999754976
LogScore BPN: -421515.142916592
LogScore HML: -4734670.97935373
LogScore ENTRWF: -5095976.5795303111
LogScore AIC: -8178941.579550111

```

Time taken to build model: 0.18 seconds
 *** Evaluation on test split ***
 Time taken to test model on test split: 0.05 seconds
 *** Summary ***
 Correctly Classified Instances 21107 84.4311 %
 Incorrectly Classified Instances 3833 15.3689 %
 Kappa statistic 0.59
 Mean absolute error 0.1748
 Root mean squared error 0.3384
 Relative absolute error 0.1643 %
 Root relative squared error 0.17332 %
 Total Number of Instances 24940
 *** Detailed Accuracy By Class ***

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0	0.791	0.159	0.612	0.761	0.690	0.599	0.915	0.744	0
1	0.861	0.209	0.937	0.861	0.898	0.599	0.915	0.975	1
Weighted Avg.	0.846	0.194	0.867	0.846	0.853	0.599	0.915	0.925	

 *** Confusion Matrix ***

	a	b	-- classified as
a	4245	1125	a = 0
b	2708	16842	b = 1

The above figure shows the results of Bayes Net Classifier using 70-30 split.

Decision Tree 70 – 30 Split

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'NaiveBayes'. The 'Test options' section shows a 'Percentage split' of 70%. The 'Classifier output' pane displays the following log:

```

proc_num = 1: 0 (8086.0/04894.0)
proc_num = 2: 1 (9121.0/04275.0)
proc_num = 3: 1 (9070.0/01646.0)
proc_num = 4: 1 (7904.0/01883.0)
proc_num = 5: 1 (7766.0/0328.0)
proc_num = 6: 1 (7764.0/01059.0)
proc_num = 7: 1 (7764.0/01120.0)
proc_num = 8: 1 (5407.0/01313.0)
proc_num = 9: 1 (21830.0/713.0)

Number of Leaves : 9
Size of the tree : 10

Time taken to build model: 1.31 seconds  

*** Evaluation on test split ***  

Time taken to test model on test split: 0.01 seconds  

*** Summary ***  

Correctly Classified Instances 19791 79.3144 %
Incorrectly Classified Instances 5159 20.6856 %
Kappa statistic 0.2398
Mean absolute error 0.2781
Root mean squared error 0.3719
Relative absolute error 0.11937 %
Root relative squared error 0.10361 %
Total Number of Instances 24940  

*** Detailed Accuracy By Class ***  


|               | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC   | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| 0             | 0.248   | 0.057   | 0.547     | 0.248  | 0.341     | 0.265 | 0.786    | 0.449    | 0     |
| 1             | 0.943   | 0.752   | 0.820     | 0.943  | 0.877     | 0.245 | 0.786    | 0.917    | 1     |
| Weighted Avg. | 0.793   | 0.602   | 0.761     | 0.793  | 0.761     | 0.245 | 0.786    | 0.816    |       |

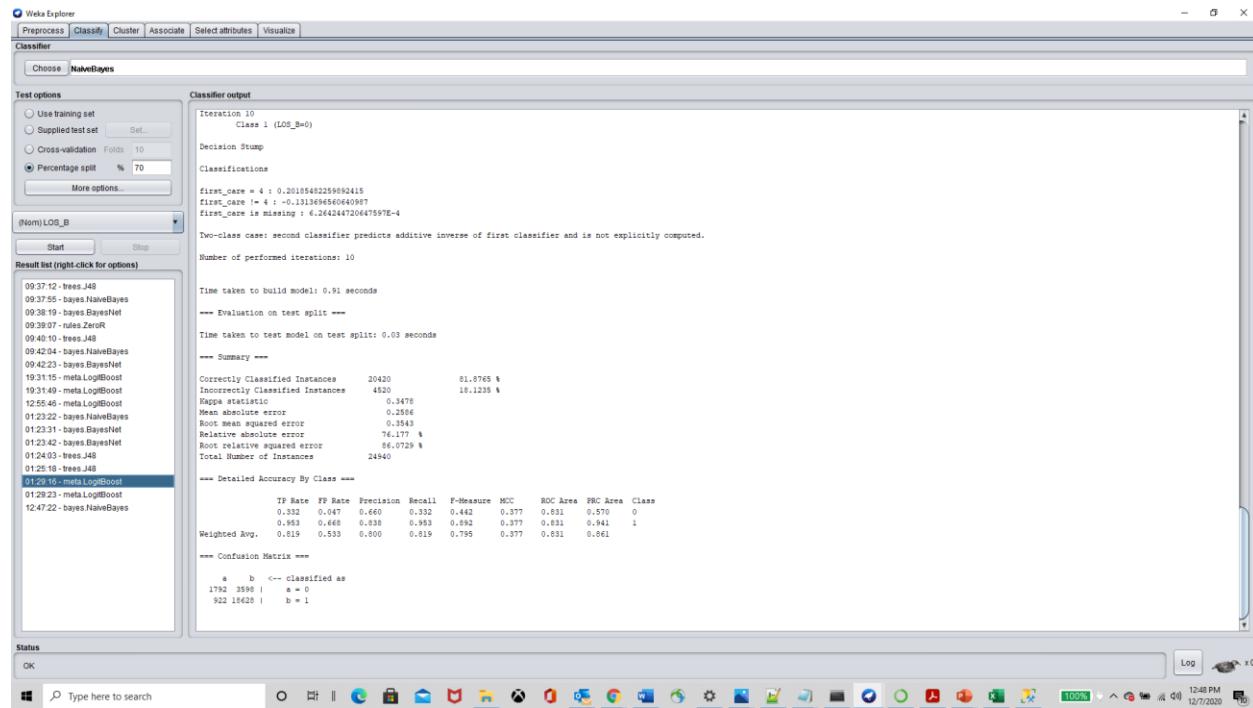

    *** Confusion Matrix ***  


|   | a    | b     | -- classified as |
|---|------|-------|------------------|
| a | 1337 | 4053  | a = 0            |
| b | 1106 | 18444 | b = 1            |


```

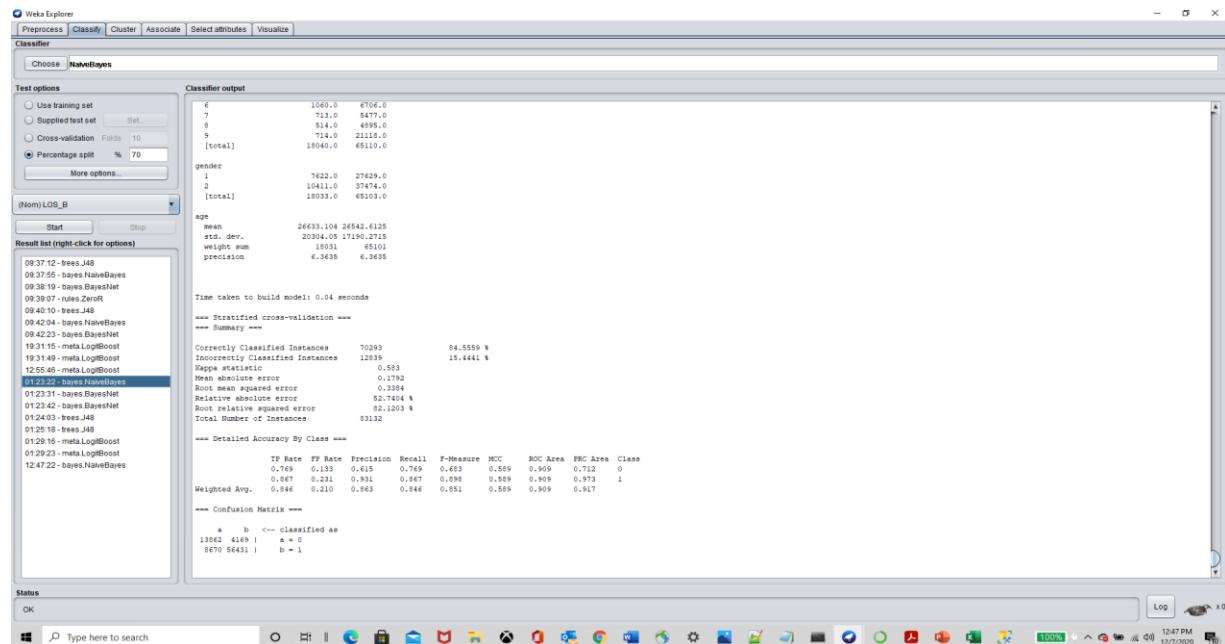
The above figure shows the results of the Decision Tree using a 70-30 split.

Additive Logistic Regression (Logit Boost) 70 – 30 Split.



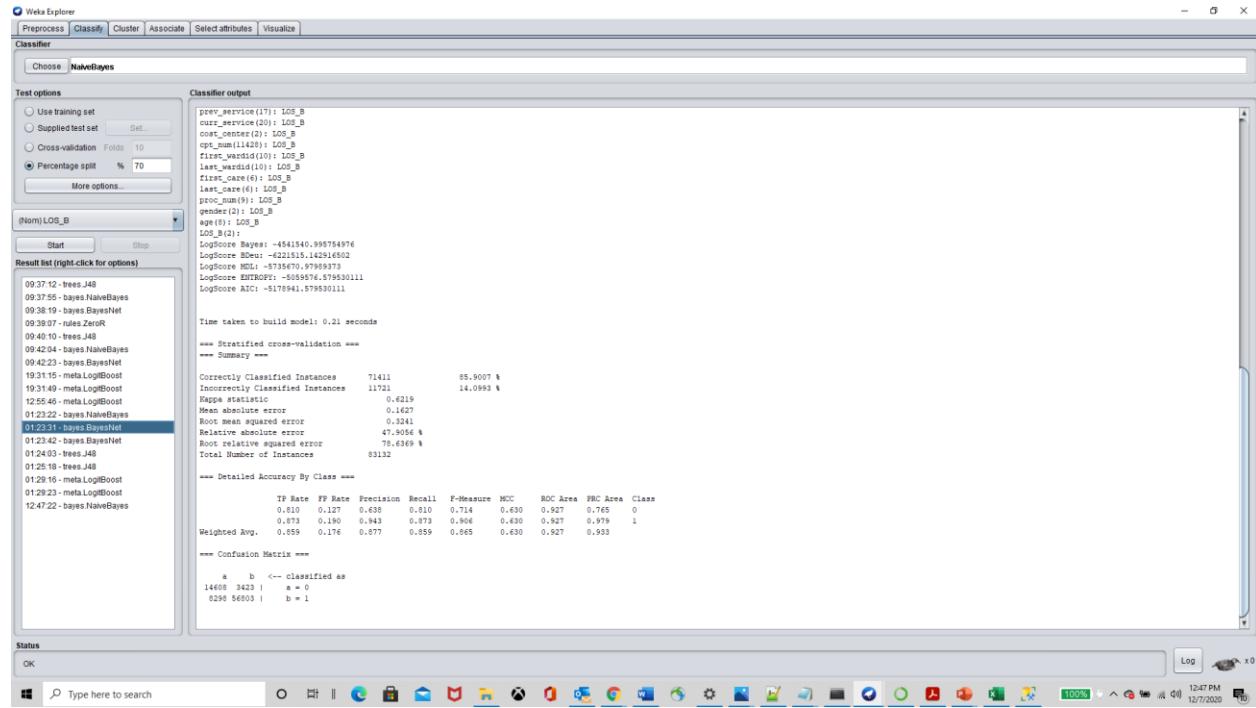
The above figure shows the results of the Logit Boost using a 70-30 split.

Naïve Bayes Classifier 10 - Fold Cross-Validation



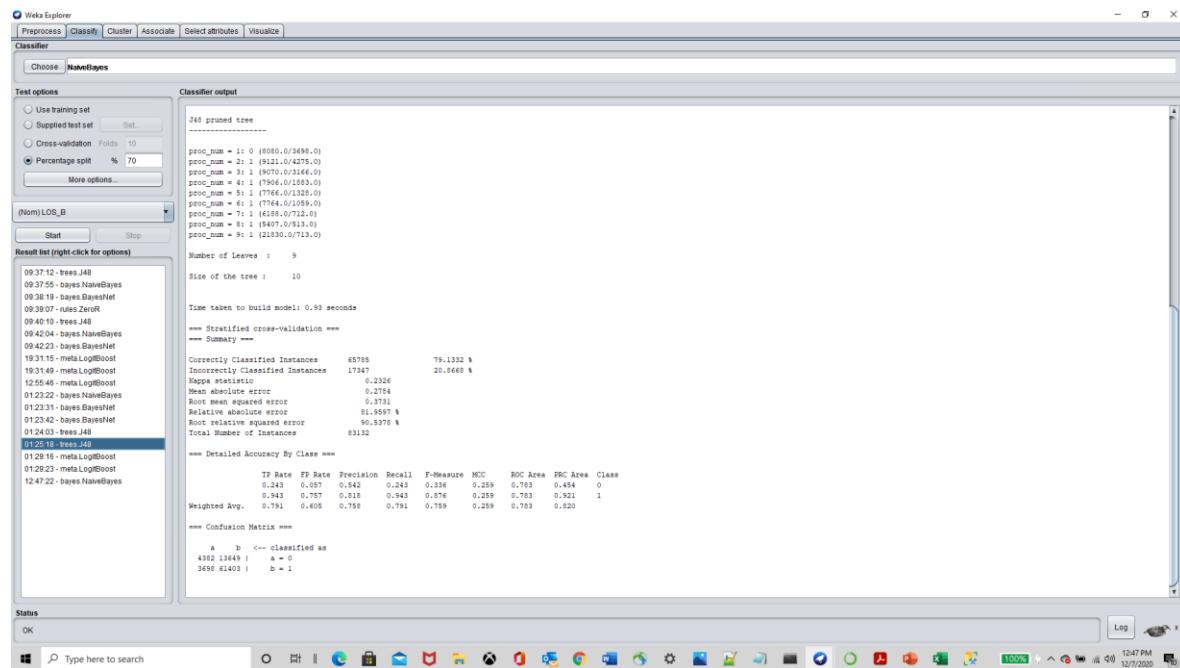
The above figure shows the results of Naïve Bayes using 10 – Fold cross-validation.

Bayes Net Classifier 10 - Fold Cross-Validation



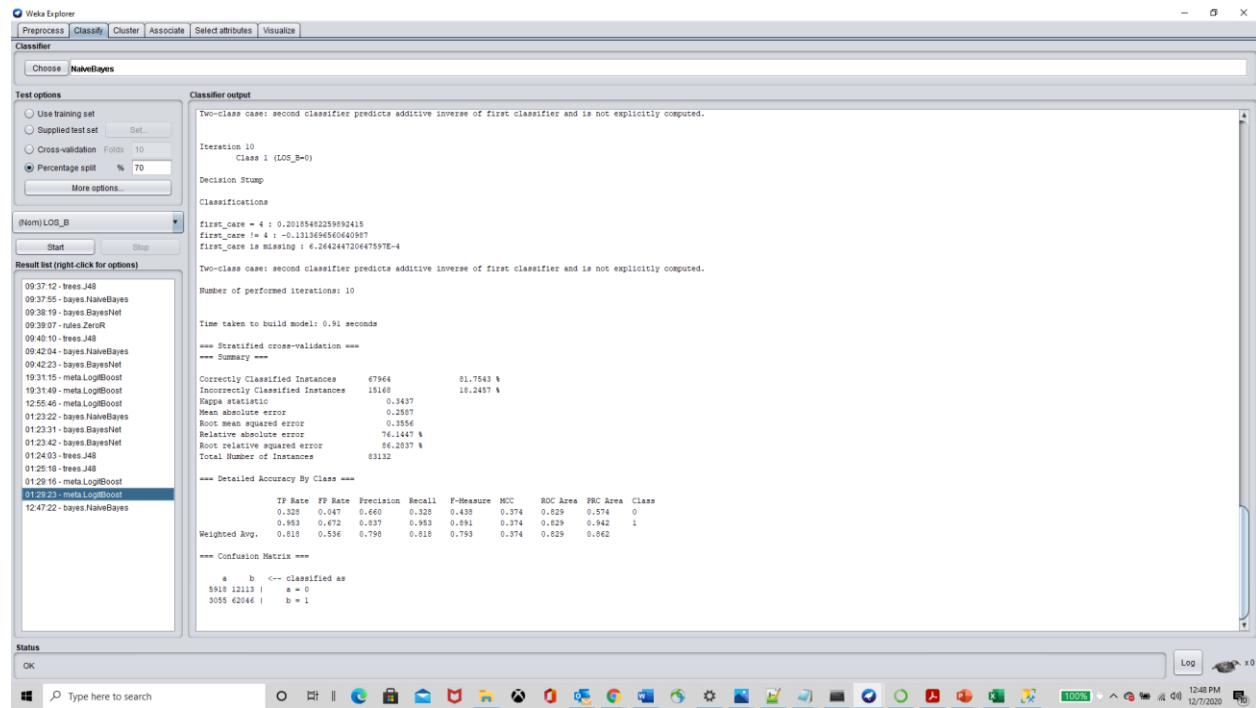
The above figure shows the results of Bayes Net using 10 – Fold cross-validation.

Decision Tree Classifier 10 - Fold Cross-Validation



The above figure shows the results of the Decision Tree using 10 – Fold cross-validation.

Additive Logistic Regression Classifier 10 - Fold Cross-Validation



The above figure shows the results of the Additive Logistic Regression (Logit Boost) using 10 – Fold cross-validation.

RESULTS

70 - 30 Split

Model	Accuracy	Recall	Model	ROC	Precision	Recall
J48	79%	94.3%	J48	0.786	82.0 %	94.3%
NaiveBayes	83%	86.3%	NaiveBayes	0.897	92.5%	86.3%
BayesNet	84%	86.1%	BayesNet	0.915	93.7%	86.1%
LogitBoost	81%	95.3%	LogitBoost	0.831	83.5%	95.3%

10 – FOLD CROSS VALIDATION

Model	ROC	Precision	Recall
J48	0.783	81.8%	94.3%
NaiveBayes	0.909	93.1%	86.7%
BayesNet	0.927	94.3%	87.3%
LogitBoost	0.829	83.7	95.3%

Conclusion

From the above, we can see that the accuracy is highest in the Bayes Net classifier using 10 – Fold Cross-validation. But I am more interested in understanding the recall of the model as when we check the model, we can understand that recall tells us that there is a chance of predicting that the patient is in the hospital after getting discharged from the ICU for less than or equal to 5 days but, the patient is actually in the hospital for more than 5 days. I would like to get a recall of 100% as it matches my problem statement that if the patient is in the hospital for more than 5 days after getting discharged from ICU then there is a great probability that the patient might get back to the ICU due to any of the multiple issues. Hence my model should not misclassify this part. When we check the precision, it says in the prediction that there are chances that the patient is in the hospital after ICU for > 5 days but actually, the patient is < or = 5 days. This case could be considered to have some error as it could not lead to/implicate serious complications to the patient's health.

In my opinion, the best model that gives me the best recall value and it is Logit Boost using 10 – Fold Cross-Validation and it also has an accuracy of 81% which is not far away from the model that has the highest accuracy (85%). Hence, I would choose this model to give it to the administration department of the hospital as it would help them in keeping the length of stay of a patient after getting discharged from the ICU ward as quickly as possible with proper treatment else if it is more than 5 days then proper attention should be given to the patient, which actually requires an experienced staff of doctors and nurse.

Future Work

In future I would like to understand the length of stay of patients with multiple diseases and single disease. Further, I would also like to explore more on accident-related cases and disease-ridden cases, which of these categories lead to greater length of stay in hospitals, that is calculated after they are discharged from the ICU or their complete hospital stay.

References

- [1] "MainHealth," [Online]. Available: <https://www.mainehealth.org/Services/Hospital-Medicine/Critical-Care-Unit-Intensive-Care-Unit>.
- [2] "USCF," [Online]. Available: <https://anesthesia.ucsf.edu/next-steps-after-icu#:~:text=After%20the%20ICU%2C%20patients%20usually,floor%20and%20then%20hopefully%20home..>
- [3] "Mimic III," [Online]. Available: <https://physionet.org/>.
]
- [4] "Society of critical Care medicine," [Online]. Available:
] <https://www.sccm.org/MyICUCare/THRIVE/Post-intensive-Care-Syndrome>.
- [5] "NIH," [Online]. Available:
] <https://pubmed.ncbi.nlm.nih.gov/24776831/#:~:text=Abstract,in%20a%20non%20ICU%20bed>.
.