HOMEWORK 1

2.

a) **SQL Query:**

```
SELECT

ad.subject_id as patient_id , x.age, ad.INSURANCE, ad.LANGUAGE, ad.RELIGION,
ad.MARITAL_STATUS, ad.ETHNICITY from admissions ad

INNER JOIN
 (
SELECT p.subject_id, MIN( ROUND( (cast(admittime as date) - cast(dob as
date)) / 365.242,0) ) AS age
FROM patients p
INNER JOIN admissions a
ON p.subject_id = a.subject_id
GROUP BY p.subject_id having p.subject_id=40080
ORDER BY p.subject_id )x
ON ad. subject_id=x.subject_id
```

Output:

patient_id	age	insurance	language	religion	marital_status	ethnicity
40080	79	Medicaid	HAIT	UNOBTAINABLE	WIDOWED	BLACK/AFRICAN AMERICAN

b) **SQL Query:**

```
select icd.subject_id as patient_id, icd.icd9_code,dicd.short_title as
short_diagnosis, dicd.long_title as long_diagnosis from diagnoses_icd  icd
inner join d_icd_diagnoses dicd  on icd.icd9_code=dicd.icd9_code where
icd.subject_id=40080 and icd.seq_num=1
```

Output:

patient_id	icd9_code	short_diagnosis	long_diagnosis
40080	42843	Ac/chr syst/dia hrt fail	Acute on chronic combined systolic and diastolic heart failure

c) **SQL Query:**

```
select patient_id,los as duration_stay,category,description,
discharge_condition from

(
(select i.subject_id as patient_id, i.los,n.category,n.description from
icustays i inner join noteevents n
on i.subject_id=n.subject_id where i.subject_id=40080 and lower(category)
like '%discharge%')los
inner join
```

```
(select split_part(a,'Discharge Instructions:',1) as Discharge_condition from
(select split_part(text,'Discharge Condition:',2) as a from noteevents
where lower(category) like '%discharge%' and subject_id=40080 )x
)di
on 1=1)
```

Output:

patient_id	duration_stay	category	description	discharge_condition
40080	4.8577	Discharge summary	Report	Mental Status: Minimally clear and coherent. Level of Consciousness: Minimally Alert and somewhat interactive. Activity Status: Bed bound - dependent hemiplegia.

d) **SQL Query:**

```
select ce.subject_id as patient_id, ce.hadm_id,max(ce.valuenum) as
max_heartrate,min(ce.valuenum) as min_heartrate  from chartevents ce
inner join d_items  di on ce.itemid=di.itemid where lower(label) ='heart
rate'  and subject_id=40080 group by 1,2
```

Output:

patient_id	hadm_id	max_heartrate	min_heartrate
40080	162107	141.0	80.0

3.

a) We can get the following insights looking at the graph:

- This plot tells us that nearly 1/5th of total patients are infants
- Excluding infants, the distribution is roughly symmetric(excluding outliers) and left skewed, and the values fall between approximately 50 and 80. This intuitively suggests that most of the patient admitted in the critical care unit are people above age of 50 or are infants because young people tend to be healthier
- The graph shows a tendency of a greater number of people being admitted in critical care as age increases and after reaching age of 80 since there are very less people alive after that age the frequency decreases
  Please note that the age of 300 signifies age greater than> 89

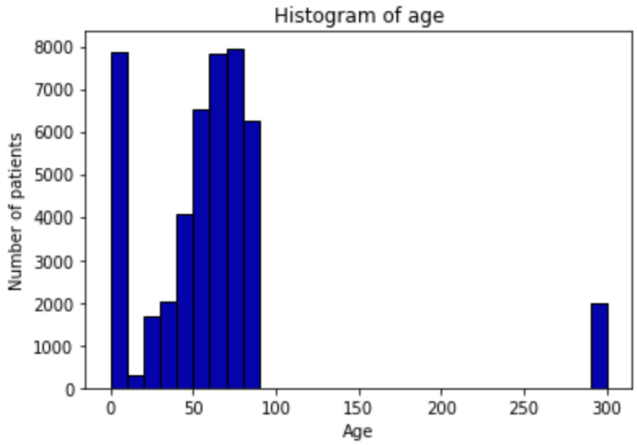

Figure 1

b) We can get the following insights from the following histogram:

- The histogram has very less frequency at very low heart rates and very high rates suggesting that very few patients have abnormal heart rates
- Most of the values lie between approximately 50-100 which is the normal heart rate

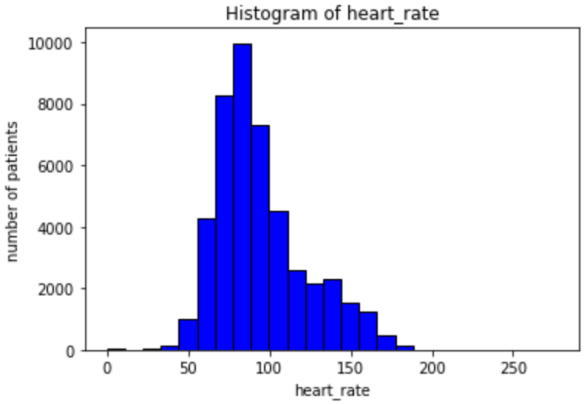

Figure 2

c) We can get the following insights from the following histogram:

- There is not a very large change in heart_rate as age increases
- Older people have wider range of heart rates

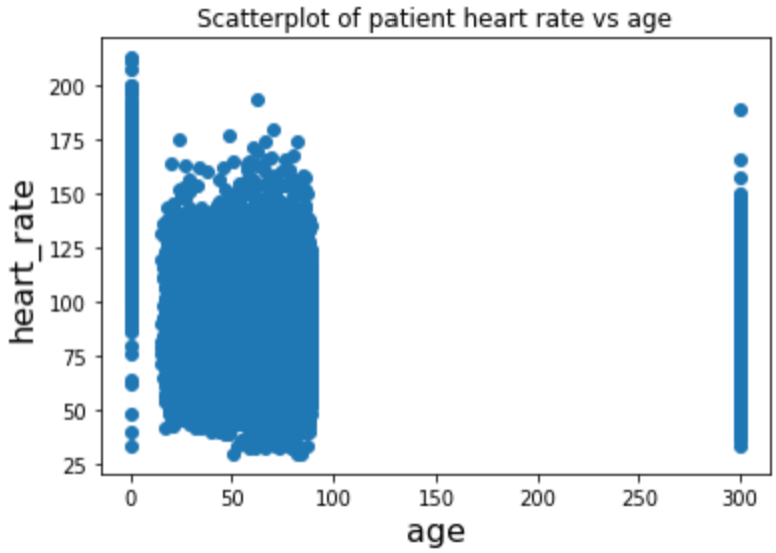

Figure 3

4

a)

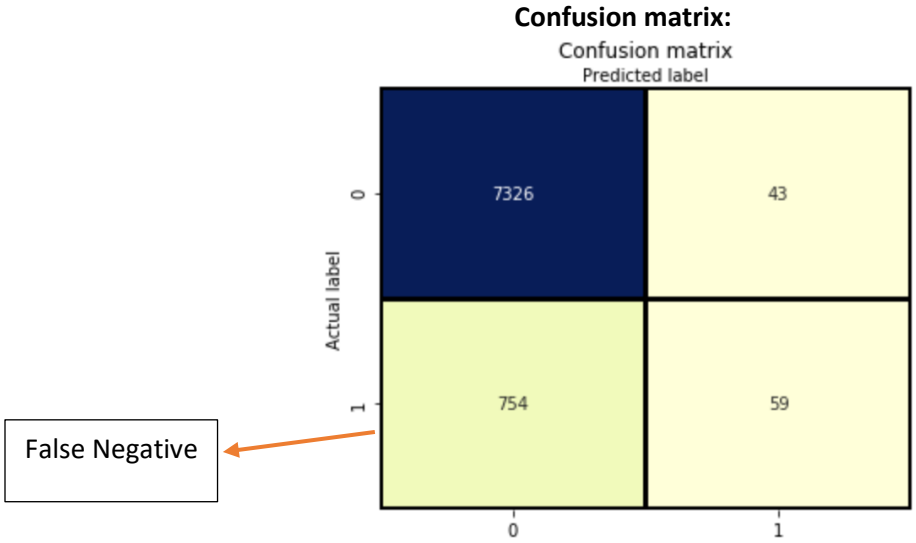

Figure 4: Confusion matrix for the model trained on adult_icu for predicting mortality

ROC Graph and AUC score:

ROC Curves summarize the trade-off between the true positive rate and false positive rate

AUC gives the rate of successful classification by the logistic model

AUC score: 0.7787

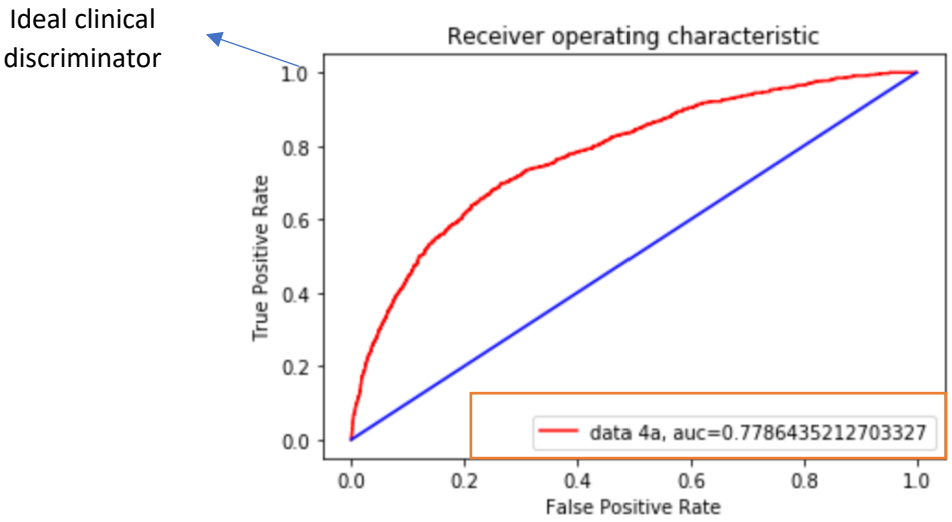

Figure 5: ROC graph for the model trained on adult_icu for predicting mortality

Other metrics that can be used:

```
Accuracy: 0.9025910535321438
Precision: 0.5784313725490197
Recall: 0.07257072570725707
No. of iterations to converge:  [231]
```

- The model has and AUC score of 0.78. This is a good score because a completely random classifier has AUC = 0.5.
- The model has a good accuracy and it can label 90 % data correctly but there is class imbalance in the dataset we used so accuracy is not an appropriate metric here

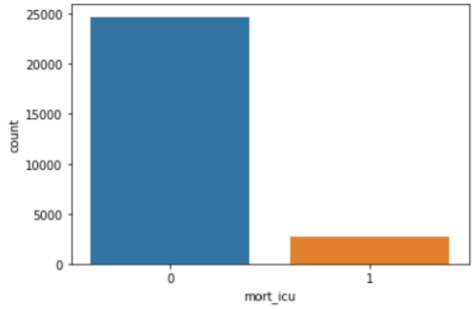

Figure 6: Analyzing data on the basis of mortality

- Since, the model is predicting mortality we should have very less false negative values, because giving extra care to patients based on their vitals and lab measurement is an important decision. As seen from figure 4 the model has 754 false negatives which means we misclassified them as safe.
- We can see that the Recall value is very low, and we should focus on identifying the positive cases because of the critical situation in ICU
- Overall the model has a moderate AUC score and weighing other factors it is not deployable

```
Top 5 risk factors:
```

	Risk factors	coef
29	tempc_max	2.315990
27	resprate_mean	2.345583
46	lactate	2.560961
28	tempc_min	3.286081
40	bilirubin	3.557763

- **Bilirubin:** is an orange-yellow substance made during the normal breakdown of red blood cells. Higher than normal levels of bilirubin may indicate different types of liver problems and can be toxic to nerves and cause brain damage. Hence, hence they are a crucial factor in determining mortality
- **Tempc_min:** When your body temperature drops, your heart, nervous system and other organs can't work normally, and the term given for it is hypothermia.
- **Lactate**: The lactate test is mainly requested to help detect and measure the severity of low levels of oxygen in the body (hypoxia). Any disorder that causes an imbalance between lactate production and clearance can lead to lactic acidosis, a serious and sometimes life-threatening condition.
- **Resprate_mean**: It is the Mean respiration rate over patients first 24 hours post admission. A change in respiratory rate is first sign of deterioration.
- **Tempc_max**: High body temperature is associated with increased heart and respiratory rates and, at extreme levels, damage to the brain, heart, lungs, kidneys, and liver. Thus, this is an important vital in determining mortality.

```
Lowest 5 risk:
```

	Risk factors	coef
11	admType_NEWBORN	0.000000
8	eth_white	0.000928
36	glucose_mean	0.001172
3	adult_icu	0.006736
1	first_hosp_stay	0.006943

- **admType_NEWBORN:** Admission type as new-born pertains to patients' birth and is not an important determinant of mortality because it is an event of birth.
- **eth_white:** Boolean value identifying whether person belong to white ethnicity or not. It should not play a major role in predicting mortality because it is just label of race.
- **glucose_mean:** Mean of Glucose levels across all readings is also not very severe for a person's health (can be controlled) hence they contribute less towards predicting mortality
- **adult_icu**: It is a binary variable confirming that the care unit is not neonatal intensive care unit and Pediatric intensive care unit. This is a label of type if ICU and should not play a very major role in identifying mortality
- **first_hosp_stay:** Binary variable indicating the first hospital stay amongst all the records for a patient

b) The confusion matrix is as follows:

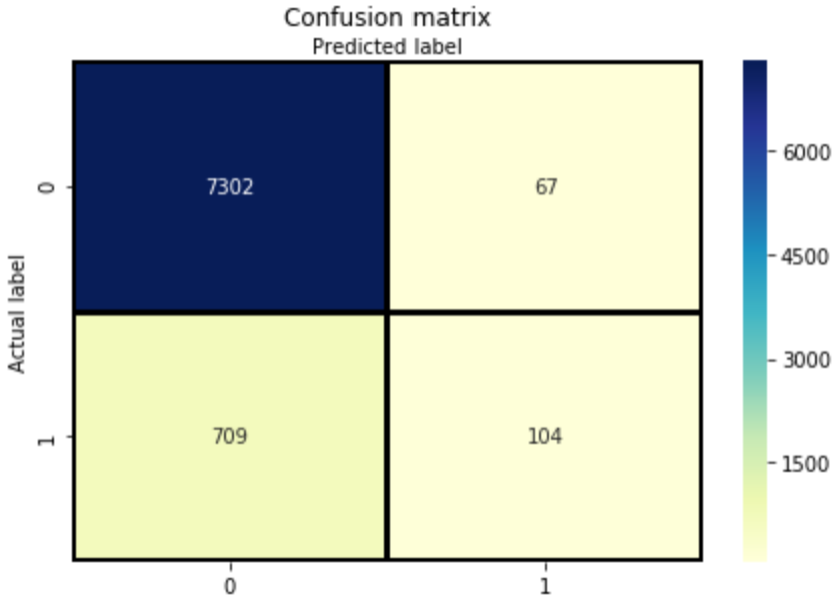

Figure 7: Confusion matrix of model trained on adult_notes for predicting mortality

ROC Graph:

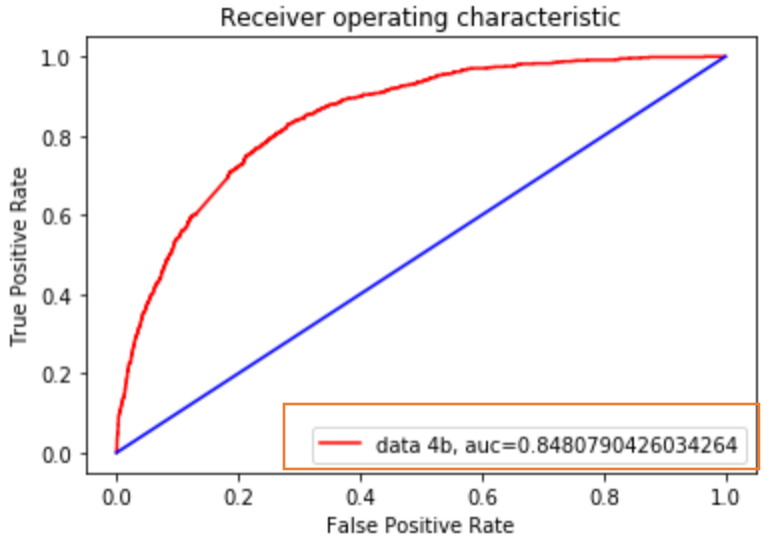

Figure 8: ROC graph for the model trained on adult_icu for predicting mortality

Top 5 words associated with high mortality

	Words	coef
171577	worsening	6.520582
91906	dnr	8.113452
81951	cmo	8.983028
85111	corneal	9.115000
140195	prognosis	12.320256

Lowest 5 words associated with high mortality

	Words	coef
98929	extubation	-14.961711
90012	diet	-12.259628
81454	clear	-10.886578
98916	extubated	-6.513061
128094	normal	-5.778392

c)

The AUC score is: 0.8612316113661884

ROC graph:

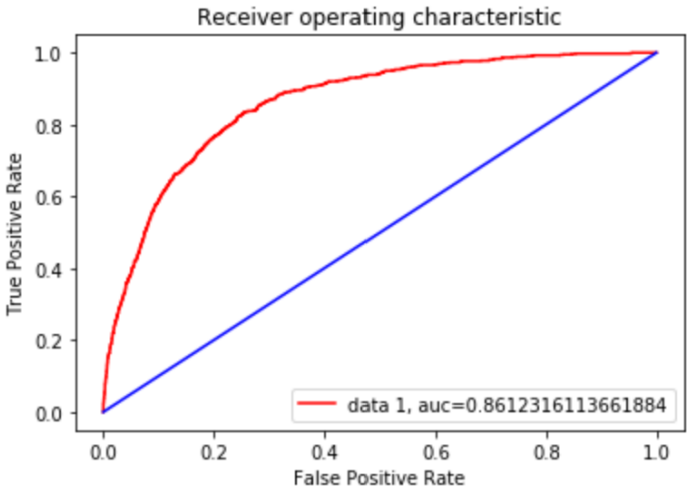

Figure 9: ROC Graph for model trained on both adult_icu and adult_notes

When we combine the models here, we are essentially making ensemble of classifiers. We have trained the models independently on different subsets of data (tabular data and clinical notes) and than combined their results by averaging the result from both the models. This method is known as bagging and it aims at reducing variance by averaging the prediction of each classifier and it helps in reducing overfitting which can be there with large number of parameters. Hence, we see an increase in AUC score.

5

a) **AUC and F1 scores of models trained by logistic regression:**

- **Heart rate:**

```
For measurement of HEART_RATE
AUC:  0.5216536470428943
F1 Score 0.0
```

- **Respiratory Rate:**

```
For measurement of RESPIRATORY_RATE
AUC:  0.5270621743398866
F1 Score 0.0021329541414859578
```

- **O2 saturation:**

```
For measurement of  O2_SATURATION
AUC:  0.5087042872546437
F1 Score 0.0
```

- **Blood pressure:**

```
For measurement of BLOOD_PRESSURE
AUC:  0.5254656624455121
F1 Score 0.059080228264518295
```

b) **AUC and F1 scores of models trained by LSTM:**

- **Heart rate:**

```
For measurement of HEART_RATE
AUC:  0.5201513753642446
F1 Score 0.0
```

- **Respiratory Rate:**

```
For measurement of RESPIRATORY_RATE
AUC:  0.5434834190626417
F1 Score 0.000711490572749911
```

- **O2 saturation:**

```
For measurement of O2_SATURATION
AUC:  0.5169902536059096
F1 Score 0.009167842031029619
```

- **Blood pressure:**

```
For measurement of BLOOD_PRESSURE
AUC:  0.5308644728846041
F1 Score 0.001439366678661389
```

c) F1 score is the harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases. Since, this is a hypertension prediction problem False Negatives are crucial and F1 is a good measure.

- **For linear regression:** Blood pressure have the highest F1 score
  This is intuitive because hypertension is defined as having a blood pressure reading of more than 140/90 mmHg over several weeks.
- **For LSTM:** O2 saturation have the highest F1 score
  This makes sense because the single most important cause of pulmonary hypertension is the narrowing of the pulmonary arteries that occurs as a result of low blood oxygen level.

The performance of both the models is not good. Linear regression models which are used as baseline have AUC score near to 0.5 which is not better than by chance score. This maybe the case because minimum, maximum and average values do not provide the complete information about a patient.

LSTM models have same performance as the baseline model. The performance of LSTM did not improve even after:

- Increasing number of layers
- Different number of neurons
- Different epochs
- Different batch size used for training

This maybe because the measurements are not taken at regular intervals.