**Predicting Hospital Length of Stay**

**HAP780 – Data Mining in Health Care**

**Volgenau School of Engineering, George Mason University**

**Chandrika Amarkhed - G01099774**

**1. Abstract**

This project predicts the duration a patient would stay at the hospital – LOS (Length of Stay) within 24 hours of admission. Length of stay is calculated as the time between hospital admission time and discharge time calculated in days. Data mining models are trained to predict if a patient stays more than 8 days or less than 8 days (8 days being the median stay of patients at the hospital). The data for the study is taken from MIMIC-III database. The queries are run in SQL Server and data mining models are run on Weka.

**Key Words**: MIMIC-III, Random Forest, Naïve Bayes, Logistic Regression, Length-of-stay, Data mining.

**2. Introduction**

Having information like how long a patient would stay at a hospital could be very useful in various ways. The hospitals can schedule and maintain wards and rooms accordingly, they can also schedule resources like (doctors, nurses) accordingly and pre-schedules procedures/surgeries can be planned more efficiently. Patients with the risk of long hospital stays can be treated better, treatments can be adjusted in ways that they are not at risk of any hospital-acquired conditions and their stay can be made more comfortable.

An estimate of length of stay can also benefit patients. They can plan their stay more efficiently and they can also get an estimate of how much the stay would cost them and they can make financial arrangements accordingly.

Predictive analytics is an important and rapidly growing tool in health analytics and the numerous machine learning(ML) methods available can help healthcare provides predict diseases, recommend treatments and plan stays and resources. This project focuses on the length of stay(LOS) of patients at the hospital. The prediction of the stay is made in 24 hours of admission.

Data from the MIMIC-III data set was chosen for this study since there is sufficient amount of information and the data is openly available. Machine learning models Logistic Regression, Random Forest and Naïve Bayes are used to predict the length of stay.

**3. Dataset**

**3.1 Description**

The data for this project was taken from MIMIC-III database. MIMIC – III (Medical Information Mart for Intensive Care III) is a freely available database containing information about 46,520 patients and information about their stays, vitals, demographics, laboratory test, procedures, etc who stayed in critical care units of the Beth Israel Deaconess Medical center and the data consists of data between 2001 and 2012.

**3.2 Selection**

One of the many challenges faced in predicting the length of stay in a hospital is determining the right factors that influence the prediction. Out of the 26 tables in the MIMIC-III dataset, this project considers information from the following tables for the study:

1. PATIENTS: Contains information about patients like Gender, Date Of Birth, Date of Death.
2. ADMISSIONS: Contains information about patient’s admission to the hospital. This table gives information about admission and discharge time, admission type, patient religion, ethnicity, marital status, insurance, and diagnosis.
3. SERVICES: Gives a list of current and previous services given to a patient which could be useful to understand if a patient stays for longer at the hospital.
4. ICUSTAYS: Gives the list of ICU stays at the hospital. This table gives information about LOS in ICUs and the care units assigned to patients.
5. PROCEDURES\_ICD: Gives the information about the ICD procedures for patients.
6. DIAGNOSES\_ICD: Gives information about the ICD9 Codes truncated to first three characters.

**4. Data Preprocessing**

The selected six tables from MIMIC dataset where preprocessed into formats required for the data mining models. The tables were then joined to get all the data required for the project.

Each table is examined and the attributes that seem to be relevant for the prediction are extracted.

1. PATIENTS – From the patients table, values like gender, and date of birth are extracted. Age is determined using DOB and the first admission date, and the gender attribute is converted to dummy variables.
2. ADMISSIONS – Attributes admission and discharge time, admission type, insurance, religion, marital status and ethnicity are extracted from the table. A new attributes LOS (Length of stay) is calculated using admission and discharge time. Ethnicity attribute has over 40 categories which are converted to more general categories. The NULL values in marital status are convert to the unknown category.

All the above categorical attributes are converted to dummy columns having 0s and 1s.

The below figure 4.1 show snapshots of the generalized categories created and the dummy columns created for each.

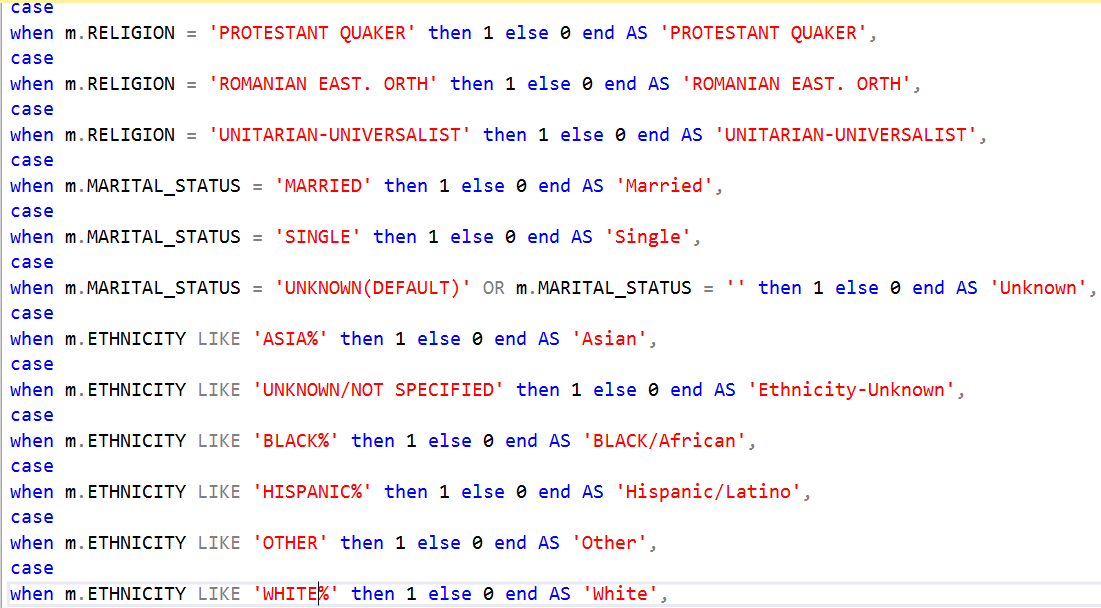
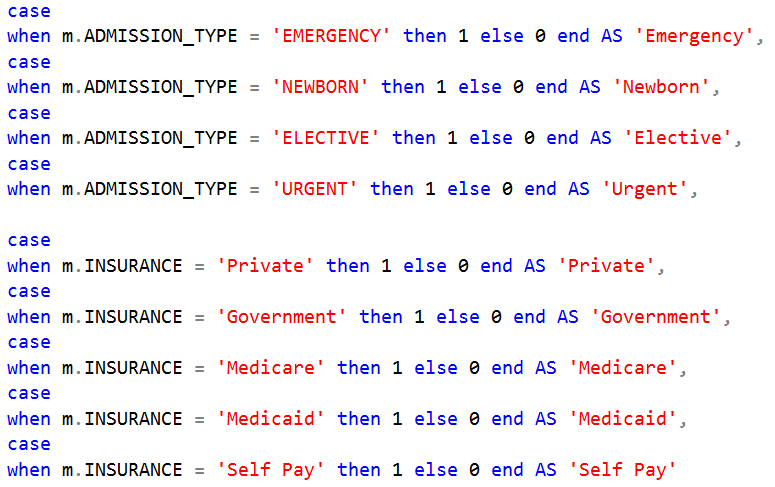
 

Fig. 4.1

1. ICUSTAYS – From this table, The stay information with the first 24 hours are considered and joined to the admissions table. The LOC in ICU is also considered for these records.
2. SERVICES – The services provided to the patients in the first 24 hours are considered from this table.
3. Procedures ICD – The number of procedures done on a patient is calculated from this table.
4. Diagnoses ICD – ICD codes were converted to 17 general ICD categories. ICD codes with V and E were deleted and not considered. The categorical columns were then converted into dummy columns.

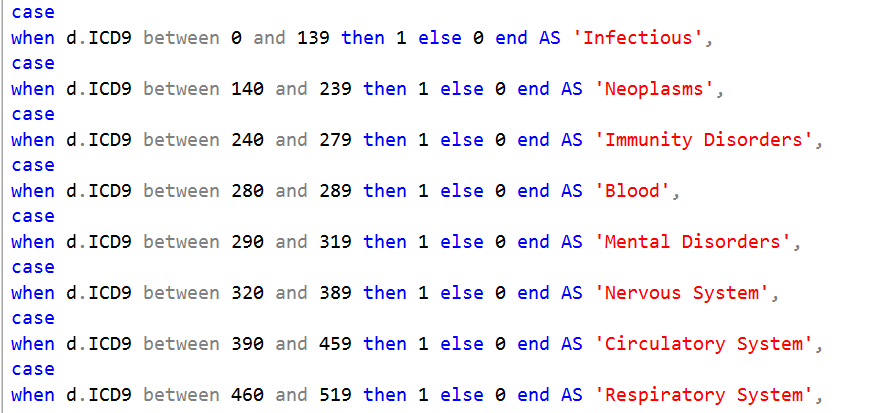
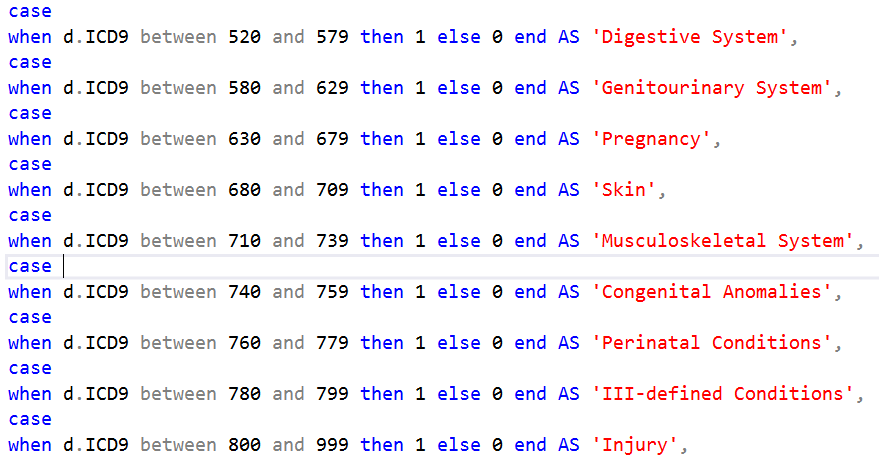
 

Fig 4.2 ICD9\_CODES

The above figure 4.2 shows the generalized categories created for the ICD9\_Codes.

The Patient and admissions tables were joined first and then the other tables are each joined after preprocessing and selecting relevant columns. Fig 4.3 shows the tables that are joined to create one dataset that is used for the predictive models.

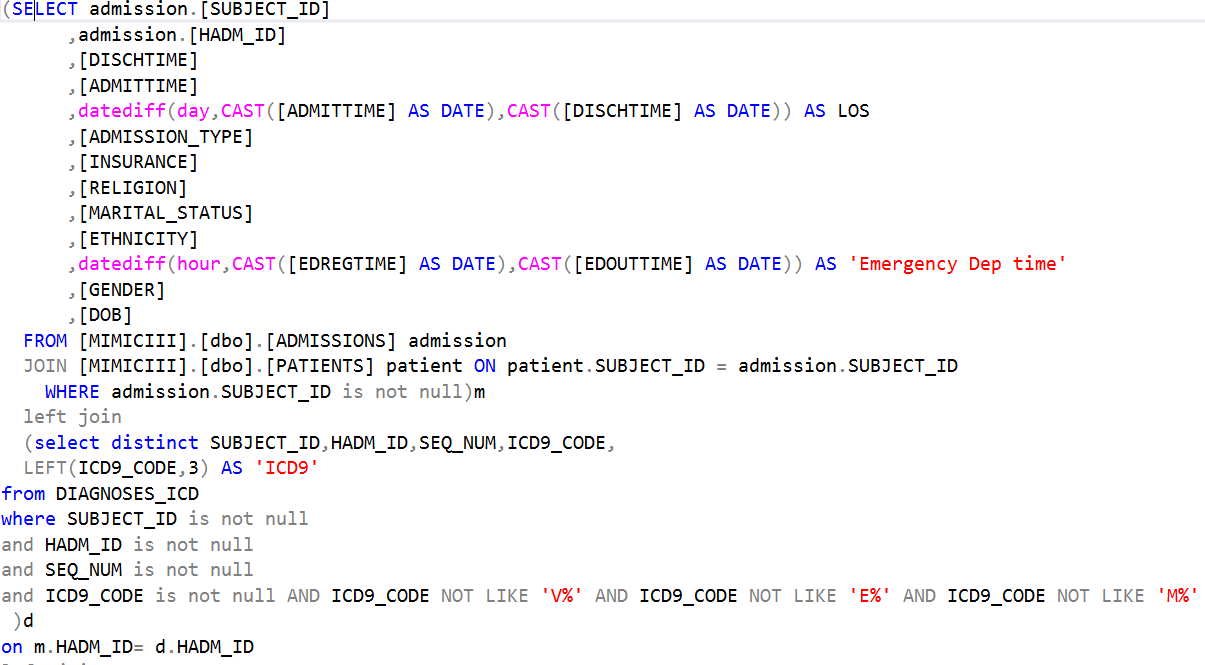
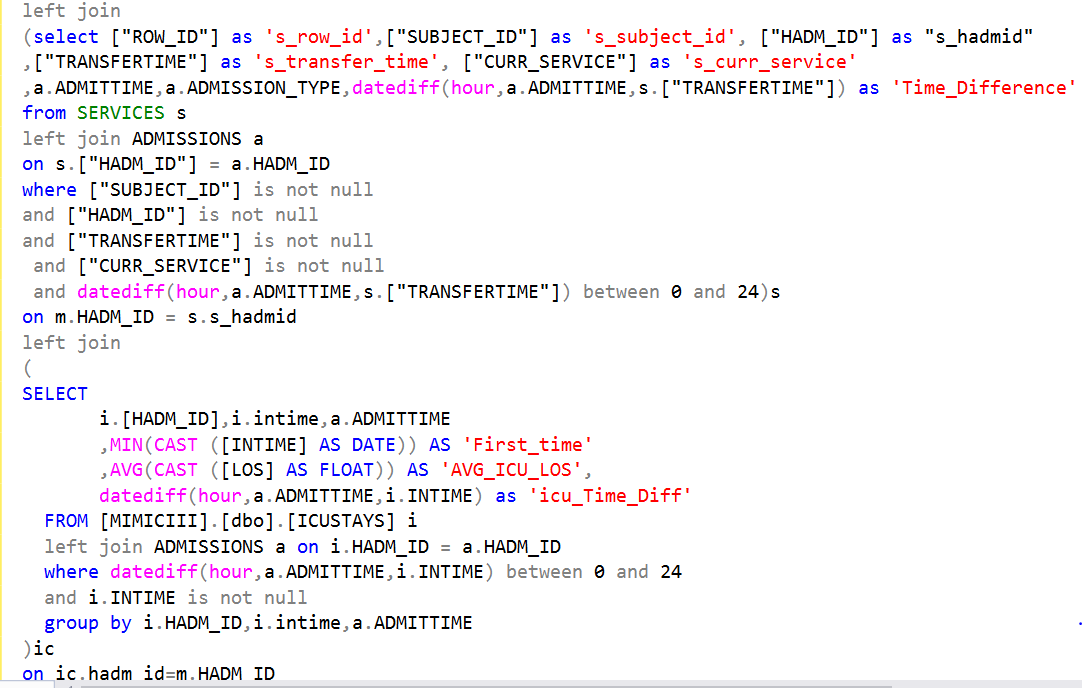
 

Fig. 4.3 Joins

After the pre-processing, the final dataset for the models has a lot of 87 attributes. One dependent variable(length of stay) which is the target variable and 86 input variables.

**5. Feature Selection**

Feature selection is the process of selecting a subset of features/variables for the use in models building. Feature selection can sometimes improve the performance of the models and it also makes the model less complex by reducing the number of features considered.

Two kinds of attribute selection/evaluation was conducted.

1. **CfsSubsetEval with BestFirst**: This evaluates the worth of a subset of features by evaluating the individual predictive ability of each feature. This feature is used with forward direction.

This helps us in evaluating the features and ranking them to see what top features contribute to predicting the length of stay in hospital.

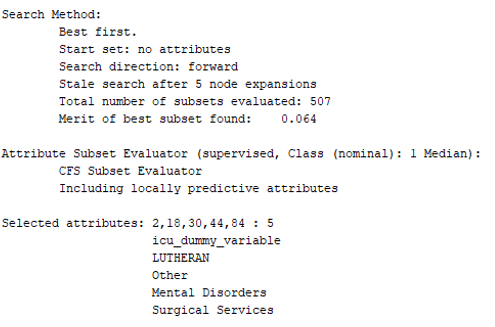


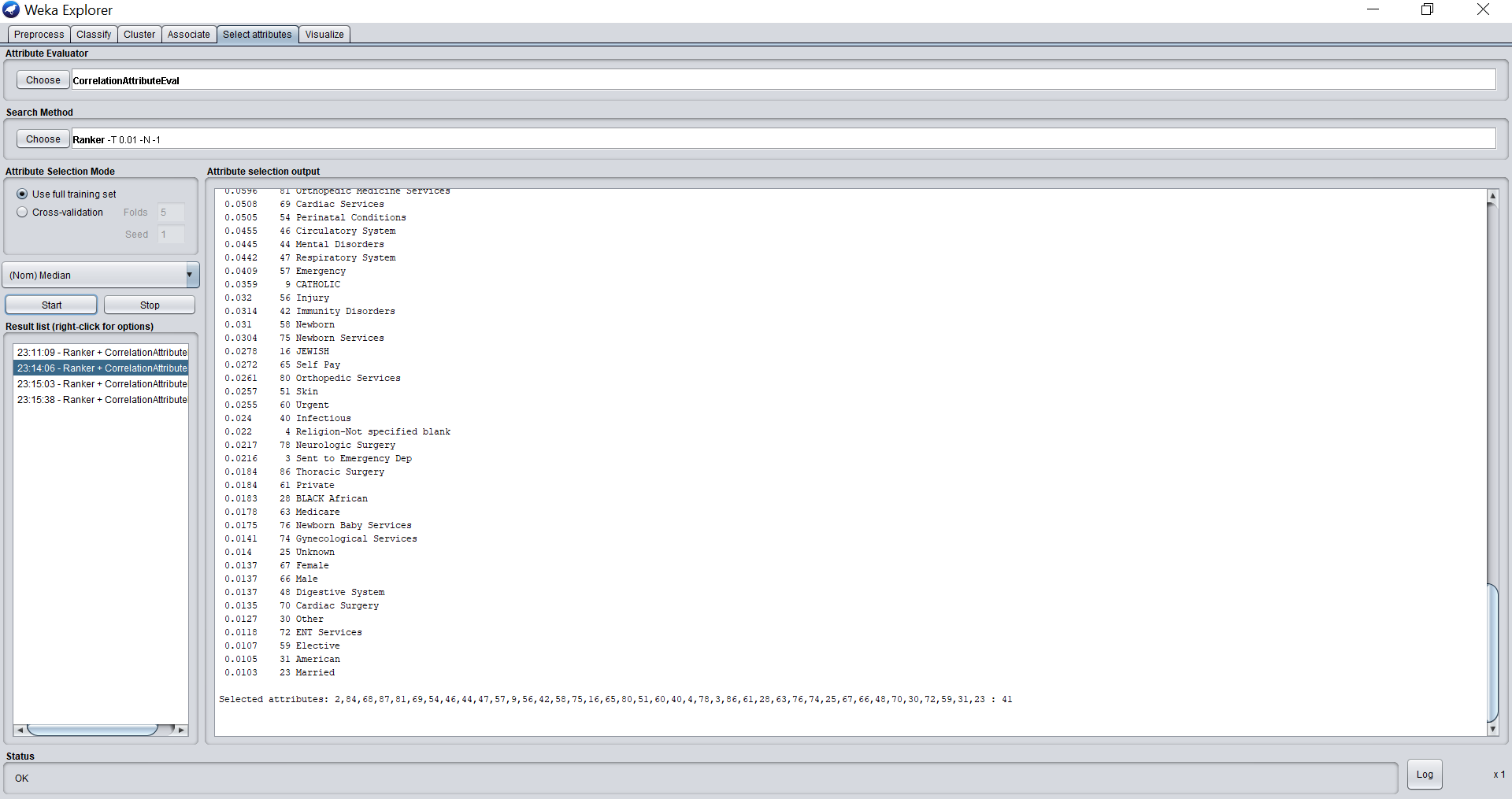
Fig 5.1 Features

Figure 5.1 gives the 5 features that are icu\_dummy\_variable which says if a patient was sent to the ICU, mental disorders and surgical category, patients with ethnicity as others and religion Lutheran.

The feature with religion category LUTHERAN is a surprising entry since most of the other ethnicity category features were removed from the list of features when CorrelationAttributeEval feature evaluator was executed.

1. **CorrelationAttributeEval feature evaluator with Ranker**: Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class.

CorrelationAttributeEval feature evaluator is used in Weka with Ranker search method and a threshold value of 0.01. From a total of 87 features, this method selects 41 features.



**Fig.5.2 CorrelationAttributeEval**

**6. Models**

Three models are selected to predict the length of stay – Bayes Network, Random Forest, Logistic Regression and Naïve Bayes. The models are built in Weka3.8 and data is split in the ratio of 70-30 for training and testing.

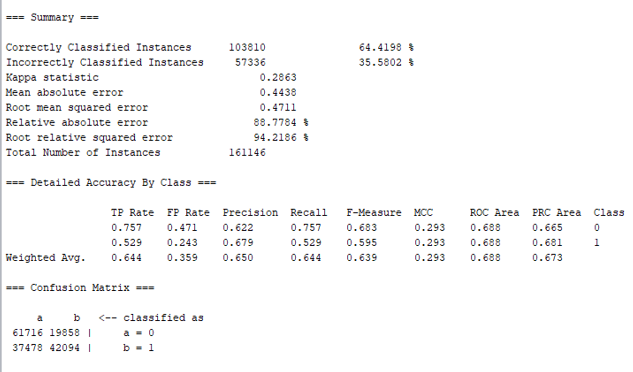
**Evaluation Metrix:** We use ROC, Precision, Recall and accuracy to evaluate the model performances.

**6.1 Logistic Regression**

Logistic Regression model was run on Weka with different methods. The feature selection method used was CorrelationAttributeEval feature evaluator and Ranker search method.

With all features and 70-30 split, we get the best accuracy of ~65%.

|  |  |
| --- | --- |
| Method | Accuracy(%) |
| 70-30 data split (All features) | **64.42** |
| 80-20 data split (All features) | 64.1 |
| Feature Selection (reduced features) | 64.32 |



6.1 Logistic Regression Best Results

**6.2 Naïve Bayes**

Naïve Bayes model was run on Weka with different methods. The feature selection method used was CorrelationAttributeEval feature evaluator and Ranker search method.

With all features selected and 80-20 split, we get the best accuracy of ~63.21%.

|  |  |
| --- | --- |
| Method | Accuracy(%) |
| 70-30 data split (All features) | 63.17 |
| 80-20 data split (All features) | **63.21** |
| Feature Selection (reduced features) | 63.2 |
| 10 fold cross validation | 63.04 |

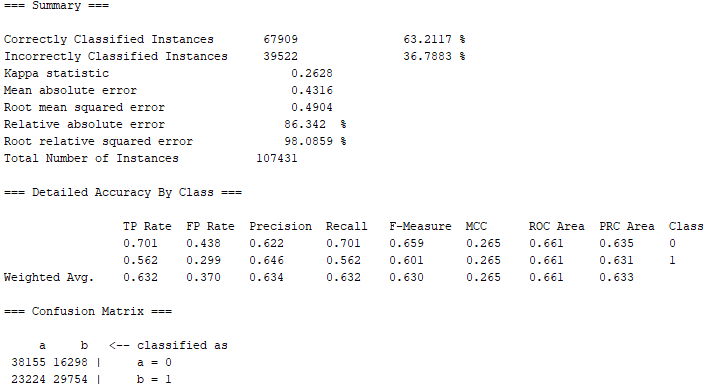


Fig 6.2 Naïve Bayes best results

**6.3 Bayes Network**

Bayes Network model was run on Weka with different methods. The feature selection method used was CorrelationAttributeEval feature evaluator and Ranker search method.

With all features and 80-20 split, we get the best accuracy of ~63.7%.

|  |  |
| --- | --- |
| Method | Accuracy(%) |
| 70-30 data split (All features) | 63.66 |
| 80-20 data split (All features) | **63.7** |
| 10 fold cross validation | 63.57 |
| Feature Selection (reduced features) | 63.63 |

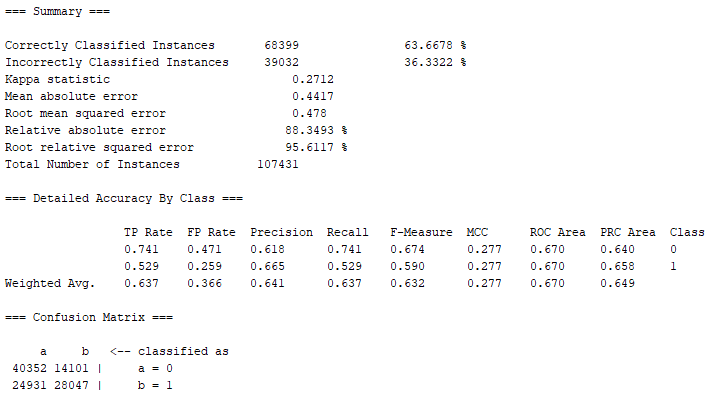


Fig 6.3 Bayes Network best results

**6.4 Random Forest**

Random Forest model was run on Weka with different methods. The feature selection method used was CorrelationAttributeEval feature evaluator and Ranker search method.

With all features and 70-30 split, we get the best accuracy of ~69.9%.

|  |  |
| --- | --- |
| Method | Accuracy(%) |
| 70-30 data split (All features) | **69.72** |
| 80-20 data split (All features) | 69.33 |
| Feature Selection (reduced features) | 68.19 |

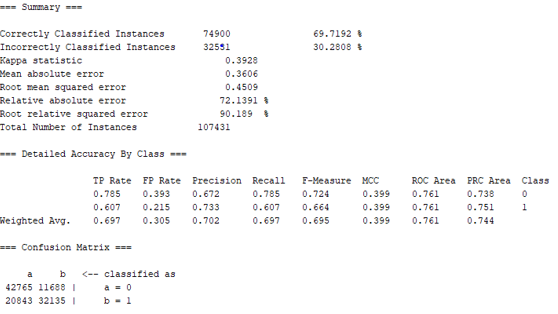


Fig 6.4 Random Forest best results

**7. Results**

4 predictive models are built to predict the length of stay of patients at the hospital. The ROC, Precision, Recall and Accuracy of all the 4 models and under all considered parameters and methods are evaluated. In general, all models perform better with all attributes considered and with a split of data in 70-30 for training and testing.

Following table represents all the evaluation metrics for all the four models and their best performances.

We can see that the Random Forest performs the best followed by Logistic Regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | ROC | Precision | Recall | Accuracy(%) |
| Logistic Regression | 0.7 | **0.66** | 0.64 | 64.42 |
| Naïve Bayes | 0.661 | 0.64 | 0.63 | 63.21 |
| Random Forest | **0.77** | 0.702 | 0.69 | **69.72** |
| Bayes Network | 0.670 | 0.641 | 0.632 | 63.67 |

Fig.7.1 Results

**8. Conclusion**

Predicting the length of stay can be beneficial in many hospital systems. The proposed prediction model can predict the length of stay of a patient in the hospital within 24 hours of their admission. Out of the 4 models in the paper, the best performing model is random Forest with an accuracy approximately 70%.

Models like this, can be used in hospitals to manage resources and improve hospital scheduling.

**9. Future Work**

MIMIC-III dataset is very vast and informative with 26 different tables having information on lab reports, stay, staff, procedures, lab events and even notes. This project made use of 6 of these tables.

Including lab events and information from lab reports about vitals could be beneficial to predict the length of stay of a patient. The transfers table could also provide useful information about the stays in the hospital which can contribute to predicting the length of stay in a hospital.

Also, there are numerous other predictive models that give good results. Comparing results to more predictive models and using different tuning parameters for the current models could also improve the performance of the prediction.

**References**

[1] Daniel Cummings (Dec 15,2018), *Predicting hospital length-of-stay at time of admission*, Retrieved from: <https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission-55dfdfe69598>

[2] Joseph Miles, PharmD (Aug 9, 2017), *Reconstruction of the MIMIC-III Database for data analytics,* Retrieved from: <http://pi.cs.oswego.edu/~jmiles3/mimic/Miles-MIMIC-Project_report.pdf>

[3]MIMIC*, MIMIC-III Critical Care Database*, Retrieved from: <https://mimic.physionet.org/about/mimic/>

[4] Thanos Gentimis, Ala’ J. Alnaser, Alex Durante, Kyle Cook, Robert Steele (Apr 02, 2018), *Predicting hospital length of stay Using Neural Network on MIMIC III Data,* IEEE, Retrieved from: <https://ieeexplore.ieee.org/document/8328535>

[5] Shirly Wang, Geeticka Chauhan, Tristan Naumann (2019), *MIMIC- Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC -III*, Retrieved from: <https://www.arxiv-vanity.com/papers/1907.08322/>