Predicting Post Procedural Complications using MIMIC-III

****

Master of Science in Computer Science

Faisal Maqbool

MSDS17027

Session: 2017 – 2019

DEPARTMENT OF COMPUTER SCIENCE

INFORMATION TECHNOLOGY UNIVERSITY

LAHORE, PAKISTAN

****

Predicting Post Procedural Complications Using MIMIC-III

A thesis submitted in partial fulfillment of the requirements for the

Degree of Master of Science in

Computer Science

Faisal Maqbool

Dr. Saeed Ul Hassan

Committee Member Name

Committee Member Name

**Declaration**

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions. I also declare that this work is the result of my own investigations, except where identified by references and free from plagiarism of the work of others.

Signature: ………….

Student Name

Date: ………………..

The undersigned hereby certify that they have read and recommend the thesis entitled “…………………………..” by ………………….. For the degree of Master of Science in Data Science.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor Name (ITU), Thesis Advisor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Committee Member Name (ITU), Thesis Committee Member

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Committee Member Name, Thesis Committee Member

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Chairperson Name, Chairperson of the Department

**Acknowledgment**

First and foremost, I would like to express sincere gratitude to my advisor Dr. Saeed Ul Hassan for being the supportive advisor I could have asked for. His precious insights, guidance and support throughout the research journey has not only helped me become a better student but also examine data for research and contribute to previously done research. Though he is very busy, he always made me feel like a priority. He took time to review my findings, answer questions, and give thoughtful feedback for which I am extremely grateful. This thesis would not have been possible without them.

I would like to thank Dr. Mohsin Ali for his help, advice and support throughout my work and for agreeing to serve on my thesis committee.

Moreover, I would like to thank my family especially my parents and my friends for me letting my follow my dreams and for their encouragement which has been the vital part of the research.

Table of Contents

Chapter: 1 11

[1.1 Introduction 11](#_Toc28221795)

[1.2 Literature Survey 12](#_Toc28221796)

[1.3 ICD, Hippa AND Complications 13](#_Toc28221797)

[1.4 Why COMPLICATIONS? 14](#_Toc28221798)

[1.5 APPROACH Followed 14](#_Toc28221799)

[1.6 Description of the Content 15](#_Toc28221800)

[Chapter: 2 MIMIC-III Critical Care Database 17](#_Toc28221801)

[2.1 About Mimic-III 17](#_Toc28221802)

[2.2 MIMIc-iii Tables 19](#_Toc28221803)

[2.3 MIMIC-III Construction 22](#_Toc28221804)

[2.4 MIMIC-III derived Concepts 22](#_Toc28221805)

[Chapter: 3 ETL and DataSet Building From DWH Mart 24](#_Toc28221806)

[3.1 Technical Process 24](#_Toc28221807)

[3.2 Relational Mapping, Batch Processing and Preprocessing 25](#_Toc28221808)

[3.3 Diagnoses 26](#_Toc28221809)

[3.4 Procedures 27](#_Toc28221810)

[3.5 Feature Space 27](#_Toc28221811)

[3.6 Features Engineering Treatment 28](#_Toc28221812)

[3.7 Sampling 29](#_Toc28221813)

[3.7.1 SMOTE: Synthetic Minority Over-Sampling Technique 29](#_Toc28221814)

[3.7.2 ADASYN: Adaptive Synthetic Sampling 30](#_Toc28221815)

[Chapter 4: Models and Experiments 32](#_Toc28221816)

[4.1 About Binary Classification 32](#_Toc28221817)

[4.2 Selected Models 32](#_Toc28221818)

[4.2.1 Logistic Regression 33](#_Toc28221819)

[4.2.2 Random Forrest 33](#_Toc28221820)

[4.2.2 Linear SVC 34](#_Toc28221821)

[4.2.3 Artificial Neural Network (ANN) 35](#_Toc28221822)

[References 37](#_Toc28221823)

**List of Tables**

[Table 1: Class Distribution of data for MIMIC-III Dataset 17](#_Toc28228809)

[Table 2: MIMIC-III Tables Summary 18](#_Toc28228810)

[Table 3 : Derived Concepts 22](#_Toc28228811)

[Table 4: Diagnoses and Procedures Count 26](#_Toc28228812)

[Table 5: Class Count 28](#_Toc28228813)

[Table 6: Results with ADASYN Data Sampling 38](#_Toc28228814)

[Table 7: Results with SMOTE Data Sampling 38](#_Toc28228815)

**List of Figures**

[Figure 1: Methodology 15](file:///D:\DataScience\Thesis-Bioinformatics-MIMICIII-\Initial%20Draft.docx#_Toc28221940)

[Figure 2: MIMIC-III Construction Model 22](file:///D:\DataScience\Thesis-Bioinformatics-MIMICIII-\Initial%20Draft.docx#_Toc28221941)

[Figure 3: Technical Chain of Steps 24](file:///D:\DataScience\Thesis-Bioinformatics-MIMICIII-\Initial%20Draft.docx#_Toc28221942)

[Figure 4: Batch Processing for Huge Files Using Python 26](#_Toc28221943)

[Figure 5 : One Hot Encoding for Categorical Features 28](#_Toc28221944)

[Figure 6: SMOTE (Sampling) 30](#_Toc28221945)

[Figure 7: Sigmoid function 33](#_Toc28221946)

[Figure 8: Logistic Vs Linear SVM 34](#_Toc28221947)

[Figure 9: Artificial Neural Network (Perceptron Basic Model) 36](#_Toc28221948)

**List of Abbreviations**

MIMIC:

PHI:

ICU:

ETL:

ICD:

HIPPA

BMI:

i2b2:

ACDF:

SMOTE

ADASYN:

**Abstract**

The advancements in bioinformatics and health care sector has inspired researchers to develop systems that can mimic work of doctors because quantifying patient health and predicting outcomes is an important problem in critical care research. But, predicting outcomes for critically ill patients admitted in intensive care units requires specific characteristics of clinical data: worth, capacity, access and dimensionality. In this Master Thesis, an analysis of the data from critical patients was carried out in order to study the influence of several factors for predicting the post-procedural complications and seeking if those complications can lead to mortality of patients. To derive insights for that, we used well-known clinical dataset named Medical Information Mart for Intensive Care III (MIMIC-III) with two types of data sampling techniques: SMOTE and ADASYN. For both techniques our results showed that Random Forrest outperforms other linear models with an accuracy of **>80%** with AUROC of **0.83.**

**Chapter: 1**

# Introduction

With the initiation of digital technology, advanced techniques are increasingly making it possible to utilize big data to more precisely risk evaluation and predict how an individual patient will behave based on a given diagnose or procedure. Intensive care unit **(ICU)** is a ward in hospital, where critically ill patients are admitted requires accurate predictors that can help doctors with the assessment of severity of illness.

In this thesis we investigate the different methodologies for extracting, transforming and loading **(ETL)** processes, which obtain data from original source to perform informative analysis and features extraction to aid model to predict post-procedural (diagnoses & procedure) complications of critically ill patients and investigating those complications if those can lead to mortality of patient or not. The methods use demographics, data from different hospital system, lab events, diagnoses, notes and other engineered information regarding each patient. The database used for the study is Medical Information Mart for Intensive Care **MIMIC-III [1]** which comes from health service with anonymized data for protecting health information **(PHI).**

Other researches about MIMIC-III data is also presented to motivate our problem, establish understanding of dataset, key findings and recommendations for future investigations. The question of predicting post-procedural complications from data science perspective and critical health perspective is not only important for doctors, administrators but also for the patient as well. For administrators this would help managing patients and required resources. Avoiding predicted complications can further be avoided if such information is known during the stay of patient at ICU.

On MIMIC-III researchers widely contributed to support the cause. **[9]** Presented an evaluation of the influence of body mass index. They hypothesize that selected severity of illness scores would perform differently if body mass index categorization was incorporated and that the performance of these score models would improve after consideration of body mass index as an additional model feature. **[10]** Presented a study recently where they characterized the relationship of Body Mass Index **(BMI)** with survival and explored gender-based interactions with surrogates of body composition of nutrition in a real world setting. Another cohort study **[11]** examine the impact of overstays of patients and discharge delays on in-hospital morbidity and mortality.

Many other researcher have contributed in prediction of other critical factors related to medicine and the health of patient. **[12]** Examined that severity scores may also lead to misclassification of critically ill obese patients. They further compared the laboratory results between obese and normal patients.

# Literature Survey

Health and medicine are one of the key sectors that requires use of new technologies to produce new possibilities and cause a greater impact in the society. Some of the recent research studies mentioned above have been conducted on **MIMIC-III** dataset. Numerous researches have also been conducted on the subject matter and overall on the usage of MIMIC-III for creating new possibilities of research and scientific areas.

Following are some of the researches that helped us motivate our problem and contribute by applying information retrieval and data science techniques.

As predicting the length of stay for the patient can provide valuable information to the management of the hospital but also for patient’s health. **[2]** Explored the use of neural networks for predicting the length of stay of patient within a time range of **(<5)** days or **(>5)** days after the patient left the Intensive care unit. They used a subset of **MIMIC-III** and written all their models in **R** and **PostgreSQL** using a supercomputer provided by the Florida Polytechnic University. Their model predictions achieved **80%** accuracy and outperformed any other linear models previously used of predicting the length of stay.

Another study in **2019** **[3]** mapped ICD-9 codes using the clinical notes from physicians, doctors and other staff against each patient automatically using deep learning. They applied multiple experiments and showed that deep learning outperforms linear models to predict top 10 ICD-9 codes with **0.6957 F1** score and accuracy of **0.8967**.

Research published in [BMC Medical Informatics and Decision Making](https://bmcmedinformdecismak.biomedcentral.com/) **[4]** proposed a new approach by combining rule-based features and knowledge guided deep learning model for effective disease classification. They evaluated their method on **i2b2** obesity challenge and demonstrated that their model outperform other method used for disease classification. Another mortality prediction case study published in Machine Learning for HealthCare Conference **[5]** demonstrated large heterogeneity in studies that product the single task of mortality prediction.

Apart from above mentioned researches a lot of other researchers contributed in predicting the complications. A study **[6]** demonstrated the performance of machine learning models in predicting postoperative complications following anterior cervical discectomy and fusion **(ACDF)**. They applied Logistic Regression, Random Forrest and ANN to achieve their results. **20,879** patients and **ANN** outperformed **LR** in predicting venous thromboembolism, wound complication, and mortality **(p < 0.05).** The SVM and RF models were no better than random chance at predicting any of the postoperative complications **(p < 0.05).** Similarly, **[7]** used machine learning to derive and validate the hospital readmission. Their machine-learning–derived model had significantly better performance (area under the receiver operating curve, **0.76**) than either the Stability or Workload Index for Transfer score (area under the receiver operating curve, 0.65. **[8]** A retrospective study, on real-time prediction of complications in critical care used **Recurrent Neural Networks** to predict complications of mortality, renal failure, and postoperative bleeding and operation revision.  47 559 intensive care admissions (corresponding to 42 007 patients), they included 11 492 (corresponding to 9269 patients). Their deep learning models yielded accurate predictions with the following **PPV** and sensitivity scores: **PPV 0·90** and **sensitivity 0·85** for mortality, **0·87** and **0·94** for renal failure, and **0·84** and **0·74** for bleeding.

Another study using MIMIC-III **[13]** implemented a deep rule based fuzzy systems for predicting the accurate in-hospital mortality. Their main contribution to the system was proposing a system which can handle multiple categorical data.

# ICD, Hippa AND Complications

The International Classification of Diseases **(ICD) [13]** is the foundation for the identification of health trends and statistics globally, and the international standard for reporting diseases and health conditions. It is the diagnostic classification standard for all clinical and research purposes. Under revision of ICD9 codes, the code 996 defines complications particular to certain specified procedures and diagnoses. Most complications are caused due to cardiac, vascular or other used devices and some of them relates to reaction caused due to a procedure performed. In our work, we are focusing on such complications and investigating if those can lead to the mortality of patient. **HIPAA** (Health Insurance Portability and Accountability Act of 1996) is United States legislation that provides data privacy and security provisions for safeguarding medical information. To protect health information MIMIC-III provided anonymized data, still we need to make sure we are following HIPPA compliance rules so that our research does not conflict with any of the standards defined.

# Why COMPLICATIONS?

Extended diagnoses and patients stay at the hospital is associated with not only the health of patient, cost, increased number of deaths but also increased number of readmissions. Each of these parameters defines the hospital performance. So, our focus is to produce insights that can complement these parameters ranging from cost to patient health.

# APPROACH Followed

As open sourced MIMIC-III data is available for research purposes for free. Our main focus was to use tools and technologies in a way that can complement already done research. Hence, we used **Python**, **R** and **PostgreSQL** for ETL process, features selection and modeling our features to extract results which I will explain in respective chapters in details.

We have used the basic Data Science (mining) model as full cycle of development in our research.



Figure 1: Methodology

The model divides several phases which are executed in iteration to get to goal which we can see in **Figure 1.1.** The ﬁrst Business Understanding phase focuses on understanding the objectives of the project. In the Data Understanding phase, the tasks related to the data collection, the exploratory analysis and the quality review of the data are included. During the Data Preparation phase, the data that will be used for the successive phases are selected, the data is cleaned if necessary, derived data are obtained from the sources and in the modeling phase we design and implement our models with appropriate parameters to extract results. In the end final deployment stage is carried as per requirement.

# Description of the Content

The thesis is organized as follows: In chapter **1** we introduce the problem, explains purpose of our work, methodologies followed and detail of contents. In Chapter **2**, we define the data source MIMIC-III in detail. In Chapter **3**, we explained the ETL process to obtain data and process it effectively for posterior use. In Chapter **4**, we explain the preprocessing and analysis of data which is used of feature engineering and model selection process. In Chapter **5**, results, conclusion and future work is been discussed.

# Chapter: 2 MIMIC-III Critical Care Database

This chapter explains the structure, context and development researchers have done on data source MIMIC-III.

# 2.1 About Mimic-III

Over the past decade, much have been written about the field of data science regarding the explosion of big data. In health care, every decision made for a critical patient requires precision by clinicians and doctors. To carry out research to aid clinicians and doctors to make better, reliable and quick decision using research and applications. This demands privacy existence of data and wide-ranging analysis.

To avoid these difﬁculties, it is available to the researcher and educational community the **MIMIC-III (Medical Information Mart for Intensive Care III)** database, openly accessible at <https://mimic.physionet.org/> This database is an evolution of the MIMIC-II database created by the Laboratory of Computational Physiology of The Massachusetts Institute for Technology (MIT) with the goal of providing tools for the creation of clinical information with the help of different techniques of data sciences. MIMIC-III is a large, freely-available relational database comprising de-identiﬁed health related data associated with over forty thousand patients who stayed in Intensive Care Units at Beth Israel Deaconess Medical Center (Boston, Massachusetts). The data spans June 2001 October2012.

MIMIC-III is a comprehensive collection of de-identified data from **53,423** distinct critical care hospital admissions from **38,597** distinct adult patients. The data has been compiled into 26 tables which contain, for example, an average of **4579** charted observations and 380 laboratory measurements for each hospital admission as well as a total of **3.8** gigabytes of unstructured textual data from various healthcare provider notes and analyses. In addition to de-identifying patient data, MIT requires training in the protection of patient data for anyone requesting access to the MIMIC dataset. After completing the prescribed training, data can be downloaded as 26 comma separated values (csv) files representing the **26** tables in the MIMIC-III database. Sample SQL code can be acquired from GitHub (<https://github.com/MIT-LCP/mimic-code>) for establishing relationships between the tables. Additionally, there is a published data dictionary which can be found at <https://mimic.physionet.org/mimictables/admissions/>. There is variability in the usage of “unique” attributes and definition of primary keys between the sample SQL code and the published data dictionary. For example, every table has an attribute called “ROW\_ID”, and the sample SQL code consistently declares this attribute as “unique” and/or as a “primary key” for every table despite the fact that tables like the “PATIENTS” table have a unique identifier (SUBJECT\_ID) that is intended to be the primary key and serve as foreign key in child relations that refer to the “PATIENTS” table.

After downloading and analyzing the MIMIC source tables, implementation occurs in 5 additional steps:

* Create tables with attribute rules (data types) and identify the primary key for each table. 1. Load records from csv files into each table.
* Declare the indexes for each table.
* Define foreign keys in each table and establish table relationships.
* Implement user interface (with appropriately granted permissions) for the database.

This database includes information on demographic data of patients, laboratory test results, vital sign measurements, procedures, medications, caregiver notes, imaging reports, mortality (both in and out of the hospital), manual evolution annotations regarding events, discharge reports, prescription, and so on. The class distribution of data shown in table **1** (adopted from <https://www.nature.com/articles/sdata201635/tables/3>)

Table 1: Class Distribution of data for MIMIC-III Dataset

| **Class of data** | **Description** |
| --- | --- |
| Billing | Coded data recorded primarily for billing and administrative purposes. Includes Current Procedural Terminology (CPT) codes, Diagnosis-Related Group (DRG) codes, and International Classification of Diseases (ICD) codes. |
| Descriptive | Demographic detail, admission and discharge times, and dates of death. |
| Dictionary | Look-up tables for cross referencing concept identifiers (for example, International Classification of Diseases (ICD) codes) with associated labels. |
| Interventions | Procedures such as dialysis, imaging studies, and placement of lines. |
| Laboratory | Blood chemistry, hematology, urine analysis, and microbiology test results. |
| Medications | Administration records of intravenous medications and medication orders. |
| Notes | Free text notes such as provider progress notes and hospital discharge summaries. |
| Physiologic | Nurse-verified vital signs, approximately hourly (e.g., heart rate, blood pressure, respiratory rate). |
| Reports | Free text reports of electrocardiogram and imaging studies. |

# 2.2 MIMIc-iii Tables

MIMIC-III is structured in a relational manner containing 26 files from which we created following tables in PostgreSQL.

Table 2: MIMIC-III Tables Summary (Adapted from: <https://mimic.physionet.org/gettingstarted/access/>)

|  |  |  |
| --- | --- | --- |
| File Name | Dimension | Summary |
| ADMISSIONS | (58976, 19) | The ADMISSIONS table gives information regarding a patient’s admission to the hospital. |
| CALLOUT | (34499, 24) | The CALLOUT table provides information about ICU discharge planning |
| CAREGIVERS | (7567, 4) | This table provides information regarding care givers. For example, it would deﬁne if a caregiver is a research nurse (RN), medical doctor (MD), and so on. |
| CHARTEVENTS | (330712483, 15) | CHARTEVENTS contains all the charted data available for a patient. |
| CPTEVENTS | (573146, 12) | The CPTEVENTS table contains a list of which current procedural terminology codes were billed for which patients. This can be useful for determining if certain procedures have been performed (e.g. ventilation). |
| D\_CPT | (134, 9) | This table gives some high level information regarding current procedural terminology (CPT) codes. Unfortunately, detailed information for individual codes is unavailable. |
| D\_ICD\_DIAGNOSES | (14567, 4) | This table deﬁnes International Classiﬁcation of Diseases Version 9 (ICD-9) codes for diagnoses. These codes are assigned at the end of the patient’s stay and are used by the hospital to bill for care provided. |
| D\_ICD\_PROCEDURES | (3882, 4) | This table deﬁnes International Classiﬁcation of Diseases Version 9 (ICD-9) codes for procedures. These codes are assigned at the end of the patient’s stay and are used by the hospital to bill for care provided. |
| D\_ITEMS | (12487, 10) | The D\_ITEMS table deﬁnes ITEMID, which represents measurements in the database. |
| D\_LABITEMS | (753, 6) | D\_LABITEMS contains deﬁnitions for all ITEMID associated with lab measurements in the MIMIC database. |
| DATETIMEEVENTS | (4485937, 14) | DATETIMEEVENTS contains all date measurements about a patient in the ICU. |
| DIAGNOSES\_ICD | (651047, 5) | This table deﬁnes ICD-9 codes for diagnoses. The ICD codes are generated for billing purposes at the end of the hospital stay. |
| DRGCODES | (125557, 8) | This table deﬁnes HCFA-DRG and APR-DRG codes which provide information regarding Diagnosis-Related Group recorded primarily for billing and administrative purposes. |
| ICUSTAYS | (61532, 12) | This table gives information regarding ICU hospital stays. |
| INPUTEVENTS\_CV | (17527935, 22) | This table contains data of ﬂuid input events (serums, intravenous medication, insulin, etc.) regarding Carevue database source associated to ICU episodes. |
| INPUTEVENTS\_MV | (3618991, 31) | This table contains input data for patients. |
| LABEVENTS | (27854055, 9) | Contains all laboratory measurements for a given patient, including outpatient data. |
| MICROBIOLOGYEVENTS | (631726, 16) | Contains microbiology information, including tests performed and sensitivities. |
| NOTEEVENTS | (2083180, 9) | This table contains all notes for patients took in a manual way by their caregivers. |
| OUTPUTEVENTS | (4349218, 13) | This table contains output data for patients. |
| PATIENTS | (46520, 8) | This table contains hospitalization-independent data for all patients such as, gender, date of birth, etc. |
| PRESCRIPTIONS | (4156450, 19) | This table contains medication related order entries, i.e. prescriptions. |
| PROCEDUREEVENTS\_MV | (258066, 25) | This table contains procedures for patients |
| PROCEDURES\_ICD | (17527935, 22) | Contains ICD procedures for patients, most notably ICD-9 procedures. The ICD codes are generated for billing purposes at the end of the hospital stay and are recorded for all patient hospitalizations. |
| SERVICES | (73343, 6) | The SERVICES table describes the service that a patient was admitted under. This service admission can be elective or caused due to a number of reasons, including bed shortage. |
| TRANSFERS | (261897, 13) | This table contains physical locations for patients throughout their hospital stay. |

# 2.3 MIMIC-III Construction

MIMIC-III was constructed based upon hospital level, patient level, ICU level & used systems level. Furthermore it includes billing, notes and reports as shown in below figure.



Figure 2: MIMIC-III Construction Model

# 2.4 MIMIC-III derived Concepts

The active researchers have contributed to already given data with additional scripts to generate new concepts and insights at MIMIC code repository **[4]** which includes views and tables as well. They also encourage other researchers to contribute to derived insights which helps to distinct between the original data and derived data and one can use as per the problem they are solving and contribute as well.

Following are the major concepts that are being used frequently by researchers.

Table 3 : Derived Concepts

|  |  |
| --- | --- |
| Class of Data | Summary |
| Comorbidity | These scripts derive binary ﬂags indicating the presence of various comorbidities using billing codes (ICD-9) assigned to the patient at hospital discharge. |
| First day | The ﬁrst day subfolder contains scripts used to calculate various clinical concepts on the ﬁrst day of a patient’s admission to the ICU, such as the highest blood pressure, lowest temperature, etc. This folder contains many useful scripts which can be adapted to capture data outside the ﬁrst day. |
| Sepsis | Deﬁnitions of sepsis, a common cause of mortality for intensive care unit patients. |
| Severity Scores | Severity of illness scores which summarize the acuity of a patient’s illness on admission to the intensive care unit (usually in the ﬁrst 24 hours). |
| Durations | Start and stop times for administration of various treatments or durations of various phenomena, including: medical agents which have a vasoactive effect on a patient’s circulatory system, continuous renal replacement therapy (CRRT), and mechanical ventilation. |
| Organ Failure | This script derives binary flags for major organ failures |

[[1]](#footnote-1)The tables are linked by identifiers which usually have the suffix “**ID**”. For example **HADM\_ID** refers to a unique hospital admission and **SUBJECT\_ID** refers to a unique patient. One exception is **ROW\_ID**, which is simply a row identifier unique to that table.

Tables pre-fixed with “**D\_**” are dictionaries and provide definitions for identifiers. For example, every row of **OUTPUTEVENTS** is associated with a single **ITEMID** which represents the concept measured, but it does not contain the actual name of the drug. By joining **OUTPUTEVENTS** and **D\_ITEMS** on **ITEMID**, it is possible to identify what concept a given **ITEMID** represents

# Chapter: 3 ETL and DataSet Building From DWH Mart

In this section, we would introduce the ETL followed by us to derive certain insights which will lead us to conclusion of stated problem. The section is divided into extraction, transformation and loading sections to reach to our features.

# 3.1 Technical Process



Figure : Technical Chain of Steps

Given the size of data mart and the volume of raw data, we devoted most of time to extraction and transformation of data.

In the first step prior to requesting access to **MIMIC**, you will need to complete the **CITI** “Data or Specimens Only Research” course by registering yourself on CITI program. After getting data access we are provided links to the **26** comma separated file containing patient, hospital and ICU related data[[2]](#footnote-2).

Following are key steps covered in technical process to engineer features:

* Tables Creation
* Relationship Mapping (Indexes and Keys)
* Materialized views from already given tables
* Trim down values for ICD-9 Codes
* Filter rows with subject id lookup and pass it to items lookup for certain diagnoses and procedure
* ICD-9 Codes for class complications which is 996
* Making it to binary classes with 1 and all other classes to 0
* Extracting derived features from chart events and lab events with batch processing
* Consolidate all other features with derived concept
* Format all features, fill out invalid fields and normalize features for model training

Code available at: <https://github.com/faisalmaqbool94/Thesis-Bioinformatics-MIMICIII->

# 3.2 Relational Mapping, Batch Processing and Preprocessing

All 26 files are relationally mapped with each other[[3]](#footnote-3). After getting these files we created a database of all those file and created respective tables. To improve the performance indexes and constraints were added. There were several of them that are huge and others are medium to tiny. The small files were dealt with PSQL but on the other hand, the big files caused a problem for not only creating table but also of processing those files in RAM. To handle such problems with huge files, we implemented Python script for asyn batch processing using Pandas[[4]](#footnote-4) which is an open sourced library to manipulate structured data and very highly efficient because of its reliable data frame objects along with transformation tools available with it.

As the data sources and research work is now publicly available. Researchers have contributed in the form of code, new concepts, and optimization of previously written script and in many other ways. Similarly we have used and created features set containing top diagnoses and procedures performed on ICU patients.

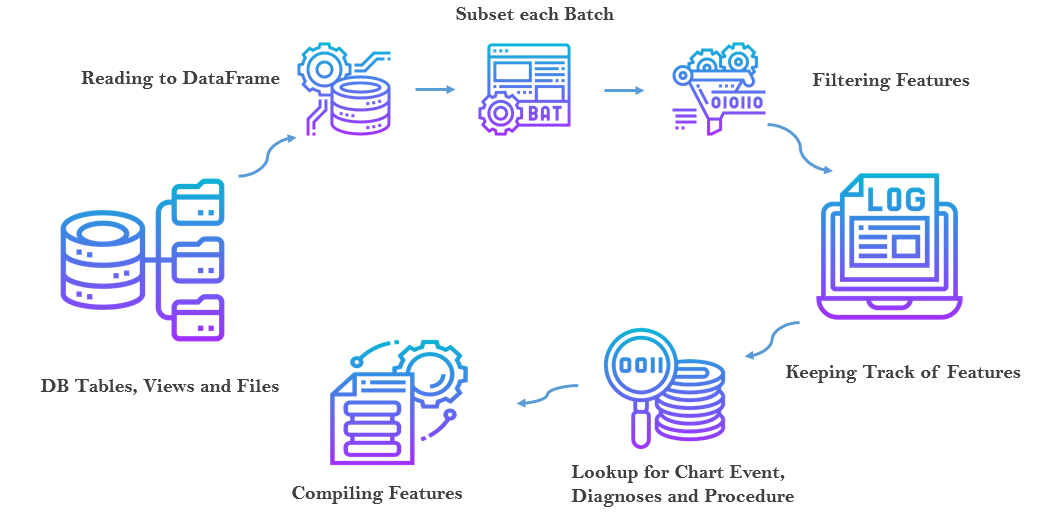


Figure 4: Batch Processing for Huge Files Using Python

To complete ETL process, PostgreSQL and Python played important role. Multiple SQL scripts were written for creation of tables, indexes, materialized views and derived tables. All of which are presented on a public repository.

<https://github.com/faisalmaqbool94/Thesis-Bioinformatics-MIMICIII->

Extraction of major chart events and lab events against each patient involved filtering of specific patients, lookup against particular diagnoses and procedures.

# 3.3 Diagnoses

From **14328** unique diagnoses that were available in MIMIC-III, we selected those diagnoses which were more common with a threshold of happening more than **30.** The number is randomly selected and experimented with. We can change number with other experiments to see if the models can outperform earlier built models. The reason behind choosing this particular number was that using these number of groups, if we use any aggregated function to fill out missing information, there will be a low probability of creating features containing near to zero variance.

# 3.4 Procedures

From **3882** unique procedures registered in MIMIC-III, we selected more common ones. Same as we did for diagnoses. For procedures there were only few fields that required any preprocessing.

In the **Figure 4** the subset creations and filtering involved lookups where we created separate files for segregating the subjects which are only appearing in above criteria of procedures and diagnoses.

|  |  |
| --- | --- |
| Item | Counts |
| Diagnoses | 14328 |
| Procedure | 3882 |

Table : Diagnoses and Procedures Count

# 3.5 Feature Space

Now that we have explored and discussed about the dataset. Now we will discuss the variables of interest. Once these features are identified we have to define the processes of our models and reach our goals.

From the original bunch of potential variables candidates that it can be included, physicians from **PIMS** hospital Islamabad and **Islamabad Diagnostic Centre**, selected and helped us engineer our features they know are important from their medical knowledge, experience or intuition. Following are our selected variables:

1. **General**: Insurance, Martial status, Hospital Expire Flag, Length of Stay, Calculated Bicarbonate, TotalCo2, Chloride, Free Calcium, Glucose, Hematocrit, Hemoglobin, Lactate, Oxygen, Oxygen Saturation, PCO2, PH, Potassium, Sodium, Temperature, Calcium Total, Centromere, Creatinine, Globulin, Blood Glucose, Blood Lipase, Blood Magnesium, Blood Potassium, Blood Sodium, Platelets Counts, Red Blood Cells, White Blood Cells, Lymphocytes
2. **Engineered Concepts (Derived from Table 3)**: Congestive Heart Failure, Cardiac Arrhythmias, Valvular Disease, Pulmonary Circulation, Peripheral Vascular, Hypertension, Paralysis, Other Neurological, Chronic Pulmonary, Diabetes Uncomplicated, Diabetes Complicated, Hypothyroidism, Renal Failure, Liver Disease, Peptic Ulcer, Aids, Lymphoma, Metastatic Cancer, Solid Tumor, Rheumatoid Arthritis, Coagulopathy, Obesity, Weight Loss, Fluid Electrolyte, Blood Loss Anemia, Alcohol Abuse, Drug Abuse, Psychoses, Depression

# 3.6 Features Engineering Treatment

Once all the featured got extracted for certain subjects and against hospital admissions. We had to distinguish between categorical and numerical variables.

For categorical features One Hot Encoding technique was performed.



Figure 5 : One Hot Encoding for Categorical Features

Missing values got treated by the average of all the diagnoses and same goes for procedures. Average value is taken because we have extracted diagnoses and procedures which are commonly occurring and average of each group was taken.

Our target variable **‘Icd9\_Code’** got converted into binary variable and mapped to (0, 1) where **0** indicates the non-complication and **1** indicates the occurrence of complication. All the complications are further subdivided into thousands of categories but we were just interested in the main class of complication which is indicated by code **996.**

|  |  |
| --- | --- |
| ****Class**** | ****Count**** |
| ****Complication**** | **2754** |
| ****No Complication**** | **30491** |

Table : Class Count

# 3.7 Sampling

As we have considered a very sensitive topic which requires a lot of domain knowledge and predicting a complication requires precision. Although we collected and engineered our dataset for targeting our goals but as the **table 5** shows that we clearly have class imbalance problem. To tackle this problem we applied over sampling and down sampling techniques which are explained below.

## 3.7.1 SMOTE: Synthetic Minority Over-Sampling Technique

SMOTE **[14]** proposed by Chawla et al., is an oversampling method. This method interpolate the minority class neighbors to construct new minority class samples randomly. The method can be described as follows. Firstly, for each minority class sample  one gets it’s -nearest neighbors from other minority class samples. Secondly, one chooses one minority class sample among the neighbors. Finally, one generates the synthetic sample  by interpolating between  and  as follows:

Where (0, 1) refers to random number between 0 and 1. In view of geometry, SMOTE can be regarded as interpolating between two minority class samples. The decision space for the minority class is expanded that allows the classifier to have a higher prediction on unknown minority class samples.

The SMOTE algorithm is simple and effective while generating synthetic samples, and the overfitting problem is avoided. It expands the decision space for the minority class.

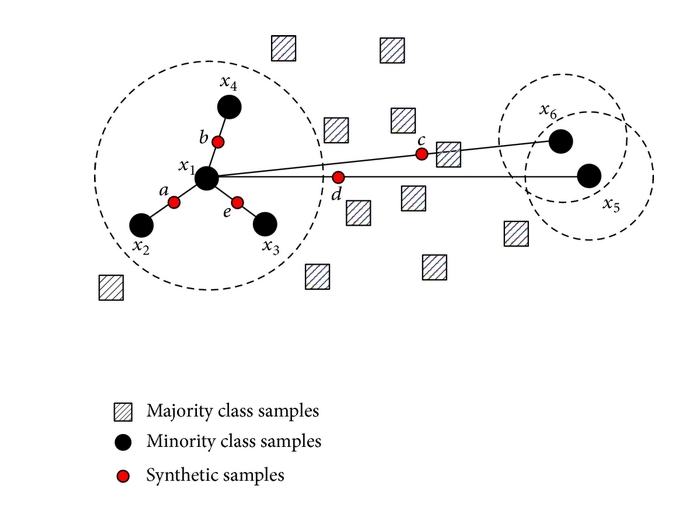


Figure : SMOTE (Sampling)

## 3.7.2 ADASYN: Adaptive Synthetic Sampling

ADASYN **[15]** is another oversampling method which interpolates new minority class samples by first calculating the number of synthetic samples of minority class and for each minority sample find the nearest neighbors by calculating the Euclidean distance.

**Input:**

D**ᵣ** with ***m*** samples with**,** ***= 1*** *to* ***m****,*where **xᵢ** is an n-dimensional vector in feature space and**yᵢ**is the corresponding class. Let  and  be the number of minority and majority class samples respectively, such that  and

**Algorithm:**

1. Calculate the Degree of Imbalance**,**
2. If (where is the preset threshold for maximum tolerated imbalance) then:
3. Calculate the number of synthetic samples to be generated from the minority class:**,**  is the balance level of the synthetic samples generated. β = 1 means there is a total balance between two classes.
4. For each minority samples, find the k-nearest neighbors based on Euclidean distance and calculate the ratio
5. Normalize**,**such that rₓ is now a density distribution.
6. Calculation of synthetic sample generated for each minority data point  
    where G is the total number of synthetic data examples that need to be generated for the minority class as defined in aforementioned Equation.
7. For each minority class data example **xᵢ**, generate **gᵢ** synthetic data examples according to the following steps:
8. Do the Loop from **1 to gᵢ**:

(a) Randomly choose one minority data example, **xᵤ**, from the K nearest neighbors for data **xᵢ**.  
(b) Generate the synthetic data example:   
where  is the difference vector in n-dimensional spaces, and λ is a random number:

The major difference between SMOTE and ADASYN is the difference in the generation of synthetic sample points for minority data points. In ADASYN, we consider a **density distribution rₓ** which thereby decides the number of synthetic samples to be generated for a particular point, whereas in SMOTE, there is a uniform weight for all minority points.

# Chapter 4: Models and Experiments

# 4.1 About Binary Classification

As we have defined our problem as binary class, either the patient would have complication or not, and defined a target feature which will be used to identify that. Depending upon the relative number of instances belonging to each class, we dealt with balanced and unbalanced labeled dataset. For unbalanced dataset we applied above mentioned two over-sampling techniques. To notion of metric performance is called accuracy which we used to validate our models defined in below section. Accuracy is defined as:

Here is the predicted class label for the Th iteration using (a defined function), is the number of ICU admission, the index represents the each single independent ICU admission. Is the indicator variable that equals one if classified correctly and 0 when classified incorrectly. Alternatively we can define accuracy metric as following:

Where TP and TN are positive and negative instances correctly classified by the model and FP are negative instances classified positive by model, similarly FN are positive instances classified negative by the model.

# 4.2 Selected Models

For our prediction task, we applied following models and are explained below:

* Logistic Regression (LR)
* Linear SVN (Linear Support Vector Classification)
* Random Forrest (Decision Tree)
* ANN (Artificial Neural Network)

To apply all these models, **Python** sklearn, imblearn, matlplotlib, Pandas and Numpy libraries were used.

## 4.2.1 Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression) using conditional probabilities mapped our features to a probability, model needs to output values in the range of and being continuous and differentiable.

The function with such properties is the **sigmoid** function, which we denote as follows:

The function is called sigmoid because of its S-shape, as illustrated in the figure below:



Figure : Sigmoid function

## 4.2.2 Random Forrest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

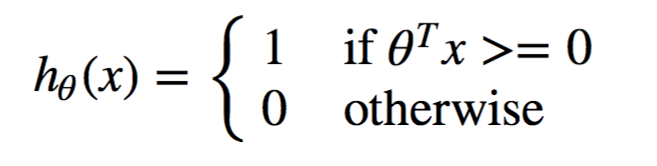
The [random forest](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm) is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a “forest”), this model uses [two key concepts](https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf) that gives it the name *random*:

1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes

## 4.2.2 Linear SVC

Linear SVC (Support Vector Classifier) (SVM) is used to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is using linear kernals.

It has the cost function like logistic regression defined below:



A comparison is given with the logistic cost function below:

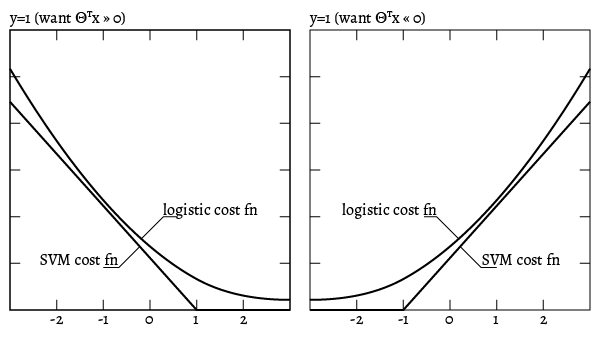


Figure : Logistic Vs Linear SVM

## 4.2.3 Artificial Neural Network (ANN)

Artificial Neural Networks (ANN) or just Neural Networks (NN) are objectively the main tool in machine learning appropriate for handling large data sets. Neural Networks are a combination of “neurons” and “synapses” consisting of three main components: An input layer, a number of hidden layers and an output layer. These three parts create what is called an n-layer Neural Network. Each layer is connected with a set of weights and a bias value to the next one. Also, in each hidden layer a choice of activation function must be defined, but if that is fixed in the beginning of the analysis, only the weights and bias values will affect the output, thus training a Neural Network is a process of fine tuning the weights and bias values to get a better accuracy through a complicated Stochastic Gradient Descent method.

Every iteration in training the neural network contains two main steps: **Backpropagation** and **Feedforward**. Feedforward is the process of calculating the predicted output and Backpropagation is the process of updating weights and biases after a specified number of iterations.

Below are the major components of a perceptron:

1) Inputs: All the features available in the training dataset become the input for a perceptron. Also, an extra value called a bias value is fed as one of the inputs.

2) Weights: The value of weights are initiated randomly (most of the times zero for all) and these values are updated accordingly by reviewing the training error

3) Weighted sum: This is the summation of all the values obtained after multiplying each weight with its associated input value and adding the bias at the end.

4) Activation function: These functions convert an input signal of a node to an output signal. Some of the commonly used activation functions are **tanh**, **sigmoid**, **relu**, **softmax**, exponential and linear. The flexibility of these activation functions is one of the reasons neural networks perform better than traditional multilinear models.

5) Output: The weighted sum is passed into the activation function and becomes the input 13 value of the next layer. As a first step, the weight vector is initialized. All the features available in the training dataset are fed as input to the perceptron. These input features are then multiplied with the corresponding weights and the values are summed up including the bias value. The new computed value is fed to the activation function in order to get the predicted output. If the predicted value doesn’t match with the actual value, the error is calculated and the weights are updated in order to reduce the error for the next iteration. This process is repeated until the error is reduced to a prescribed level, or if a certain number of steps is achieved.

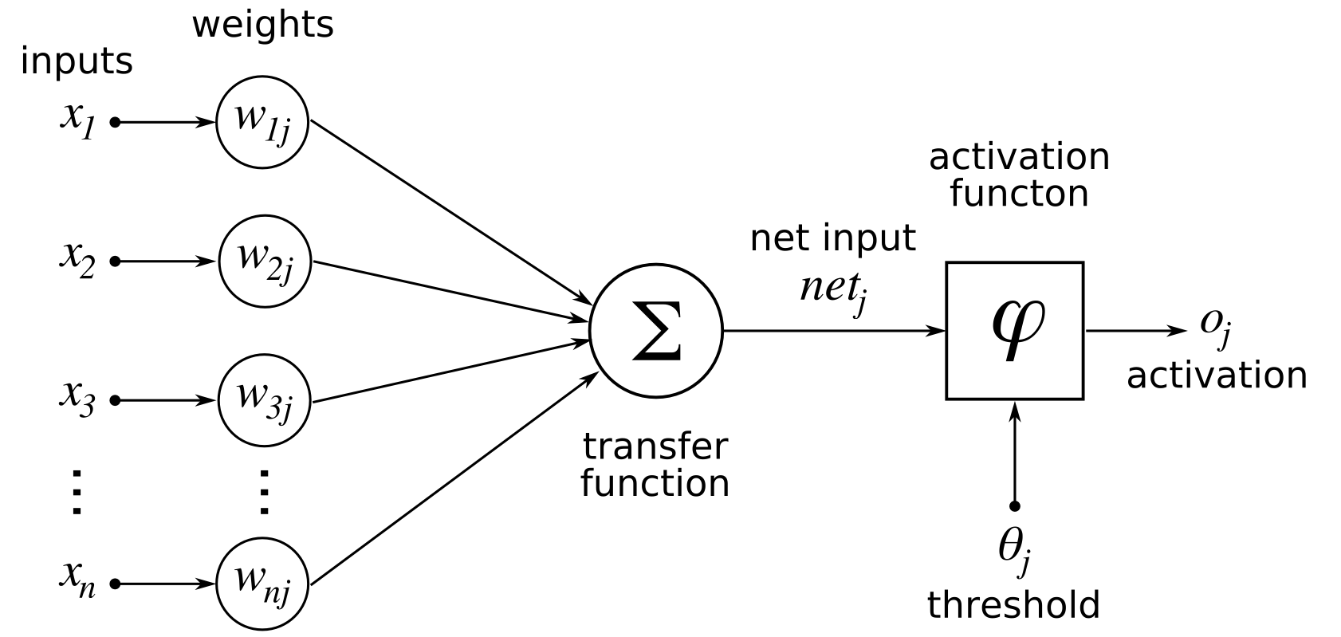
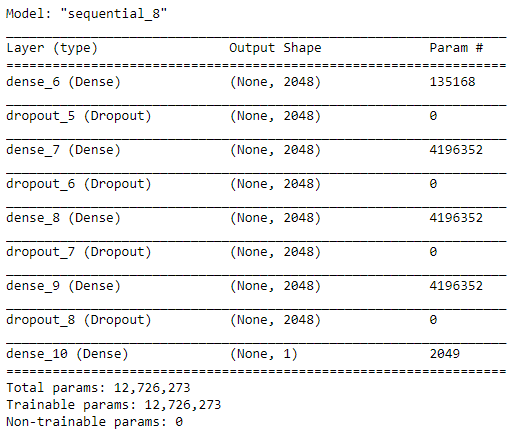


Figure : Artificial Neural Network (Perceptron Basic Model)

Our ANN model in compiled form given below shows number of nodes and hidden layers.



# Chapter 5: Experiments and Results

## 5.1 Statistics

In this section, we present some of the basic numerical descriptors of our dataset and the results of our initial analysis. Following are some statistics from our data.

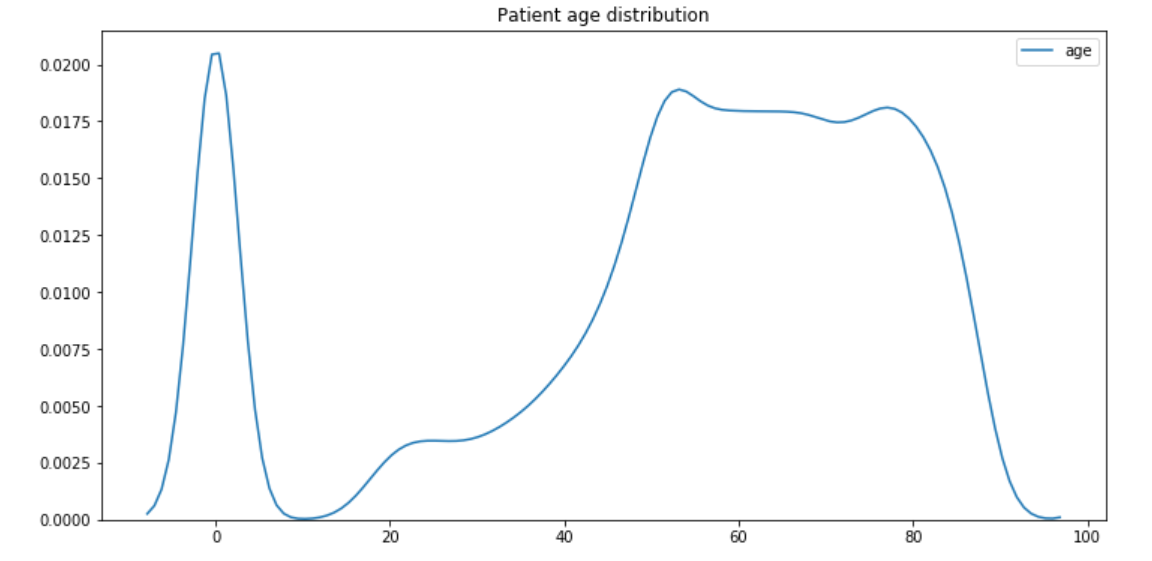


Figure : Patient Age Distribution

The above figure shows the patient age distribution which clearly impacted every feature of patient.

Similarly, we plotted the length of stay distribution for patients. The figure below shows that:

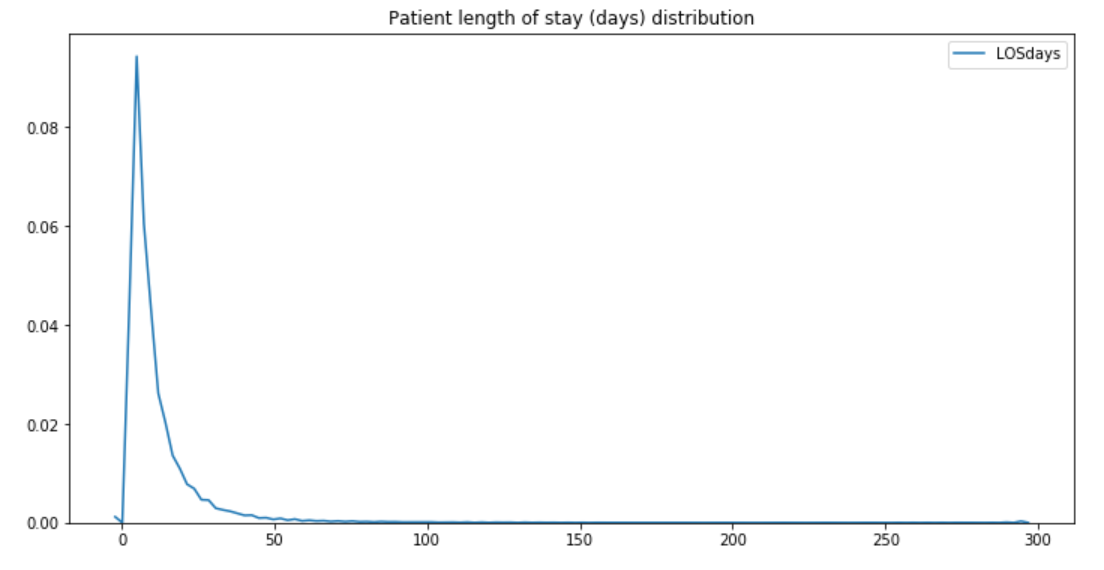


Figure : Patient Length of Stay Distribution

As we have considered the insurance types as well. Insurance type do impact in cases of complications because the complications are not only based on certain medication only, they also are based upon certain tools, systems or an items used on patients due to which a particular complication occurred. For example, complications related to stents. The insurance covers such type of things that a patients will get what kind of services hence we did not neglect this feature. Following is the distribution on types of insurances.

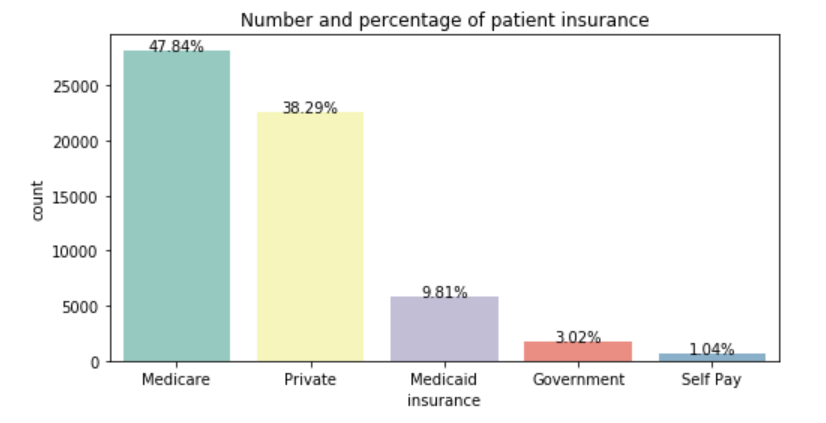


Figure : Insurance Types Distribution

## 5.2 Results

After the data preparation, problem statement and the decision of performance metric there were several model candidates to run. The candidate models that we selected are Logistic Regression, Linear SVC (SVM), Random Forrest and ANN.

First we standardized our data and normalized our data frames to be passed to models to predict the complications. We used standard min-max Scaler for normalization of our data. In case of the neural network classifier we used a multilayer perceptron for binary classification that uses the **ReLU** activation function.

Before each model, two data sampling processes were executed. SMOTE and ADASYN as we had class imbalance problem. After getting the interpolated data we passed that dataset to each of the model to predict the complications.

The different result metric resulted from each of the model against each data sampling technique is shown in below table:

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | AUROC |
| Logistic Regression | 65 % | 0.72 |
| Linear SVC | 66% | 0.72 |
| Random Forrest | 86% | 0.83 |
| ANN | 81% | 0.82 |

Table : Results with ADASYN Data Sampling

The above table showed result after applying the ADASYN data sampling technique. After that we implement SMOTE as well to see if the models vary with the type of static sampling instead of interpolating data based on the distribution as in ADAYSN

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | AUROC |
| Logistic Regression | 67 % | 0.72 |
| Linear SVC | 67% | 0.73 |
| Random Forrest | 85% | 0.86 |
| ANN | 84% | 0.83 |

Table : Results with SMOTE Data Sampling

The later table shows result with SMOTE data sampling technique. It clearly shows that RF still outperforms other models.

# References

**[1]** Johnson, Alistair EW, Tom J. Pollard, Lu Shen, H. Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3 (2016): 160035.

**[2]** Gentimis, Thanos, Alnaser Ala'J, Alex Durante, Kyle Cook, and Robert Steele. "Predicting hospital length of stay using neural networks on mimic iii data." In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pp. 1194-1201. IEEE, 2017.

**[3]** Huang, Jinmiao, Cesar Osorio, and Luke Wicent Sy. "An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes." *Computer Methods and Programs in Biomedicine* 177 (2019): 141-153.

**[4]** Yao, Liang, Chengsheng Mao, and Yuan Luo. "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks." *BMC medical informatics and decision making* 19, no. 3 (2019): 71.

**[5]** Johnson, Alistair EW, Tom J. Pollard, and Roger G. Mark. "Reproducibility in critical care: a mortality prediction case study." In *Machine Learning for Healthcare Conference*, pp. 361-376. 2017.

**[6]** Arvind, Varun, Jun S. Kim, Eric K. Oermann, Deepak Kaji, and Samuel K. Cho. "Predicting Surgical Complications in Adult Patients Undergoing Anterior Cervical Discectomy and Fusion Using Machine Learning." *Neurospine* 15, no. 4 (2018): 329.

**[7]** Rojas, Juan C., Kyle A. Carey, Dana P. Edelson, Laura R. Venable, Michael D. Howell, and Matthew M. Churpek. "Predicting intensive care unit readmission with machine learning using electronic health record data." *Annals of the American Thoracic Society* 15, no. 7 (2018): 846-853.

**[8]** Meyer, A., Zverinski, D., Pfahringer, B., Kempfert, J., Kuehne, T., Sündermann, S.H., Stamm, C., Hofmann, T., Falk, V. and Eickhoff, C., 2018. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, *6*(12), pp.905-914.

**[9]** Deliberato, Rodrigo Octavio, Ary Serpa Neto, Matthieu Komorowski, David J. Stone, Stephanie Q. Ko, Lucas Bulgarelli, Carolina Rodrigues Ponzoni, Renato Carneiro de Freitas Chaves, Leo Anthony Celi, and Alistair EW Johnson. "An evaluation of the influence of body mass index on severity scoring." *Critical care medicine* 47, no. 2 (2019): 247-253.

**[10]** Naik, Girish S., Sushrut S. Waikar, Alistair EW Johnson, Elizabeth I. Buchbinder, Rizwan Haq, F. Stephen Hodi, Jonathan D. Schoenfeld, and Patrick A. Ott. "Complex inter-relationship of body mass index, gender and serum creatinine on survival: exploring the obesity paradox in melanoma patients treated with checkpoint inhibition." *Journal for immunotherapy of cancer* 7, no. 1 (2019): 89.

**[11]** Bose, Somnath, Alistair EW Johnson, Ari Moskowitz, Leo Anthony Celi, and Jesse D. Raffa. "Impact of intensive care unit discharge delays on patient outcomes: a retrospective cohort study." *Journal of intensive care medicine* 34, no. 11-12 (2019): 924-929.

**[12]** Deliberato, Rodrigo Octávio, Stephanie Ko, Matthieu Komorowski, M. A. Armengol de La Hoz, Maria P. Frushicheva, Jesse D. Raffa, Alistair EW Johnson, Leo Anthony Celi, and David J. Stone. "Severity of illness scores may misclassify critically ill obese patients." *Critical care medicine* 46, no. 3 (2018): 394-400.

**[13]** Davoodi, Raheleh, and Mohammad Hassan Moradi. "Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier." *Journal of biomedical informatics* 79 (2018): 48-59.

**[14]** Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

**At the end**

**ICD 9** <https://www.cdc.gov/nchs/icd/>

2) ICD 3) CrispDM4) derived concepts , In conclusion remaining,

**THESIS GUIDELINES**

**Thesis Cover:**

For thesis, the following color scheme will be used:

* Bachelor Thesis: Blue
* Master Thesis: Dark Maroon
* Ph.D. Thesis: Black

**Format for Spine:**

* For spine of the thesis, the student will put ***Degree title i.e. MSCS***, ***Thesis title*** and mention ***Session***.

**Font Size and Style:**

* The ***Headings*** in the thesis will be in ***Times New Roman*** with font size ***16.***
* The **Body** of the thesis will be in ***Times New Roman*** with font size ***12.***
* The line spacing will be *1.5.*

**Referencing:**

* The student should follow ***IEEE*** referencing style.

**Appendices:**

* Provide appendices at the end of the thesis labeling as appendix A, appendix B and so on.

1. <https://mimic.physionet.org/gettingstarted/overview/> [↑](#footnote-ref-1)
2. <https://mimic.physionet.org/gettingstarted/access/> [↑](#footnote-ref-2)
3. <https://mit-lcp.github.io/mimic-schema-spy/> [↑](#footnote-ref-3)
4. <https://pandas.pydata.org/> [↑](#footnote-ref-4)