# Predicting Diabetics Readmission using Machine learning Algorithms

## Introduction

A hospital readmission is when a patient discharged from the hospital is readmitted within a specified time frame. Time frame generally varies from 30 day to 90 days also 1 year sometimes. The first admission in an hospital is called as "index admission" and the subsequent admission is called as "readmission". In general, readmission indicates the ineffectiveness of treatment and lack of healing during past hospital care.
Given the prevalence of diabetes in the general population and hospitalized patients, diabetes is a major contributor to acute care re-utilization and associated costs.In 2016, over 10 million hospital discharges involved patients with diabetes. The annual cost of readmissions within 30 days of discharge (30-day readmission) is $20–25 billion based on the most inclusive readmission rates of 16.0 to 20.4% among patients with diabetes in the USA Hospital .Studies show that diabetes increases the risk of 30 -day readmission by at least 17% and up to 2.5 times compared to patients without diabetes.readmission is an unavoidable problem for both patients and hospital.Machine Learning (ML) has appeared as a powerful Readmission prediction tool and offered data based on active intervention data. Recent research (2019-2024) on the prediction of the ML-based readmission, focusing on key methodologies, challenges and future directions, is exploring this literature. Machine Learning (ML) has appeared as a powerful Readmission prediction tool and offered data based on active intervention data. Recent research (2019-2024) on the prediction of the ML-based readmission, focusing on key methodologies, challenges and future directions, is exploring this literature.

**History:**
Readmissions in the hospital first appeared in medical literature in 1953 at the work of Moya Woodside investigating the results in psychiatric patients in London. Research of the health services gradually examined the hospital readmissions, increasing costs of health care and recognition that certain groups of patients were high consumers of health care services.
In 2007, the Centers for Medicare and Medicaid Services (CMS) reported congress that in 2005, 17.6% of patients readmitted in hospital in a 30 day time frame at a cost of $15 billion on government. They started tracking the hospital readmission rate to improve the healthcare treatment and facilities.

In 2009, CMS started sharing the hospital readmission rate of heart attacks, vascular surgery and heart failure problems to the hospitals. In order to this, CMS contracted with the Yale-New Haven Services Corporation/Center for Outcomes Research and Evaluation (CORE) to develop a hospital-wide readmission (HWR) measure, which it began publicly reporting on Hospital Compare in 2013.

When patients are readmitted shortly after their initial hospital stay, it indicates suboptimal quality of care, inadequate patient education, and challenges transitioning from hospital to home. In the last decade, more programs have been introduced to reduce hospital readmissions, as this has become a top priority in U.S. healthcare reform.
Understanding the root causes of readmission in the hospital is essential for the development of effective strategies of hospital readmission reduction program. While numerous readmissions cannot be prevented, they may represent the progression of the disease, a separate problem not related to acceptance or difficulties in adherence to the discharge plan.There can be various reasons for readmission.

Some reasons for hospital readmissions are:
**During Hospital:**
1. Poor Handoffs:Doctors and nurses sometimes cannot transmit important information about patients to move responsibility for care. Many subsequent doctors do not receive a complete summary of the release of the hospital in time, which can lead to errors.

2. Medication problems: Changing medications during the treatment can lead to harmful reactions in the body causing side effects . Polypharmacy increases risk of complications.
3. Poor hospital management: Sometimes, due to untidy hospital the risk of more complications increases in patients due to the unnecessary infections carried from one patient to other.
4. Poor healing technique: Sometimes after surgery, the hospital do not take good care in the process of healing leading to a new health issue.

**At Discharge:**

1. Therapeutic Mistakes: Errors in treatment instructions (incorrect doses or duplicate regulations) can lead to serious health problems after leaving the hospital.
2. Early discharge: Patients who are discharged too soon may not be fully treated, which increases the chance of return of same disease.

**After Discharge:**

1. Insufficient, delayed or absent monitoring: Only half of the recipients of Medicare who had to be readmitted within 30 days of release had a subsequent visit to the doctor.
2. Inadequate care: patients do not have to receive appropriate follow -up care such as treatment, physical therapy or home health care. This can worsen their condition and require a return to the hospital.
3. Misunderstanding communication: poor communication between healthcare providers and insufficient care coordination can lead to readmissions. Also, the language barrier among health care providers, employees and patients can lead to further misinformation because some things are "lost in translation".
4. Illiterate patients: There are many readmissions because patients do not fully understand their conditions or how to drive them. Insufficient education of patients may lead to non -compliance with treatment plans and aftercare.
5. Social determinants Health: Social and economic circumstances of patients may affect their ability to follow treatment recommendations. Factors such as transport problems, food uncertainty and housing instability can increase the risk of backward takeover.

## Major Risk Factors For Diabetes:

Many factors increase the risk of readmission, including:

- **Sociodemographics** (e.g., age, race, income).
- **Comorbidities** (e.g., heart disease, kidney disease).
- Insulin use.
- Length of stay (LOS) in the hospital.
- History of prior readmissions.

## LITERATURE REVIEW:

---

### 1. Effective hospital readmission prediction models using machine-learned features

Sacha Davis1*, Jin Zhang2, Ilbin Lee2, Mostafa Rezaei4, Russell Greiner1,5, Finlay A. McAlister3 and Raj Padwal3

Data development and features:

A large data set containing 428,669 index hospitalizations has been analyzed, with the exception of birth and psychiatric cases. The data file was divided into 11 of the same parts, with one subset is used for

testing , while the remaining 10 subsets were used for 10-fold-cross-validation. This robust validation technique has ensured that predictive performance of various models can be reliably compared.

Feature Engineering :

To effectively represent patient data, the study used both the selection of features of manual features and an unsupervised feature encoding technique- Word2VEC. Word2vec, originally designed for the processing of natural language, was applied to coding longitudinal information about patients by continuous implementation of vacuum-sovelligers from the Python NLTK library. The inclusion of Word2vec functions allowed a richer representation of the patient's history and complemented manually selected functions.

Model selection and comparison:

Two machine learning models were considered to predict readmission: logistics regression (LR) and gradient Boost(GBM). In addition, a logistics regression model based on LACE score was used as a base line for comparison. The results showed that the combination of manual and Word2vec features brought the highest predictive accuracy across all models:

LR with combined features achieved AUC 0.786 ± 0.0058.

GBM with combined features reached AUC 0.814 ± 0.0045.

The tuned GBM further improved the AUC 0.825 ± 0.0045.

Word2vec function itself overcame manual functions within the LR model, achieving AUC 0.757 compared to 0.747 for manual functions. On the other hand, within the GBM model, manual functions performed better (AUC = 0.804) than Word2vec (AUC = 0.768). In particular, the logistical regression regression line based on lace -based, with only 0.655, with only 0.655, which strengthens the superiority of machine noble elements in Readmission prediction.

LIMITATIONS:

The study predicts readmissions also at discharge because their models used some variables from the index hospitalization episode in addition to records from before the index admission. But, it is more desirable to predict readmission at the time of admission than the time of discharge, also the prediction timing of most past studies is at discharge.Although Word2vec improved AUC, it need careful tuning also the quality of dataset should be relevant to it.While the study provides promising results, real-time implementation in clinical settings was not evaluated.

**2. Predicting and Preventing Acute Care Re-Utilization by Patients with Diabetes**

Daniel J. Rubin1 & Arnav A. Shah2

Data exploration:

Predictive models for the hospital readmission are strongly relying on data from electronic health records (EHRS), which include patient demography, clinical history, laboratory results and patterns of use of health care. Especially machine learning models (ML) use large data to analyze complex relationships and improve prediction accuracy. However, the challenges of these models can limit challenges such as incomplete or inaccurate data in the EHR.

Model learning models:

The ML models revolutionized the prediction of readmissions by analyzing large data sets and identification of formulas that could miss traditional models. These models, such as random forests and neural networks, have achieved high accuracy (C-statistics up to 0.97) by incorporating thousands of variables. However, many ML models lack external verification, which means that their efficiency in the real world outside the original data file remains uncertain.

Limitations :

Despite their potential, predictive models face several restrictions. First of all, many models prefer non-modifiable risk factors (eg age, comorbidity) that limit their usefulness for designing targeted interventions. Second, integrating these tools into clinical workflows requires significant investments in technology and training. Finally, concerns about personal data protection and compliance with regulations such as HIPAA can prevent the use of data on predictive modeling.

### 3. Performance Characteristics of a Machine-Learning Tool to Predict 7-Day Hospital Readmissions

John M Morrison, MD, PhD,a,b Brittany Casey, MD,b

Dataset Information:

The dataset, obtained from the UCI Machine Learning repository, contains clinical care data of 10 years (1999-2008) at 130 US hospital networks. It has 50 features and 101766 patient records with outcomes. Only information of encounters that satisfied the following criteria were considered:
- It is an inpatient case, i.e. the patient was admitted in a hospital, for at least 1 day (and at most 14 days).
- Diabetes was diagnosed in the patient (irrespective of the associated diagnoses).
- Medications were given and laboratory tests were performed on the patient.

The dataset contains additional attributes i.e. patient race, age, gender, time in hospital, medical specialty of associated disorders, A1c test result, admission type (outpatient, inpatient or emergency), number of laboratory tests performed, number and nature of previous visits etc.

Methods:

They checked the prevalence of readmission within 30 days in the dataset using the following formula:
Prevalence % = [(Number of instances of specific trait) / (Number of instances measured in total)] x100
Machine learning methods for predicting hospital readmissions include **KNN** (simple but sensitive to imbalanced data), **Logistic Regression** (interpretable but struggles with non-linear data), **Naïve Bayes** (fast but assumes feature independence), **Decision Trees** (easy to interpret but prone to overfitting), **Gradient Boosting** (high accuracy but computationally intensive), and **Random Forest** (robust but less interpretable). Feature reduction improved model performance, with Gradient Boosting and Random Forest often achieving the best results.

### 4. Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients

Saumya Salian1, Dr. G. Harisekaran2

Data Collection:

**Hospital Readmissions Diabetes Data Set**

Dataset Details:

The data that is obtained is highly noisy and variant in nature. It is necessary to closely examine the data and select the right kind of attributes that can help in obtaining accurate results. This data set was obtained from the UCI Repository of Machine Learning Databases. The data set was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. The data set is first loaded into Hadoop File System (HDFS).The preprocessing of data is required to modify the data into classifiable data. This includes removing null values, missing data, and outliers using tools like HiveQL, which runs on Hadoop Distributed File System (HDFS) for scalable storage and processing.Data is structured into a suitable format, and binary class labels (e.g., 1 for readmission, 0 for no readmission) are defined using HiveQL queries. Tools like Hive and MapReduce enable efficient data summarization, querying, and transformation, ensuring the data is clean and ready for machine learning models.

## Methodology:

The process of predictive modeling includes the use of hadoop and R studio to analyze data stored in HDF. Data file diabetes from the Hive warehouse is used to create predictive models. Classification techniques such as logistics regression, tree decision -making and support vector machine (SVM) are used, with data divided into training ratio to test 7: 3. The decision tree and SVM showed the lowest level of errors. The correlation matrix identifies the relevant properties and the final model is selected by the decision -making tree, with branches representing functions and sheets representing classification. The model predicts the patient reading and random forest is used to identify key predictors such as glucose, body weight index, age, pregnant and pedigree as predictors of the first 5 readmisions.

## Result:

The analysis begins with a correlation matrix for identifying relationships between features and reverse. Plasma glucose shows the highest correlation (0.466), while blood pressure has the lowest (0.065). The tree of the decision -making model is created using all 8 functions that generate classification rules based on sill values for glucose, age and body weight index. For example, if plasma glucose <127.5 and age <28.5, the patient is classified as a “tested_negative” for readmission. It is then evaluated variable with plasma glucose (0.7881), age (0.687) and body weight (0.686) identified as the most important predictors, while insulin is least important. Finally, recursive elimination of elements ranks the top 5 predictors: plasma glucose, age, body weight index, pregnant and pedigree, which are used to specify the model for better performance.

**DATA-SPECIFIC INFORMATION FOR: [diabetic_data.csv]**

- Number of variables: 50
- Number of instances/rows: 101766
- Target Variable is Readmitted. The Values which are <30, if the patient was readmitted in less than 30 days and which are >30 ,if the patient was readmitted in more than 30 days, and “No” for no record of readmissionFeatures.

## 5. Machine learning in predicting pediatric readmission

## Data development :

Rodriguez et al. (2022) conducted a retrospective cohort study for the development of machine learning models (ML) to predict a 30 -day pediatric readmission. Studies analyzed 9,080 patients under 18 years of

age from a tertiary university hospital using demographic, clinical and biochemical data. The data file was divided into training (75%) and testing (25%) subgroups, with techniques such as Downsampling, repeated cross validation and tuning of hyperparameters applied to increase the performance of the model.

Model learning models (ml)

- Six ML algorithms were tested, including:
  - **XGBOOST** (a powerful ML method),
  - Random Forest,
  - Logistic Regression, and others.
- Techniques like **downsampling**, **cross-validation**, and **hyperparameter tuning** were used to improve model performance.

Key Findings:

The XGBOOST did best and achieved AUC 0.814 (accuracy rate where 1.0 is perfect).
Important predictors of Readmission are:Diagnosis of cancer,Age, number of red blood cells,Sodium levels,Emergency admission and complex health problems.The study shows that ML models, especially XGBOOST, can help hospitals to identify children with a high risk of taking back. If these models are integrated into electronic health record systems (EHR), doctors and nurses could use them to provide better care and prevent unnecessary hospital stays.

Limitations:

The study was performed in one hospital, so the model may not work as well in other environments.
The integration of these tools into hospital systems requires solving technical challenges and concerns about personal data protection.

Conclusion:

The findings emphasize the efficacy of ML, in particular XGBOOST, in predicting pediatric readmission and enabling timely identification of high-risk patients. The authors suggest that the integration of these models into EHR systems could optimize healthcare interventions and reduce unnecessary hospitalization. Future research should focus on verifying the model in various hospital environments and patient populations to ensure wider usability.

## DATA EXPLORATION:

---

UCI Machine Learning Datasets Repository
The dataset represents 7 years (2018 - present) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.

1. It is an inpatient encounter (a hospital admission)
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

**Predictors:**

1. **Encounter ID** – Unique identifier of an encounter.
2. **Patient number** – Unique identifier of a patient.
3. **Race** – Categories: Caucasian, Asian, African American, Hispanic, and Other.
4. **Gender** – Categories: Male, Female, and Unknown/Invalid.
5. **Age** – Grouped in 10-year intervals: (0-10), (10-20), …, (90-100).
6. **Weight** – Patient's weight in pounds.
7. **Admission type** – Integer identifier corresponding to 9 values (e.g., Emergency, Urgent, Elective, Newborn, Not Available).
8. **Discharge disposition** – Integer identifier corresponding to 29 values (e.g., Discharged to home, Expired, Not Available).
9. **Admission source** – Integer identifier corresponding to 21 values (e.g., Physician referral, Emergency room, Transfer from a hospital).
10. **Time in hospital** – Number of days between admission and discharge.
11. **Payer code** – Integer identifier for 23 payer types (e.g., Blue Cross/Blue Shield, Medicare, Self-pay).
12. **Medical specialty** – Integer identifier for 84 physician specialties (e.g., Cardiology, Internal Medicine, General Practice).
13. **Number of lab procedures** – Total number of lab tests performed during the encounter.
14. **Number of procedures** – Total number of non-lab procedures performed.
15. **Number of medications** – Number of distinct medications administered during the encounter.
16. **Number of outpatient visits** – Number of outpatient visits in the year before the encounter.
17. **Number of emergency visits** – Number of emergency visits in the year before the encounter.
18. **Number of inpatient visits** – Number of inpatient visits in the year before the encounter.
19. **Diagnosis 1** – Primary diagnosis (first three digits of ICD9 code).
20. **Diagnosis 2** – Secondary diagnosis (first three digits of ICD9 code).
21. **Diagnosis 3** – Additional secondary diagnosis (first three digits of ICD9 code).
22. **Number of diagnoses** – Total number of diagnoses recorded.
23. **Glucose serum test result** – Values: ">200", ">300", "normal", or "none" (if not measured).
24. **A1c test result** – Values: ">8" (>8%), ">7" (between 7% and 8%), "normal" (<7%), or "none" (not measured).
25. **Change of medications** – Indicates if there was a change in diabetes medication. Values: "change" or "no change".
26. **Diabetes medications** – Indicates if any diabetes medication was prescribed. Values: "yes" or "no".
27. **Medication details (24 features)** – For specific diabetes medications (e.g., metformin, insulin, etc.), the values indicate:

- "up" (dosage increased)

- "down" (dosage decreased)
- "steady" (dosage unchanged)
- "no" (not prescribed).

## Technical Requirements

---

**1. Data Preprocessing:**

- We identified unique values in each column and determined unique columns.
- We conducted exploratory data (EDA) analysis with suitable visualizations for categorical variables and backward trends.
- Also manipulated the missing values by replacing the average values.
- We removed columns with excessive missing values such as "weight", "Payer_code" and "Medical_specialty".
- Removed columns with constant values such as "Citoglipton" and "examide".
- Removed values with unknown sex and missing information about diagnostics.
- Excluded encounters leading to hospice discharge or patient death to avoid bias.

**2. Imputation of missing values**

- We replaced the missing values with average values for numerical attributes.
- Categorical variables with unknown values were standardized (eg "?" The race changed to "unknown").
- Consolidated similar types of admissions (eg urgent and trauma in emergency).
- Mapped null and unmapped types of admission to one category ("not available").

**3. The function of engineering**

- We created a list of categorical variables for coding.
- Also Identified and deleted duplicate records.
- Detected and removed outliers beyond three standard deviations to reduce model bias.
- Calculated correlation matrix and removed highly correlated or redundant function.
- Defined predictor variables (x) and target variable (s) when examining class imbalances.

**4. Addressing Imbalanced Data and Resampling Techniques**

- Resample the training set
- Oversampling
- Undersampling
- Repeated Stratified K fold Cross validation
- Resample with different ratios
- Ensemble different resampled datasets
- Use the right evaluation metrics