BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

WORK INTEGRATED LEARNING PROGRAMMES (WILP)

MID-SEMESTER REPORT

Dissertation Work Title:

Automated Machine Learning (AutoML) for Healthcare Risk Prediction

Course No.: DSECLZG628T

Course Title: Dissertation / Project / Project Work

Dissertation / Project /Project Work Done by:

Student Name: Keerthi Kumar B
BITS ID: 2023DC04097
Degree Program: M.Tech. in Data Science & Engineering
Research Area: Health Care Management
Dissertation / Project Work carried out at: Individual Research Project (Executed Using Personal Computational Resources)

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

VIDYA VIHAR, PILANI, RAJASTHAN - 333031.

December 2025

Abstract

Healthcare datasets contain valuable information that can be leveraged to identify individuals at high risk of serious medical conditions such as stroke. However, developing accurate and reliable machine learning models for healthcare applications typically involves extensive manual experimentation, careful preprocessing, and expert-driven model selection. Automated Machine Learning (AutoML) aims to reduce this complexity by automating key stages of the machine learning lifecycle, thereby improving development efficiency and reproducibility.

This dissertation project focuses on healthcare risk prediction using a publicly available stroke prediction dataset. The objective is to establish a strong analytical baseline through exploratory data analysis, preprocessing, and manual model development, which will later be used to evaluate the effectiveness of AutoML-based approaches. During the mid-semester phase, exploratory data analysis was conducted using Python and Tableau to understand feature distributions, class imbalance, and clinically relevant risk factors.

The dataset analysis revealed a significant class imbalance, with 4,861 non-stroke cases and 249 stroke cases, highlighting the need for careful evaluation strategies. Key risk indicators such as age, body mass index (BMI), and average glucose level showed clear differences between stroke and non-stroke populations. Data preprocessing steps included missing value treatment, categorical feature encoding, feature scaling, and stratified train–test splitting.

A baseline Logistic Regression model was implemented and evaluated using appropriate classification metrics suited for imbalanced healthcare data. The results obtained during this phase establish a reproducible reference point for comparison with AutoML models in subsequent stages. The mid-semester work validates the feasibility of the proposed approach and lays a strong foundation for automated modeling, explainability, and deployment in the final phase.

Signature of the Student	**Signature of the Supervisor**
Name: Keerthi Kumar B	**Name: Tejas Kambale**
Date: 21-12-2025	**Date: 21-12-2025**
Place: Bangalore	**Place: Bangalore**

Contents

1. Introduction and Motivation

The rapid digitization of healthcare systems has resulted in the generation of large volumes of patient-related data, including demographic details, clinical measurements, and lifestyle information. When analyzed effectively, this data has the potential to support early disease detection, risk stratification, and preventive healthcare planning. One such critical application is the prediction of stroke risk, where timely identification of high-risk individuals can significantly reduce mortality and long-term complications.

Machine learning techniques have been widely adopted in healthcare analytics to uncover hidden patterns in patient data. However, building reliable machine learning models for healthcare applications is a complex process. It typically involves extensive exploratory data analysis, careful handling of missing values, selection of suitable algorithms, feature engineering, and repeated hyperparameter tuning. These steps require significant domain knowledge and manual effort, making the development process time-consuming and difficult to scale.

Automated Machine Learning (AutoML) has emerged as a promising approach to address these challenges. AutoML frameworks aim to automate critical stages of the machine learning pipeline, including model selection, feature handling, and hyperparameter optimization. By reducing human intervention, AutoML can accelerate model development while maintaining competitive performance. This is particularly valuable in healthcare environments, where rapid experimentation and reproducibility are essential.

Despite the growing popularity of AutoML, its effectiveness in healthcare risk prediction must be evaluated carefully. Healthcare datasets often exhibit class imbalance, missing values, and strict requirements for interpretability. Therefore, it is important to establish a strong manual machine learning baseline before adopting AutoML solutions. This project is motivated by the need to systematically compare traditional manually developed models with AutoML-based approaches in the context of stroke risk prediction.

The work completed during the mid-semester phase focuses on building this foundation. Through exploratory data analysis using Python and Tableau, key risk factors and data challenges were identified. Data preprocessing and a baseline Logistic Regression model were implemented to provide a transparent and interpretable reference. This groundwork ensures that subsequent AutoML experiments can be evaluated meaningfully in terms of performance, efficiency, and clinical relevance.

2. Problem Definition and Research Objectives

The problem addressed in this project is the prediction of stroke risk using patient demographic, clinical, and lifestyle data. The task is formulated as a binary classification problem, where the objective is to predict whether a patient is likely to experience a stroke (1) or not (0) based on available features such as age, hypertension, BMI, and average glucose level.

Healthcare datasets pose specific challenges, including class imbalance, missing values, and mixed feature types. Addressing these challenges while maintaining interpretability and reproducibility is a key requirement of this project.

3. Dataset Description

The dataset used in this project is the Stroke Prediction Dataset, a publicly available healthcare dataset obtained from Kaggle. It contains patient-level demographic, clinical, and lifestyle attributes and is suitable for binary classification tasks in healthcare risk prediction.

3.1 Dataset Overview

- **Total records:** 5,110
- **Target variable**: stroke
- **Non-stroke (0):** 4,861 records
- **Stroke (1):** 249 records
- **Problem type:** Binary classification

The distribution of the target variable clearly indicates class imbalance, with stroke cases forming a small minority of the dataset. This observation influenced the choice of stratified sampling and ROC-AUC as a primary evaluation metric.

3.2 Feature Description

The dataset includes the following categories of features:

- **Demographic features:** age, gender, residence type
- **Clinical features:** hypertension, heart disease, BMI, average glucose level
- **Lifestyle features:** smoking status, work type

3.3 Dataset Challenges

The following challenges were identified during exploratory analysis:

- **Class imbalance:** Stroke cases account for less than 5% of the data.
- **Missing values:** BMI contains missing entries.

- **Mixed data types:** Combination of numerical and categorical variables.

These challenges guided the preprocessing and modeling strategies adopted in the project.

4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to understand the structure of the dataset, identify data quality issues, and examine relationships between input features and stroke occurrence. Python libraries were used for detailed analysis, while Tableau was used for visual exploration and presentation of key insights.

4.1 Class Distribution Analysis

The analysis revealed a significant class imbalance in the dataset. Out of 5,110 records, 4,861 correspond to non-stroke cases, while only 249 records represent stroke cases. This imbalance is typical in healthcare datasets and highlights the importance of using appropriate evaluation metrics and sampling strategies.

4.2 Age Distribution

Age-based analysis showed that stroke cases are predominantly concentrated in higher age groups. Very few stroke occurrences were observed in younger age ranges, while a noticeable increase in stroke cases was observed for patients above approximately 50 years of age. This trend aligns with established clinical understanding that stroke risk increases with age.

4.3 BMI and Stroke Relationship

The relationship between BMI and stroke was examined using box-and-whisker plots. Stroke cases exhibited a higher median BMI compared to non-stroke cases, along with greater variability and the presence of high-BMI outliers. This suggests that BMI may act as an important contributing risk factor in stroke prediction.

4.4 Glucose Level and Stroke Relationship

Analysis of average glucose levels indicated that patients who experienced a stroke generally had higher glucose levels compared to non-stroke patients. The upper quartile of glucose values for stroke cases was significantly higher, reinforcing the role of metabolic factors in stroke risk.

4.5 Tableau Dashboard

To improve interpretability and communication of insights, a Tableau dashboard was created that consolidated stroke distribution, age distribution, BMI comparison, and glucose level comparison into a single visual view. The dashboard provided an intuitive overview of key risk patterns and supported the findings from Python-based analysis.

5. Data Preprocessing

Data preprocessing was carried out to convert the raw healthcare dataset into a machine-learning-ready format while preserving clinical relevance and minimizing bias. The preprocessing steps were designed based on insights obtained during exploratory data analysis.

5.1 Missing Value Handling

The dataset contained missing values in the BMI attribute. Since BMI is a clinically important feature and removing records could lead to loss of valuable information—especially stroke cases—median imputation was applied. Median imputation was chosen as it is robust to outliers and preserves the central tendency of the data.

5.2 Categorical Feature Encoding

Several input variables, such as gender, work type, residence type, and smoking status, were categorical in nature. These features were converted into numerical form using one-hot encoding. This approach avoids introducing artificial ordinal relationships and is suitable for both linear models and automated machine learning frameworks.

5.3 Feature Scaling

Numerical features were standardized using standard scaling to ensure that all features contributed equally during model training. Feature scaling was applied only after splitting the data to avoid data leakage and was particularly important for linear models such as Logistic Regression.

5.4 Train–Test Split

The dataset was divided into training and testing sets using an 80:20 split. Stratified sampling was employed to maintain the original class distribution of stroke and non-stroke cases in both sets. This ensured a fair and reliable evaluation of model performance on unseen data.

5.5 Processed Data Management

To maintain reproducibility, intermediate datasets were stored separately. A cleaned dataset (after missing value handling) and an ML-ready dataset (after encoding) were saved in the processed data directory. This separation supports consistent experimentation and enables future AutoML comparisons without repeating preprocessing steps.

6. Baseline Machine Learning Model

A manual baseline machine learning model was developed to establish a reference point for evaluating AutoML-based approaches in later stages of the project. The baseline model provides transparency, interpretability, and a clear understanding of how the data contributes to prediction outcomes.

6.1 Model Selection

Logistic Regression was selected as the baseline model due to its simplicity, interpretability, and widespread adoption in healthcare risk prediction tasks. Logistic Regression provides probabilistic outputs, which are useful for estimating patient risk levels rather than making only binary decisions.

6.2 Model Training

The model was trained using the preprocessed training dataset. Feature scaling was applied prior to training to ensure stable convergence and balanced feature influence. The model was configured with an increased iteration limit to ensure proper convergence given the number of encoded features.

6.3 Role of the Baseline Model

The baseline Logistic Regression model serves two key purposes:

It establishes a transparent and interpretable benchmark.

It provides a performance reference against which AutoML-generated models will be compared.

By developing a manual baseline, the project ensures that any performance gains achieved through AutoML are meaningful and not the result of improper preprocessing or evaluation strategies.

7. Model Evaluation and Results

The performance of the baseline Logistic Regression model was evaluated using metrics that are appropriate for imbalanced healthcare datasets. Rather than relying solely on accuracy, multiple evaluation measures were considered to better understand model behavior and clinical relevance.

7.1 Evaluation Metrics

The following metrics were used to evaluate the baseline model:

- **Accuracy:** Measures overall prediction correctness.
- **Confusion Matrix:** Provides insight into true positives, false positives, true negatives, and false negatives.
- **ROC-AUC Score:** Evaluates the model's ability to distinguish between stroke and non-stroke cases across different classification thresholds.

Given the significant class imbalance observed during exploratory analysis, ROC-AUC was treated as a more reliable indicator of model performance than accuracy alone.

7.2 Results Interpretation

The confusion matrix analysis revealed that while the model performs well in identifying non-stroke cases, detecting stroke cases remains more challenging due to their minority representation. This behavior is expected in healthcare risk prediction problems and highlights the importance of careful threshold selection and future model enhancement.

The ROC curve demonstrated that the baseline Logistic Regression model performs better than random guessing, indicating that the selected features and preprocessing steps contribute meaningful information for stroke risk prediction. Although the baseline model is not optimized for maximum recall of stroke cases, it establishes a credible reference for further improvement.

7.3 Significance of Results

The evaluation results validate the preprocessing pipeline and confirm that clinically relevant features such as age, BMI, and average glucose level play a meaningful role in prediction. These findings justify proceeding to the next phase, where AutoML techniques will be employed to explore more complex models and improved performance.

8. Challenges and Learnings

During the mid-semester phase, several practical and conceptual challenges were encountered. One of the primary challenges was handling class imbalance, as stroke cases constituted a very small portion of the dataset. This required careful selection of evaluation metrics and stratified sampling to ensure fair model assessment.

Another challenge involved data quality, particularly missing values in clinically important features such as BMI. Addressing this required domain-aware preprocessing decisions to avoid loss of critical information. Additionally, the presence of mixed data types necessitated appropriate encoding strategies to ensure compatibility with machine learning models.

From an implementation perspective, the project reinforced the importance of maintaining a clear and reproducible data pipeline, including separation of raw and processed datasets. The exploratory analysis phase also highlighted the value of combining Python-based analysis with Tableau visualizations for better interpretability and communication of insights.

Overall, the mid-semester work provided valuable learning in healthcare data handling, evaluation of imbalanced classification problems, and the importance of establishing a strong baseline before adopting automated modeling approaches.

9. Work Remaining and Future Plan

The mid-semester phase of the project focused on establishing a strong analytical foundation through data understanding, preprocessing, and baseline model development. The remaining work will build upon this foundation to achieve the final project objectives.

The next phase of the project will involve the implementation of Automated Machine Learning (AutoML) frameworks such as PyCaret or H2O AutoML. These frameworks will be used to automatically explore multiple machine learning algorithms, optimize hyperparameters, and generate ensemble models. The performance of AutoML-generated models will be systematically compared with the manually developed baseline model.

Following model development, explainable AI techniques such as SHAP will be applied to interpret model predictions and identify the most influential features contributing to stroke risk. This step is particularly important in healthcare applications, where transparency and trust are essential.

Finally, a Streamlit-based interactive application will be developed to demonstrate the healthcare risk prediction system. The final phase will also include comprehensive documentation, result analysis, and submission of the complete dissertation.

9.1 Tentative Timeline

Weeks	Activities
Weeks 1–3	Implementation of AutoML frameworks (PyCaret / H2O AutoML) and automated model experimentation
Weeks 4–5	Performance comparison between manual baseline model and AutoML-generated modelscompare with Logistic Regression.

Weeks 6–7	Explainable AI analysis using SHAP for model interpretability and feature importance
Weeks 8–9	Development of Streamlit-based interactive application for healthcare risk prediction
Week 10–11	Result analysis, validation, and refinement of models
Week 12–13	Final dissertation report writing and formatting
Week 14	Final review, corrections, and submission preparation

Table 1: The tentative timeline for completing the remaining phases of the project

10. Conclusion

This mid-semester report presents the progress made in developing a healthcare risk prediction system using machine learning techniques. The work completed during this phase includes dataset selection, exploratory data analysis using Python and Tableau, data preprocessing, and development of a manual baseline machine learning model.

The exploratory analysis provided valuable insights into class imbalance and key clinical risk factors such as age, BMI, and average glucose level. Data preprocessing ensured that the dataset was converted into a machine-learning-ready format while maintaining reproducibility. The baseline Logistic Regression model established a transparent and interpretable reference point for evaluating more advanced AutoML-based approaches in the subsequent phase.

Overall, the project is progressing as planned and is well-positioned for the implementation of AutoML techniques, explainability analysis, and deployment in the next stage. The work completed till mid-semester satisfies the academic requirements and lays a strong foundation for the successful completion of the dissertation.

11. References

- F. Soriano, *Stroke Prediction Dataset*, Kaggle, 2021.
  Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
- Tableau Software, *Tableau Desktop Documentation*,
  Available: https://help.tableau.com/current/pro/desktop/en-us/
- Géron, A., *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., O’Reilly Media, 2019.

- Molnar, C., *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022.
  Available: https://christophm.github.io/interpretable-ml-book/