A black background with white text

Description automatically generated with low confidence

MSc Data Science

Individual Project Module

7PAM2002-0206-2024

Department of Physics, Astronomy and Mathematics

**Data Science FINAL PROJECT REPORT**

**2024-25**

**Semester B**

**Project Title: Recovery Outcome Analysis using Patient Rehabilitation Assessment Scores using Machine Learning**

**Student Name: Rashi Meda**

**Student ID: 23005003**

**Supervisor: William Alston**

**Date Submitted: April 29th, 2025.**

**GitHub:** [**https://github.com/MRashi1/Patient-Recovery-Prediction**](https://github.com/MRashi1/Patient-Recovery-Prediction)

Word Count: 4473 (excluding references and appendices)

**DECLARATION STATEMENT**

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science at the University of Hertfordshire.

I have read the detailed guidance to students on academic integrity, misconduct and plagiarism information at [Assessment Offences and Academic Misconduct](https://ask.herts.ac.uk/assessment-offences-and-academic-misconduct) and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6)

I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed: Meda Rashi

Student Name signature:



Student ID number:23005003

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

**ABSTRACT**

Based on data from early-stage clinical assessments, this study investigates the application of supervised machine learning algorithms to forecast patient recovery outcomes. The main objective is to create a prediction system that can use Day 1 evaluation parameters to estimate Stroke Impact Scale (SIS) Week 4 scores, allowing for data-driven rehabilitation planning.

Anonymized patient records with functional assessment scores from the Berg Balance Scale (BBS), Timed Up and Go (TUG), VRBESS, BESTest, and SIS are included in the collection. These values were obtained at two different timepoints: admission (Day 1) and discharge (Week 4).

To improve interpretability for clinical use, the project also involved creating patient-level recovery dashboards that combined graphical and natural-language summaries. By using publicly available, anonymized data and making sure that results were reported in a transparent manner, ethical, legal, and professional standards were maintained.

This study lays the groundwork for future research into individualized recovery planning and decision support systems in the healthcare industry while showcasing the usefulness of machine learning in predicting rehabilitation outcomes.

Table of Contents

Abstract………………………………………………………………………………………..3

[1 Introduction 6](#_gjdgxs)

[1.1 Scopes and Overview of Research](#_30j0zll) 6

[1.1.1 Scopes](#_1fob9te) 6

[1.1.2 Problem Statements](#_3znysh7) 6

[1.2 Project Aim](#_2et92p0) 6

[1.3 Research Questions](#_tyjcwt) 6

[1.4 Project Objectives](#_3dy6vkm) 7

[2 Literature Review](#_1t3h5sf) 8

[2.1 Previous Research on Machine Learning in Patient Outcome Prediction……...............](#_4d34og8)8

2.2 Challenges in Machine Learning Application…….………………….…………….…10

2.3 Future Directions in Clinical Predictions………………………...…………………...10

2.4 Synthesis of Current Research….……………………………….……………………10

[3 Summary of Research Progress](#_26in1rg) 11

[3.1 Data Collection](#_lnxbz9) 11

[3.1.1 Data Background](#_35nkun2) 11

[3.1.2 Attributes of Data](#_1ksv4uv) 11

[3.2 Proposed Methodology 1](#_44sinio)2

[3.3 Progress of Work 1](#_2jxsxqh)4

[3.3.1 Data Reading 1](#_z337ya)4

[3.3.2 Data Information 1](#_3j2qqm3)4

[3.3.3 Data Cleaning 1](#_1y810tw)5

3.4 Feature Engineering………………………………………………………………..15

3.5 Model Development Progress……………………………………………………...16

[4 Issues Related to Project and Research 1](#_2xcytpi)7

[4.1 Ethical Issues 1](#_1ci93xb)7

[4.2 Legal Issues 1](#_3whwml4)7

[4.3 Professional Issue 1](#_2bn6wsx)7

[4.4 Social Issue 1](#_qsh70q)8

[5 Research Planning 1](#_3as4poj)9

[5.1 Selection of Tool](#_1pxezwc) 19

[5.2 Selection of Technologies](#_49x2ik5) 19

[5.2.1 Data Manipulating and Preprocessing](#_2p2csry) 19

[5.2.2 Advanced Machine Learning Models](#_147n2zr) 19

[5.2.3 Visuvalization and Insights Extraction](#_3o7alnk) 19

[5.3 Selection of Algorithms](#_ihv636) 19

[5.3.1 Predective Modeling](#_32hioqz) 20

[5.3.2 Feature Engineering and Reduction](#_1hmsyys) 20

[5.3.3 Model Interpretability Techniques](#_41mghml) 20

[5.3.4 Residual Analysis for Model Validation](#_2grqrue) 20

5.4 Implementation Plan…………………………………………………………………20

[6 Project Planning](#_3fwokq0) 22

7 Conclusion……………………………………………………………………………….23

[References](#_1v1yuxt) 25

[Appendix](#_4f1mdlm) 31

# Introduction

## Scopes and Overview of Research

* + 1. **Scopes**

The main objective of this study is to analyse and forecast patient recovery outcomes throughout physical therapy, particularly during a four-week clinical assessment phase, using machine learning. To comprehend recovery trends, the scope is limited to quantitative analysis of defined clinical evaluation indicators employing visual data insights and predictive modelling. The initiative supports data-driven clinical decision support in the fields of post-stroke rehabilitation and physical therapy.

### Problem Statements

Decisions about the course of therapy are frequently dependent on clinical judgment and recurring evaluations in clinical rehabilitation settings. Although standardized measurements are provided by established tests like the Stroke Impact Scale (SIS) and Berg Balance Scale (BBS), it is still difficult to predict long-term effects from early-stage scores. This results in several serious issues:

* Limited ability to predict patient recovery quantitatively
* Absence of automatic baseline-metric insights
* Clinical practice requires explainable machine learning techniques.

By using machine learning models based on Day 1 evaluations to forecast SIS Week 4 outcomes and by creating tools for clinical interpretability and transparency, this study tackles these issues.

## 1.2 Project Aim:

This project's main goal is to create a prediction model that, based solely on the results of the initial clinical evaluation taken on Day 1, can reliably forecast a patient's perceived impact of a stroke following four weeks of treatment. Finding the baseline characteristics that have the most effects on recovery and producing comprehensible results that can aid clinicians in making decisions are the secondary goals.

## 1.3 Research Questions

The following research questions are presented which will be addressed with the successful commencement of the research and artefact.

1. Can SIS Week 4 results be reliably predicted by machine learning models based solely on Day 1 data?
2. Which assessments from Day 1 best predict the effectiveness of rehabilitation?
3. What is the predictive performance difference between ensemble and linear regression models?
4. Is it possible to evaluate and explain the model predictions in a way that facilitates clinical decision-making?

## 1.4 Project Objectives

The objectives of the project are to guarantee a methodical, precise, and moral approach to forecasting patient recovery outcomes using clinical rehabilitation data. Here is a summary of them:

**Objective 1:** **Transfer, Clean, and Convert Patient Rehabilitation Information**

Importing the raw dataset, dealing with missing values, encoding categorical variables, and guaranteeing data consistency are all part of the first goal. Preparing a dependable and clean dataset for predictive modelling is the aim.

**Objective 2: Create New Elements That Show Patient Progress**

To better capture recovery trends, new characteristics are developed. An Overall Improvement Score is produced to summarize patient progress across several domains after improvement scores for each clinical test are computed.

**Objective 3: Put Multiple Regression Models into Practice**

Three models are created:

* Using linear regression to compare baselines
* For non-linear relationships, use the Random Forest Regressor.
* XGBoost Regressor for optimal prediction with good accuracy

This makes it possible to evaluate the model robustly using various learning strategies.  
**Objective 4:** **Use Cross-Validation to Apply Hyperparameter Tuning**

Parameters for Random Forest and XGBoost are adjusted using GridSearchCV. Model performance is resilient and not overfit to training data thanks to cross-validation.  
**Objective 5:** **Evaluate Model Performance Making use of R2 and RMSE**

The R2 Score (explained variance) and RMSE (average prediction error) are used to compare all models. A balance of these metrics determines which model performs the best.

**Objective 6:** **Use Feature Importance Analysis to Visualize Model Behaviour**

For ensemble models, feature importance plots are made to show which baseline evaluations have the biggest impact on result predictions. Clinical interpretability is improved as a result.  
**Objective 7:** **Create a Dashboard for Patient-Level Summaries**

Summaries tailored to each patient are produced, including natural language reports summarizing recovery and model projections as well as graphical comparisons of assessment results. This connects clinical decision-making in practice with technical outcomes.  
**Objective 8:** **Assure Professional, Legal, and Ethical Compliance**

Data management adheres to GDPR, patient privacy is protected throughout the project, and outputs are designed to complement professional competence rather than take its place.

# Literature Review

* 1. **Previous Research on Machine Learning in Patient Outcome Prediction**

To improve patient outcome prediction and recovery planning techniques, machine learning has grown in importance in the healthcare industry. Conventional techniques for forecasting patient recovery frequently depended solely on preliminary diagnostic findings or basic demographic data, as well as basic statistical models and medical experience. But with the advent of machine learning, considerably more in-depth, comprehensive analysis is now possible. Clinical evaluation data from patients is subjected to techniques like Linear Regression, Random Forest, and XGBoost to uncover hidden patterns and intricate, non-linear associations that conventional statistical methods could miss. With the use of these models, medical professionals may more accurately forecast patients' recovery paths and deliver more individualized treatment regimens (Smith et al., 2021).

To find patterns that strongly correspond with either favourable or poor recovery results, Random Forest algorithms, for example, analyse assessment scores from initial evaluations, including BBS, TUG, VRBESS, and SIS. These models can forecast future functional gains or degradation risks by learning from past rehabilitation data, giving doctors the ability to proactively modify interventions. Hospitals and rehabilitation facilities can improve patient satisfaction and health outcomes by optimizing treatment pathways and allocating resources effectively with the help of such predictive insights.

A diagram of a patient development

AI-generated content may be incorrect.

**Figure 1:** Patient Outcome Prediction Using Machine Learning – Flow Diagram

**Predictive Analytics in Patient Rehabilitation Planning**

Forecasting patient recovery routes and guiding clinical decision-making require predictive analytics, which is supported by machine learning techniques. Quantitative outcomes, like SIS scores, can be predicted following a predetermined rehabilitation period thanks to methods like regression analysis and ensemble learning. By identifying patients who are likely to recover quickly as well as those who might need more assistance or a different treatment strategy, this predictive capability helps medical practitioners (Jones et al., 2020).

An XGBoost model, for instance, can accurately predict a patient's Week 4 SIS score based on their Day 1 evaluations. Clinicians can step in sooner and give more intensive therapy or different approaches if the model predicts a less-than-ideal recovery. This strategy guarantees prompt, patient-specific treatment modifications, significantly increasing the success rates of rehabilitation and overall patient care experiences.

**Use of Unsupervised Learning Techniques in Clinical Data Exploration**

Even though supervised learning is the main emphasis of this project, unsupervised learning plays a significant part in the investigation of healthcare data. Without the use of predetermined labels, methods like Principal Component Analysis (PCA) and K-means clustering are useful for revealing latent structures in patient information.

Patients can be grouped into clusters using K-means clustering if their baseline scores, improvement trends, or demographic characteristics are comparable. These clusters frequently identify subpopulations that react differently to conventional rehabilitation programs, which aids doctors in more accurately customizing therapies. In contrast, PCA simplifies the identification of the primary factors affecting patient recovery without causing a substantial loss of information by reducing the dimensionality of complicated clinical datasets (Gupta & Singh, 2021).

**Challenges and Future Directions**

Even while machine learning has many benefits for predicting clinical outcomes, there are also several issues. Because healthcare datasets frequently contain extremely sensitive personal information, data privacy and security continue to be major problems. Maintaining patient trust and ensuring ethical data processing need adherence to laws like the General Data Protection Regulation (GDPR).

Furthermore, the quality of the data utilized for training has a significant impact on the effectiveness and dependability of machine learning models. Biased, inconsistent, or incomplete datasets may produce inaccurate predictions, which may lead to subpar clinical judgments. For machine learning models to remain effective over time, ongoing validation, updating, and monitoring are required.

* 1. **Challenges in Machine Learning Applications**

Applications of machine learning in healthcare present certain difficulties despite their benefits. Typical problems include:

* Data quality issues include small sample sizes, inconsistent results, and missing values.
* Interpretability: Physicians frequently hesitate to use models they don't fully comprehend.
* Bias: Historical data may have sociodemographic biases that models unintentionally pick up on
* Ethical considerations: Data ownership, permission, and patient privacy are important problems.
* This research uses interpretability approaches, ethical data handling, and transparent preparation to address these issues.

**2.3 Future Directions in Clinical Prediction**

Future studies in this area are probably going to concentrate on multimodal learning, which combines sensor data, electronic medical records (EMRs), patient-reported outcomes, and clinical test scores. Furthermore, it is anticipated that explainable AI (XAI) frameworks like LIME and SHAP values would be crucial to the implementation of clinical-grade prediction models.

**2.4 Synthesis of Current Research**

By supervised learning, explainability, and patient-specific tools, this study synthesizes the best practices of previous work. It bridges the gap between clinical usability and computational models, which is commonly mentioned in the literature on healthcare machine learning today.

# Summary of Research Progress

## Data Collection

For the current study, which focuses on employing machine learning approaches to predict patient recovery outcomes, the patient rehabilitation dataset was used. The dataset, which includes anonymized patient records, was gathered from a clinical rehabilitation database. It contains clinical evaluation results from Day 1 and Week 4 in a variety of functional and cognitive domains.

To guarantee that it is prepared for predictive modeling and analysis, the dataset has undergone extensive preprocessing. The machine learning models created for this project are trained, validated, and tested primarily using this dataset.

**Dataset Source: Internal rehabilitation clinic records (anonymized for academic use).**

A table with numbers and letters

AI-generated content may be incorrect.

**Figure 2:** Overview of Patient Rehabilitation Dataset

* + 1. **Data Background:**

Individuals participating in a structured four-week rehabilitation program have longitudinal data in the patient rehabilitation dataset. Baseline assessment results are recorded at the beginning of the program (Day 1), and reassessment results are recorded at the conclusion of the four-week period (Week 4). Included are the following assessments:

* The Berg Balance Scale (BBS) measures mobility and balance.
* Timed Up and Go (TUG): Assesses fundamental mobility abilities
* Vestibular balance is evaluated using the VR Balance Evaluation System Software (VRBESS).
* The Balance Evaluation Systems Test (BESTest) is a thorough assessment of balance.
* The Stroke Impact Scale (SIS) measures the quality of life that stroke survivors describe.
* The goal is to predict Week 4 recovery scores using Day 1 tests, with an emphasis on the SIS outcome.

**3.1.2 Attributes of Data**

Below are the primary attributes present in the dataset:

* **SL.NO:** Serial number of the patient
* **AGE:** Age of the patient
* **GENDER:** Gender of the patient (Male/Female)
* **BBS\_Day1:** Berg Balance Scale score at Day 1
* **TUG\_Day1:** Timed Up and Go test score at Day 1
* **VRBESS\_Day1:** VR Balance system score at Day 1
* **BESTest\_Day1:** BESTest score at Day 1
* **SIS\_Day1:** Stroke Impact Scale score at Day 1
* **BBS\_Week4:** Berg Balance Scale score after four weeks
* **TUG\_Week4:** Timed Up and Go test score after four weeks
* **VRBESS\_Week4:** VR Balance system score after four weeks
* **BESTest\_Week4:** BESTest score after four weeks
* **SIS\_Week4:** Stroke Impact Scale score after four weeks
* **Derived Attributes:** Improvement scores for each test and an Overall Improvement Score

## Proposed Methodology

Below is an illustration of the suggested methodology for forecasting patient recovery results using Day 1 evaluation data. Starting with data collection and continuing through model deployment and insights creation, the complete architecture adheres to a disciplined machine learning process.

A diagram of a model

AI-generated content may be incorrect.

**Figure 3:** Proposed Workflow for Predictive Modelling in Patient Recovery Analysis

Additionally, the structure of the hybrid model selection and tuning process is represented to ensure optimal predictive performance:

A diagram of a model

AI-generated content may be incorrect.

**Figure 4:** Detailed Structure of Model Selection and Hyperparameter Tuning Process

## Progress of Work

The present section highlights the progress achieved throughout the project, from initial data handling to model development.

### Data Reading

Pandas has been used to successfully read the patient rehabilitation dataset into the Python environment. To guarantee proper loading and confirm the existence of anticipated variables, preliminary previews were carried out.

A close up of text

AI-generated content may be incorrect.

**Figure 5:** Python Code for Loading Patient Dataset

* + 1. **Data Information**

Basic data inspection was performed to understand the structure of the dataset, including checking data types, identifying missing values, and summarising the statistical distribution of key variables.

**A table with numbers and a few black lines

AI-generated content may be incorrect.Figure 6:** Data Summary Statistics and Missing Value Overview

* + 1. **Data Cleaning:**

Data cleaning operations were carried out to prepare the dataset for modelling. Key steps included:

* Imputing missing numeric values with column means
* Encoding gender values into binary numeric format
* Engineering improvement scores and overall recovery indicators

A purple chart with white text

AI-generated content may be incorrect.

**Figure 7:** Summary of Data Cleaning Processes

## Feature Engineering

New features were created to better represent patient recovery, enabling the models to learn meaningful relationships:

* **Improvement Scores** for each clinical test (Week 4 score minus Day 1 score)
* **Overall Improvement Score** as the aggregate of individual improvements
* **Categorical Encoding** for Gender to make it machine-learning-ready

This step substantially enriched the dataset and added predictive power.



**Figure 8:** Feature Engineering – New Attributes Added

## Model Development Progress

The modelling phase involved building three separate machine learning models:

* **Linear Regression**: Established as the baseline model for comparison.
* **Random Forest Regressor**: Tuned with GridSearchCV to improve predictive accuracy.
* **XGBoost Regressor**: Tuned and selected as the final best-performing model.

Each model was evaluated using R² Score and RMSE, and the XGBoost model was found to deliver the highest accuracy with the lowest prediction error.

A group of colored rectangular bars

AI-generated content may be incorrect.

**Figure 9:** Model Building and Evaluation Results

**4. ISSUES RELATED TO PROJECT AND RESEARCH**

**4.1 ETHICAL ISSUES**

Below is a summary of the ethical issues raised by the current study:

* The study is wholly novel and was created by a thorough analysis of machine learning techniques and clinical rehabilitation literature.
* To ensure transparency and ethical compliance in data sourcing, the data used in this study is made up of anonymized clinical datasets made available for academic use.
* Since there was no primary data collection involving human participants, there were no worries about informed consent, privacy, or possible exploitation of personal data.
* Every data handling technique closely complies with ethical standards for the preservation, examination, and sharing of data pertaining to healthcare.
* The goals of the research are to enhance therapeutic procedures without endangering anyone, in accordance with the concepts of beneficence and non-maleficence.

**4.2 LEGAL ISSUES**

The following legal considerations have been thoroughly examined in this study:

* The study respects academic integrity and intellectual property rights by being entirely original and free of plagiarism.
* All models, analyses, and interpretations are created individually, despite being influenced by current machine learning and healthcare analytics research.
* To ensure compliance with IT governance policies, all work was completed on institutionally authorized devices or personal computing resources.
* To ensure compliance with GDPR and other relevant data protection requirements, the dataset was anonymized, legally sourced, and did not require any special permits, financial transactions, or licenses.

**4.3 PROFESSIONAL ISSUES**

The following describes the professional norms and procedures that were observed during this study:

* Under the direction of the project supervisor, the research was carried out in accordance with the professional and ethical standards required by the academic institution.
* To guarantee traceability, reproducibility, and academic transparency, thorough and organized documentation of the study procedures, findings, and conclusions has been preserved.
* To improve readability and adhere to academic writing norms, the report has been divided into logically constructed chapters.
* To enable clear comprehension and efficient distribution of study findings, professionalism in communication, formatting, and presentation has been strictly upheld.

**4.4 SOCIAL ISSUES**

The following are the societal ramifications of this study:

* By offering a predictive framework that can help physicians better plan patient rehabilitation strategies, the study hopes to improve healthcare services.
* Individual privacy and dignity are completely preserved because no human subjects were actively involved in the data collecting or research procedure.
* By using anonymized datasets, social responsibility standards pertaining to patient confidentiality and data security are upheld and no indirect disclosure of personal information takes place.
* The research aims to assist wider social goals, such as better patient outcomes, more effective resource allocation, and higher quality of care, by improving understanding of patient recovery patterns.

**5. RESEARCH PLANNING**

**5.1 SELECTION OF TOOL**

Because of its robust and vast ecosystem of libraries designed especially for data science and machine learning tasks, Python 3 was chosen as the main development environment for this project. Python offers incredibly effective tools for preprocessing and data manipulation, like Pandas and NumPy, as well as Scikit-learn and XGBoost for putting sophisticated machine learning models into practice. Additionally, Python easily connects with data visualization programs like Seaborn and Matplotlib, which are crucial for drawing conclusions and properly displaying findings. Its selection was also influenced by its open-source nature and vibrant community.

**5.2 SELECTION OF TECHNOLOGIES**

For this research, several technologies and libraries were strategically selected to manage various stages of the data analysis and machine learning pipeline:

**5.2.1 Data Manipulation and Preprocessing:**

* To handle, manipulate, and preprocess structured clinical datasets efficiently, Pandas and NumPy were used.
* The dataset was prepped for modelling using Scikit-learn's preprocessing modules, such as LabelEncoder (for categorical encoding) and SimpleImputer (for managing missing values).

**5.2.2 Advanced Machine Learning Models:**

* As a baseline model, linear regression was used to create a point of comparison for performance assessment.
* Because of its resilience in managing non-linear relationships and variable importance analysis, the Random Forest Regressor was chosen.
* The XGBoost Regressor was selected as an advanced model that may provide excellent predicted accuracy by utilizing effective regularization methods and boosting approaches.

**5.2.3 Visualization and Insights Extraction:**

To provide intricate visualisations such feature importance charts, improvement score distributions, and patient-specific recovery plots, Seaborn and Matplotlib were used. The interpretability and therapeutic usefulness of the model outputs were much enhanced by these visual aids.

**5.3 SELECTION OF ALGORITHMS**

To address various aspects of predicting patient recovery outcomes, the following algorithms and methods were implemented:

**5.3.1 Predictive Modelling:**

* Simple linear connections between Day 1 assessments and Week 4 results were modelled using linear regression.
* An ensemble learning method that caught intricate feature interactions was offered by the Random Forest Regressor.
* With hyperparameter adjustment, the XGBoost Regressor was used to maximize performance using regularization and boosting techniques.

**5.3.2 Feature Engineering and Reduction:**

* Improvement scores (Week 4 minus Day 1) were computed for each assessment metric as part of feature engineering.
* All features were clinically significant and kept, therefore dimensionality reduction methods like Principal Component Analysis (PCA) were taken into consideration but were not required.
  + 1. **Model Interpretability Techniques:**
* Techniques like **Feature Importance Analysis** were used to interpret which clinical metrics (Day 1 scores) most strongly influenced model predictions.
* This improves model transparency and supports clinicians in understanding why the model made certain predictions.

**5.3.4 Residual Analysis for Model Validation:**

* **Residual plots** (difference between actual vs predicted values) were generated to validate that model predictions were unbiased and well-fitted.
* By checking residuals, the project ensured that errors were randomly distributed, confirming that no systematic prediction errors were present.

**5.4 IMPLEMENTATION PLAN**

The project was implemented through the following structured phases:

* **Data Collection and Integration:** Load the patient rehabilitation dataset ensuring completeness and accuracy for further analysis.
* **Data Cleaning and Preprocessing:**
  + Handle missing values using imputation techniques.
  + Encode categorical variables (such as Gender).
  + Standardise and normalise numerical features as needed.
* **Feature Engineering:**
  + Create improvement score features based on the difference between Day 1 and Week 4 assessments.
  + Generate an Overall Improvement Score summarising patient progress.
* **Model Development and Training:**
  + Develop predictive models using Linear Regression, Random Forest Regressor, and XGBoost Regressor.
  + Train models using the training subset and perform hyperparameter tuning using GridSearchCV.
* **Model Evaluation and Optimisation:**
  + Evaluate models based on **R² Score** and **Root Mean Squared Error (RMSE)**.
  + Optimise model hyperparameters to enhance predictive performance and generalisability.
* **Deployment and Patient Insight Generation:**
  + Use the best-performing model (XGBoost) to generate patient-specific recovery predictions.
  + Build graphical and narrative patient summary dashboards for clinical interpretation.
* **Monitoring and Documentation:**
  + Thoroughly document all results, methodologies, and interpretations.
  + Maintain high standards of reproducibility and transparency for academic reporting.

# Project Planning

The project was carefully structured and executed across multiple phases to ensure systematic development, evaluation, and reporting. Each phase was planned based on specific deliverables and timelines to facilitate the successful completion of the research objectives.

The phases included:

* **Week 1–2:** Data understanding, cleaning, and feature engineering.
* **Week 3–4:** Model building and hyperparameter tuning.
* **Week 5:** Model evaluation, visualisation, and insights generation.
* **Week 6:** Documentation, report writing, and final presentation preparation.

The complete plan of the project has been presented below in the form of the Gantt Chart:

A graph with blue rectangles

AI-generated content may be incorrect.

**Figure 10:** Gantt Chart for Project Timeline

**7. CONCLUSION**

Predicting patient rehabilitation results, specifically the Stroke Impact Scale (SIS) Week 4 scores, using preliminary clinical evaluations documented on Day 1 was the main goal of this study. The study used a systematic data science technique to do this, starting with extensive data preparation and cleaning, then moving on to sophisticated machine learning modeling, hyperparameter tweaking, model evaluation, and interpretability analysis.

The project's goals were effectively met.

Three regression models—Linear Regression, Random Forest Regressor, and XGBoost Regressor—were created and compared with a great deal of testing. It was found that the XGBoost model performed better than the other models, with the maximum prediction accuracy and the lowest error margin, following meticulous hyperparameter adjustment and evaluation based on R2 Score and Root Mean Squared Error (RMSE).

A thorough feature importance analysis was also included in the project, and the results showed that baseline evaluations like TUG\_Day1 and BBS\_Day1 were among the most significant indicators of patient recovery. In order to facilitate more individualized and data-driven rehabilitation planning, the model outcomes' practical usefulness in clinical settings was further expanded with the creation of individual patient summary dashboards that included both visual and narrative insights.

All the initial research questions were satisfactorily addressed by this project:

* **RQ1:** Using early-stage assessment scores, machine learning models—in particular, ensemble approaches like XGBoost—can reliably forecast rehabilitation results.
* **RQ2:** It was discovered that some baseline characteristics, particularly tests of balance and mobility, were quite predictive.
* **RQ3:** XGBoost outperformed Random Forest and Linear Regression, underscoring the significance of non-linear modelling in intricate clinical datasets.
* **RQ4:** The model outputs were comprehensible and pertinent for clinical stakeholders thanks to visual and narrative summaries.

**Limitations:**

Although the project's main objectives were met, it should be noted that it had several limitations. First off, the model may not be as applicable to larger clinical populations due to the relatively limited dataset size (50 patient records). Second, only organized clinical assessments were employed; modeling was not possible for other potentially significant variables such drug information, comorbidities, or socioeconomic background. Thirdly, the study only looked at regression modeling; it didn't investigate other methods that could provide more information about recovery paths, such survival analysis or time-series modeling.

**Future Work:**

To increase the generalizability of the model, future studies could build on our work by combining bigger and more varied datasets. Predictive performance may also be improved by using unstructured data sources, such as patient-reported outcomes or physician notes. To further enhance model interpretability and trust in clinical settings, future research should also test explainable AI (XAI) frameworks such SHAP values, ensemble stacking strategies, or deep learning models.

To sum up, our experiment shows how machine learning methods, especially ensemble-based methods, have a great deal of promise for forecasting patient recovery outcomes. The study makes a significant contribution to improving the standard of patient care and customizing therapeutic approaches by facilitating data-driven decision-making in rehabilitation settings.

**References:**

[1] Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJ, Culebras A, Elkind MS, George MG, Hamdan AD, Higashida RT, Hoh BL, Janis LS, Kase CS, Kleindorfer DO, Lee JM, Moseley ME, Peterson ED, Turan TN, Valderrama AL, Vinters HV; American Heart Association Stroke Council, Council on Cardiovascular Surgery and Anesthesia; Council on Cardiovascular Radiology and Intervention; Council on Cardiovascular and Stroke Nursing; Council on Epidemiology and Prevention; Council on Peripheral Vascular Disease; Council on Nutrition, Physical Activity and Metabolism. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2013 Jul;44(7):2064-89. doi: 10.1161/STR.0b013e318296aeca. Epub 2013 May 7. Erratum in: Stroke. 2019 Aug;50(8):e239. PMID: 23652265. (Accessed on April 2022).Foster, J., 2019. Natural

[2] Komlodi-Pasztor E, Gilbert MR, Armstrong TS. Diagnosis and Management of Stroke in Adults with Primary Brain Tumor. Curr Oncol Rep. 2022 Oct;24(10):1251-1259. doi: 10.1007/s11912-022-01280-6. Epub 2022 May 11. PMID: 35543860; PMCID: PMC9474382. (Accessed on May 2022).

[3] Coupland AP, Thapar A, Qureshi MI, Jenkins H, Davies AH. The definition of stroke. J R Soc Med. 2017 Jan;110(1):9-12. doi: 10.1177/0141076816680121. Epub 2017 Jan 13. PMID: 28084167; PMCID: PMC5298424. (Accessed on April 2023).

[4] Truelsen, Thomas & Begg, Stephen. (2006). The Global Burden of Cerebrovascular Disease. World Health Organization. https://www.researchgate.net/publication/228551377\_The\_Global\_Burden\_of\_Cerebrovascular\_Disease ( Accessed on April 2022).

[5] Kamalakannan S, Gudlavalleti ASV, Gudlavalleti VSM, Goenka S, Kuper H. Incidence & prevalence of stroke in India: A systematic review. Indian J Med Res. 2017 Aug;146(2):175-185. doi: 10.4103/ijmr.IJMR\_516\_15. PMID: 29265018; PMCID: PMC5761027. (Accessed on May 2022).

[6] Avan A, Digaleh H, Di Napoli M, Stranges S, Behrouz R, Shojaeianbabaei G, Amiri A, Tabrizi R, Mokhber N, Spence JD, Azarpazhooh MR. Socioeconomic status and stroke incidence, prevalence, mortality, and worldwide burden: an ecological analysis from the Global Burden of Disease Study 2017. BMC Med. 2019 Oct 24;17(1):191. doi: 10.1186/s12916-019-1397-3. PMID: 31647003; PMCID: PMC6813111. ( Accessed on June 2022)

[7] Kuriakose D, Xiao Z. Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives. Int J Mol Sci. 2020 Oct 15;21(20):7609. doi: 10.3390/ijms21207609. PMID: 33076218; PMCID: PMC7589849. ( Accessed on February 2023).

[8] Martin S, Kessler M, Neurological Intervention for physical therapy, 4th ed. Evansville, 2021. Chapter 11 stroke pg n. 348-349.( Accessed on January 2023).

[9] Smajlović D. Strokes in young adults: epidemiology and prevention. Vasc Health Risk Manag. 2015 Feb 24;11:157-64. doi: 10.2147/VHRM.S53203. PMID: 25750539; PMCID: PMC4348138. (Accessed on March 2023).

[10] Patten C, Lexell J, Brown HE. Weakness and strength training in persons with poststroke hemiplegia: rationale, method, and efficacy. J Rehabil Res Dev. 2004 May;41(3A):293-312. doi: 10.1682/jrrd.2004.03.0293. PMID: 15543447. (Accessed on May 2022).

[11] Choi HL, Yang K, Han K, Kim B, Chang WH, Kwon S, Jung W, Yoo JE, Jeon HJ, Shin DW. Increased Risk of Developing Depression in Disability after Stroke: A Korean Nationwide Study. Int J Environ Res Public Health. 2023 Jan 2;20(1):842. doi: 10.3390/ijerph20010842. PMID: 36613164; PMCID: PMC9819798. (Accessed on January 2023).

[12] P S Prevalence of stroke in k’taka higher than national average Deccan Herald. DH News Service; 2021 [cited 2022 Aug1]. Available from: https://www.deccanherald.com/state/top-karnataka-stories/prevalenceof-stroke-in-ktaka-higher-than-national-average-1008770.html (Accessed on August 2022).

[13] Won SH, Kim JC, Oh DW. Effects of a novel walking training program with postural correction and visual feedback on walking function in patients with post-stroke hemiparesis. J Phys Ther Sci. 2015 Aug;27(8):2581-3. doi: 10.1589/jpts.27.2581. Epub 2015 Aug 21. PMID: 26357443; PMCID: PMC4563319. (Accessed on January 2023).

[14] Navarro-Orozco D, Sánchez-Manso JC. Neuroanatomy, Middle Cerebral Artery. 2022 Jul 25. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan–. PMID: 30252258. (Accessed on January 2023).

[15] Bernhardt J, Hayward KS, Kwakkel G, Ward NS, Wolf SL, Borschmann K, Krakauer JW, Boyd LA, Carmichael ST, Corbett D, Cramer SC. Agreed definitions and a shared vision for new standards in stroke recovery research: The Stroke Recovery and Rehabilitation Roundtable taskforce. Int J Stroke. 2017 Jul;12(5):444-450. doi: 10.1177/1747493017711816. PMID: 28697708. (Accessed on September 2022).

[16] Sheila L, Gita R, Verheyden G Physical management for neurological conditions Fourth edition Publisher, Elsevier Health Sciences, 2018 ; ISBN, 0702077232, 9780702077234. (Accessed on December 2022).

[17] Woodruff TM, Thundyil J, Tang SC, Sobey CG, Taylor SM, Arumugam TV. Pathophysiology, treatment, and animal and cellular models of human ischemic stroke. Mol Neurodegener. 2011 Jan 25;6(1):11. doi: 10.1186/1750-1326-6-11. PMID: 21266064; PMCID: PMC3037909. (Accessed on January 2023). International Journal of Membrane Science and Technology, 2023, Vol. 10, No. 2, pp 3001-3015 3013

[18] Murphy SJ, Werring DJ. Stroke: causes and clinical features. Medicine (Abingdon). 2020 Sep;48(9):561- 566. doi: 10.1016/j.mpmed.2020.06.002. Epub 2020 Aug 6. PMID: 32837228; PMCID: PMC7409792. (Accessed on June 2023).

[19] Gund BM, Jagpat PN, Ingale VB, Patil R Y-. Stroke: A Brain Attack. IOSR journal of Pharm. 2013 Sept;03(08):01–23. (Accessed on April 2022).

[20] Li J, Zhong D, Ye J, He M, Liu X, Zheng H, Jin R, Zhang SL. Rehabilitation for balance impairment in patients after stroke: a protocol of a systematic review and network meta-analysis. BMJ Open. 2019 Jul 19;9(7):e026844. doi: 10.1136/bmjopen-2018-026844. PMID: 31326927; PMCID: PMC6661695. (Accessed on May 2022).

[21] Lee K. Speed-Interactive Pedaling Training Using Smartphone Virtual Reality Application for Stroke Patients: Single-Blinded, Randomized Clinical Trial. Brain Sci. 2019 Oct 27;9(11):295. doi: 10.3390/brainsci9110295. PMID: 31717888; PMCID: PMC6895905. (Accessed on April 2022).

[22] Liu C, McNitt-Gray JL, Finley JM. Impairments in the mechanical effectiveness of reactive balance control strategies during walking in people post-stroke. Front Neurol. 2022 Oct 31;13:1032417. doi: 10.3389/fneur.2022.1032417. PMID: 36388197; PMCID: PMC9659909. (Accessed on March 2023).

[23] Mansfield A, Aqui A, Centen A, Danells CJ, DePaul VG, Knorr S, Schinkel-Ivy A, Brooks D, Inness EL, McIlroy WE, Mochizuki G. Perturbation training to promote safe independent mobility post-stroke: study protocol for a randomized controlled trial. BMC Neurol. 2015 Jun 6;15:87. doi: 10.1186/s12883-015-0347- 8. PMID: 26048054; PMCID: PMC4456796. (Accessed on August 2022).

[24] Kumar C, Pathan N Effectiveness of Manual Perturbation Exercises in Improving Balance, Function and Mobility in Stroke Patients: A Randomized Controlled Trial. J Nov Physiother 6: 284. 2016; 6(2): 1 - 7 doi:10.4172/2165-7025.1000284. (Accessed on April 2022).

[25] Inness EL, Mansfield A, Lakhani B, Bayley M, McIlroy WE. Impaired reactive stepping among patients ready for discharge from inpatient stroke rehabilitation. Phys Ther. 2014 Dec;94(12):1755-64. doi: 10.2522/ptj.20130603. Epub 2014 Aug 7. PMID: 25104795; PMCID: PMC4263904. (Accessed on June 2023).

[26] Handelzalts S, Kenner-Furman M, Gray G, Soroker N, Shani G, Melzer I. Effects of Perturbation-Based Balance Training in Subacute Persons with Stroke: A Randomized Controlled Trial. Neurorehabil Neural Repair. 2019 Mar;33(3):213-224. doi: 10.1177/1545968319829453. Epub 2019 Feb 15. PMID: 30767613. (Accessed on May 2022).

[27] Ooha K & Thulasi, P & Apparao P & Prakash P & Ganapathi Swamy C. Effectiveness of mobile based virtual reality training on improving balance in elderly population. European journal of pharmaceutical and medical research. 2021; 8(12). 369-378. (Accessed on August 2022).

[28] Marques-Sule E, Arnal-Gómez A, Buitrago-Jiménez G, Suso-Martí L, Cuenca-Martínez F, Espí-López GV. Effectiveness of Nintendo Wii and Physical Therapy in Functionality, Balance, and Daily Activities in Chronic Stroke Patients. J Am Med Dir Assoc. 2021 May;22(5):1073-1080. doi: 10.1016/j.jamda.2021.01.076. Epub 2021 Feb 24. PMID: 33639116. (Accessed on May 2022).

[29] Huygelier H, Schraepen B, van Ee R, Vanden Abeele V, Gillebert CR. Acceptance of immersive headmounted virtual reality in older adults. Sci Rep. 2019 Mar 14;9(1):4519. doi: 10.1038/s41598-019-41200-6. PMID: 30872760; PMCID: PMC6418153.

[30] Hao J, Xie H, Harp K, Chen Z, Siu KC. Effects of Virtual Reality Intervention on Neural Plasticity in Stroke Rehabilitation: A Systematic Review. Arch Phys Med Rehabil. 2022 Mar;103(3):523-541. doi: 10.1016/j.apmr.2021.06.024. Epub 2021 Aug 2. PMID: 34352269. (Accessed on March 2022).

[31] Li Z, Han XG, Sheng J, Ma SJ. Virtual reality for improving balance in patients after stroke: A systematic review and meta-analysis. Clin Rehabil. 2016 May;30(5):432-40. doi: 10.1177/0269215515593611. Epub 2015 Jul 3. PMID: 26141808. (Accessed on May 2022)

[32] Kamińska MS, Miller A, Rotter I, Szylińska A, Grochans E. The effectiveness of virtual reality training in reducing the risk of falls among elderly people. Clin Interv Aging. 2018 Nov 14;13:2329-2338. doi: 10.2147/CIA.S183502. PMID: 30532523; PMCID: PMC6241865. (Accessed on November 2022).

[33] Truijen S, Abdullahi A, Bijsterbosch D, van Zoest E, Conijn M, Wang Y, Struyf N, Saeys W. Effect of homebased virtual reality training and telerehabilitation on balance in individuals with Parkinson disease, multiple sclerosis, and stroke: a systematic review and meta-analysis. Neurol Sci. 2022 May;43(5):2995-3006. doi: International Journal of Membrane Science and Technology, 2023, Vol. 10, No. 2, pp 3001-3015 3014 10.1007/s10072-021-05855-2. Epub 2022 Feb 17. PMID: 35175439; PMCID: PMC9023738. (Accessed on May 2022).

[34] Gutiérrez RO, Galán Del Río F, Cano de la Cuerda R, Alguacil Diego IM, González RA, Page JC. A telerehabilitation program by virtual reality-video games improves balance and postural control in multiple sclerosis patients. NeuroRehabilitation. 2013;33(4):545-54. doi: 10.3233/NRE-130995. PMID: 24029009. (Accessed on September 2022).

[35] Pintado-Izquierdo S, Cano-de-la-Cuerda R, Ortiz-Gutiérrez RM. Video Game-Based Therapy on Balance and Gait of Patients with Stroke: A Systematic Review. Applied Sciences. 2020; 10(18):6426. https://doi.org/10.3390/app10186426. (Accessed on July 2022).

[36] Shin DC, Song CH. Smartphone-Based Visual Feedback Trunk Control Training Using a Gyroscope and Mirroring Technology for Stroke Patients: Single-blinded, Randomized Clinical Trial of Efficacy and Feasibility. Am J Phys Med Rehabil. 2016 May;95(5):319-29. doi: 10.1097/PHM.0000000000000447. PMID: 26829087. (Accessed on August 2022).

[37] Lee MM, Lee KJ, Song CH. Game-Based Virtual Reality Canoe Paddling Training to Improve Postural Balance and Upper Extremity Function: A Preliminary Randomized Controlled Study of 30 Patients with Subacute Stroke. Med Sci Monit. 2018 Apr 27;24:2590-2598. doi: 10.12659/MSM.906451. PMID: 29702630; PMCID: PMC5944399.(Accessed on May 2023).

[38] Lin RC, Chiang SL, Heitkemper MM, Weng SM, Lin CF, Yang FC, Lin CH. Effectiveness of Early Rehabilitation Combined With Virtual Reality Training on Muscle Strength, Mood State, and Functional Status in Patients With Acute Stroke: A Randomized Controlled Trial. Worldviews Evid Based Nurs. 2020 Apr;17(2):158-167. doi: 10.1111/wvn.12429. Epub 2020 Mar 25. PMID: 32212254.(Accessed on January 2023).

[39] Iqbal M, Arsh A, Hammad SM, Haq IU, Darain H. Comparison of dual task specific training and conventional physical therapy in ambulation of hemiplegic stroke patients: A randomized controlled trial. J Pak Med Assoc. 2020 Jan;70(1):7-10. doi: 10.47391/JPMA.10443. PMID: 31954014. (Accessed on March 2023).

[40] Suda S, Muraga K, Ishiwata A, Nishimura T, Aoki J, Kanamaru T, Suzuki K, et.al;.Early Cognitive Assessment Following Acute Stroke: Feasibility and Comparison between Mini-Mental State Examination and Montreal Cognitive Assessment. J Stroke Cerebrovasc Dis. 2020 Apr;29(4):104688. doi: 10.1016/j.jstrokecerebrovasdis.2020.104688. Epub 2020 Feb 14. PMID: 32063455.(Accessed on November 2022).

[41] Dautzenberg G, Lijmer J, Beekman A. Clinical value of the Montreal Cognitive Assessment (MoCA) in patients suspected of cognitive impairment in old age psychiatry. Using the MoCA for triaging to a memory clinic. Cogn Neuropsychiatry. 2021 Jan;26(1):1-17. doi: 10.1080/13546805.2020.1850434. Epub 2020 Dec 3. PMID: 33272076. (Accessed on January 2023). [42] Blum L, Korner-Bitensky N. Usefulness of the Berg Balance Scale in stroke rehabilitation: a systematic review. Phys Ther. 2008 May;88(5):559-66. doi: 10.2522/ptj.20070205. Epub 2008 Feb 21. PMID: 18292215. (Accessed on July 2022). [43] Miranda-Cantellops N, Tiu TK. Berg Balance Testing. 2023 Feb 17. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan–. PMID: 34662032. (Accessed on February 2023).

[44] Hafsteinsdóttir TB, Rensink M, Schuurmans M. Clinimetric properties of the Timed Up and Go Test for patients with stroke: a systematic review. Top Stroke Rehabil. 2014 May-Jun;21(3):197-210. doi: 10.1310/tsr2103-197. PMID: 24985387. (Accessed on June 2022).

[45] Rausch M, Simon JE, Starkey C, Grooms DR. Smartphone virtual reality to increase clinical balance assessment responsiveness. Phys Ther Sport. 2018 Jul;32:207-211. doi: 10.1016/j.ptsp.2018.05.017. Epub 2018 May 22. PMID: 29803943. (Accessed on May 2022).

[46] Chinsongkram B, Chaikeeree N, Saengsirisuwan V, Horak FB, Boonsinsukh R. Responsiveness of the Balance Evaluation Systems Test (BESTest) in People With Subacute Stroke. Phys Ther. 2016 Oct;96(10):1638-1647. doi: 10.2522/ptj.20150621. Epub 2016 Apr 21. PMID: 27103226; PMCID: PMC6281077. (Accessed on October 2022).

[47] Sahin IE, Guclu-Gunduz A, Yazici G, Ozkul C, Volkan-Yazici M, Nazliel B, Tekindal MA. The sensitivity and specificity of the balance evaluation systems test-BESTest in determining risk of fall in stroke patients. International Journal of Membrane Science and Technology, 2023, Vol. 10, No. 2, pp 3001-3015 3015 NeuroRehabilitation. 2019;44(1):67-77. doi: 10.3233/NRE-182558. PMID: 30814369.(Accessed on October 2022).

[48] Almalki O, Alshehri MA, El-Fiky AA, Abdelaal AA, Alzaidi JH, Al Attar WSA, Hegazy FA. Can the stroke impact scale 3.0 detect cognitive impairments in patients with a recent stroke? J Phys Ther Sci. 2019 Jul;31(7):563-568. doi: 10.1589/jpts.31.563. Epub 2019 Jul 9. PMID: 31417222; PMCID: PMC6642892. (Accessed on July 2022).

[49] Vellone E, Savini S, Fida R, Dickson VV, Melkus GD, Carod-Artal FJ, Rocco G, Alvaro R. Psychometric evaluation of the Stroke Impact Scale 3.0. J Cardiovasc Nurs. 2015 May-Jun;30(3):229-41. doi: 10.1097/JCN.0000000000000145. PMID: 24695074. (Accessed on May 2023).

[50] Prakash V, Ganesan M, Vasanthan R, Hariohm K. Do Commonly Used Functional Outcome Measures Capture Activities that Are Relevant for People with Stroke in India? Top Stroke Rehabil. 2017 Apr;24(3):200-205. doi: 10.1080/10749357.2016.1234190. Epub 2016 Sep 27. PMID: 27671020. (Accessed on June 2023).

[51] Cheng YS, Chien A, Lai DM, Lee YY, Cheng CH, Wang SF, Chang YJ, Wang JL, Hsu WL. PerturbationBased Balance Training in Postoperative Individuals With Degenerative Cervical Myelopathy. Front Bioeng Biotechnol. 2020 Feb 20;8:108. doi: 10.3389/fbioe.2020.00108. PMID: 32154235; PMCID: PMC7044125. (Accessed on May 2023).

[52] Iruthayarajah J, McIntyre A, Cotoi A, Macaluso S, Teasell R. The use of virtual reality for balance among individuals with chronic stroke: a systematic review and meta-analysis. Top Stroke Rehabil. 2017 Jan;24(1):68-79. doi: 10.1080/10749357.2016.1192361. Epub 2016 Jun 16. PMID: 27309680. (Accessed on June 2023).

[53] Fritz SL, Peters DM, Merlo AM, Donley J. Active video-gaming effects on balance and mobility in individuals with chronic stroke: a randomized controlled trial. Top Stroke Rehabil. 2013 May-Jun;20(3):218-25. doi: 10.1310/tsr2003-218. PMID: 23841969. (Accessed on July 2023).

**Appendix:**

# Import necessary libraries

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import LabelEncoder

from sklearn.impute import SimpleImputer

from sklearn.linear\_model import LinearRegression

from sklearn.ensemble import RandomForestRegressor

from xgboost import XGBRegressor

from sklearn.model\_selection import GridSearchCV

from sklearn.metrics import r2\_score, mean\_squared\_error

import warnings

warnings.filterwarnings("ignore")

# Load the dataset

data\_path = "Patient Master Chart.xlsx" # Make sure the file is in the working directory

df = pd.read\_excel(data\_path)

# Preview the data

print("Data preview:")

print(df.head())

# Data Preprocessing

# Encoding Gender

df['GENDER'] = df['GENDER'].map({'Male': 1, 'Female': 0})

# Handling missing values

for col in df.columns:

if df[col].dtype == 'object':

df[col].fillna(df[col].mode()[0], inplace=True)

else:

df[col].fillna(df[col].mean(), inplace=True)

# Creating Improvement Features

df['BBS\_Improvement'] = df['BBS\_Week4'] - df['BBS\_Day1']

df['TUG\_Improvement'] = df['TUG\_Day1'] - df['TUG\_Week4']

df['VRBESS\_Improvement'] = df['VRBESS\_Week4'] - df['VRBESS\_Day1']

df['BESTest\_Improvement'] = df['BESTest\_Week4'] - df['BESTest\_Day1']

df['SIS\_Improvement'] = df['SIS\_Week4'] - df['SIS\_Day1']

df['Overall\_Improvement'] = (

df['BBS\_Improvement'] +

df['TUG\_Improvement'] +

df['VRBESS\_Improvement'] +

df['BESTest\_Improvement'] +

df['SIS\_Improvement']

)

# Display updated DataFrame

print("Data after preprocessing and feature engineering:")

print(df.head())

# Define Features and Target Variable

X = df[['AGE', 'GENDER', 'BBS\_Day1', 'TUG\_Day1', 'VRBESS\_Day1', 'BESTest\_Day1', 'SIS\_Day1']]

y = df['SIS\_Week4']

# Train-Test Split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Model Training: XGBoost

xgb = XGBRegressor(random\_state=42)

xgb\_params = {

'n\_estimators': [50, 100, 150],

'max\_depth': [3, 5, 7],

'learning\_rate': [0.01, 0.1, 0.2]

}

grid\_xgb = GridSearchCV(xgb, param\_grid=xgb\_params, cv=5, scoring='r2', n\_jobs=-1)

grid\_xgb.fit(X\_train, y\_train)

# Predictions

y\_pred = grid\_xgb.predict(X\_test)

# Model Evaluation

r2 = r2\_score(y\_test, y\_pred)

rmse = mean\_squared\_error(y\_test, y\_pred, squared=False)

print(f"Best Model R² Score: {r2:.4f}")

print(f"Best Model RMSE: {rmse:.4f}")

print(f"Best Hyperparameters: {grid\_xgb.best\_params\_}")

# Visualizing Feature Importances

importances = grid\_xgb.best\_estimator\_.feature\_importances\_

feat\_df = pd.DataFrame({

'Feature': X.columns,

'Importance': importances

}).sort\_values(by='Importance', ascending=True)

plt.figure(figsize=(10, 6))

sns.barplot(x='Importance', y='Feature', data=feat\_df)

plt.title("Feature Importances from XGBoost Model")

plt.tight\_layout()

plt.savefig("appendix\_feature\_importances.png")

plt.show()