

**Faculty of Natural and
Mathematical Sciences**
Department of Informatics

King's College London
Strand Campus, London,
United Kingdom



7CCSMPRJ

Individual Project Submission 2024 - 2025

Name: Siddharth Kishor Samarth
Student Number: K24012370
Degree Programme: MSc. Advanced Computing
Project Title: GluCORRECT - Harnessing Artificial Intelligence to
scrutinize Hypoglycemia in hospitalised patients with
diabetes to classify, anticipate and analyse hypoglycemic
episodes [Knowledge Exchange Project with NHS England]
Supervisor: Dr. Rita Borgo
Word Count: ===== Word count goes here =====

RELEASE OF PROJECT

Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

- ☒ I agree to the release of my project
☐ I do not agree to the release of my project

Signature:

A handwritten signature in blue ink, appearing to be "Sks", written over a horizontal line.

Date: August 6, 2025



Department of Informatics
King's College London
United Kingdom

7CCSMPRJ Individual Project

**GluCORRECT - Harnessing Artificial
Intelligence to scrutinize Hypoglycemia in
hospitalised patients with diabetes to classify,
anticipate and analyse hypoglycemic episodes
[Knowledge Exchange Project with NHS
England]**

Name: **Siddharth Kishor Samarth**
Student Number: K24012370
Course: MSc. Advanced Computing

Supervisor: Dr. Rita Borgo

This dissertation is submitted for the degree of MSc Advanced Computing.

Acknowledgements

I would like to express my sincerest gratitude towards my project supervisor, Dr. Rita Borgo, for her invaluable advice and consistent direction throughout the course of this project. Her mentorship and ideas have been instrumental in shaping the development of this work, leading to its successful completion.

I am also deeply thankful & appreciative of my industry advisor, Dr. Piya Sen Gupta, for providing the dataset that has served as the foundation of this work. Her contributions have significantly enhanced the practical relevance and quality of this project.

Ultimately I would like to thank my friends and my parents, especially my dad, without whose sacrifices I would not be where I am today.

Abstract

Project Variant: Variant 4 - Develop a weighted score and design score to predict risk of a hypoglycaemic episode before it occurs.

It is well known that hypoglycemia as well as hyperglycemia are common adverse events in patients who receive blood sugar control medication, and they are also one of the most frequently cited causes of hospital admissions in people with diabetes. National quality improvement programmes from the Healthcare Quality Improvement Partnership (HQIP) and the study of ambulance call-out data have shown that ***lack of awareness*** by both affected individuals and their attendants is associated with a dramatically increased rate of complications, amongst other factors. Guy's & St. Thomas' NHS Foundation Trust (hereafter referred to as GSTT) has found, after careful deliberation and departmental review, that hypoglycemic episodes have been occurring with unusual frequency. The Trust now seeks to take measures to resolve such problems with a greater focus on prevention combined with early corrective action. This research project has been undertaken in close collaboration with GSTT, one of the largest NHS trusts in the UK and an indispensable element of London's healthcare system, with almost 24000 staff across 5 major hospitals, handling over 3 million patients a year and generating an annual turnover of over £3 billion.

This analytical study serves as a foundation and proof-of-concept to aid GSTT in pre-emptively reducing hypoglycemia within hospitalised inpatients, by utilising statistics & machine learning techniques. Through exploratory data analysis I draw out relevant conclusions about the dataset around patient age, ethnicity and

I go on to identify the significant factors responsible for hypoglycemia within the dataset provided by the industry advisor from GSTT through exploratory data analysis, while also . I explore how they can be utilized to devise a risk score, to classify patients based on their risk of hypoglycemia.

RESULTS: HbA1c values identified as risky - 56 or so eGFR ethnicity major is glucose value

In conclusion, exhibit my findings with potential ways of applying them in practise in hospitals.

All abbreviations and symbols used in the report must be listed and defined in alphabetic order.

Nomenclature

| | |
|-------------------|---|
| AUROC | Area Under Receiver Operating Characteristic Curve - a measure of model performance |
| GSTT | Guy's and St Thomas' NHS Foundation Trust |
| HQIP | Healthcare Quality Improvement Partnership |
| "Hypo" or "Hypos" | Hypoglycemic episode(s) |
| "Inpatient" | Referring to the the fact that a patient is required to stay overnight in order to be treated (in case of surgeries or long term observation for example) |
| NCAPOP | National Clinical Audit & Patient Outcomes Programme |
| NDA | National Database Audit |
| NDISA | National Diabetes Inpatient Safety Audit |
| NHS | The publicly funded healthcare system of the United Kingdom, the National Health Service. |

| | |
|-----|---|
| a | The number of angels per unit area |
| A | The area of the needle point |
| c | Speed of light in a vacuum inertial frame |
| h | Planck constant |
| LMI | Linear Matrix Inequalities |
| N | The number of angels per needle point |

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Clinical Overview | 1 |
| 1.2 | Background | 1 |
| 1.3 | Aims and Objectives | 3 |
| 1.4 | Report Structure | 3 |
| 1.4.1 | Dissertation Length | 4 |
| 2 | Literature Survey & Review | 5 |
| 2.1 | The Diabetes Management Tightrope | 5 |
| 2.2 | Methods Of Monitoring Blood Glucose Over Time | 5 |
| 2.3 | Review Of Relevant Literature and Similar Research | 6 |
| 2.4 | A Novel Contribution to ML-based Glycaemia management in the NHS . | 7 |
| 3 | Dataset | 8 |
| 3.1 | Raw Data | 8 |
| 3.2 | Cleaned Dataset | 9 |
| 4 | Methodology and Implementation | 11 |
| 4.1 | Research Strategy and Approach | 11 |
| 4.2 | Dealing With Imbalanced Data | 13 |
| 4.3 | Machine Learning Theories | 13 |
| 4.3.1 | Decision Trees | 14 |
| 4.3.2 | Random Forest | 15 |
| 4.3.3 | GridSearchCV | 16 |
| 5 | Main Results and Findings | 17 |
| 5.1 | Interpretation Analysis and Evaluation | 17 |
| 6 | Math equations | 17 |
| 6.1 | Maths | 17 |
| 6.2 | Figures | 17 |
| 7 | Legal, Social, Ethical and Professional Issues | 19 |
| 8 | Conclusion | 20 |
| | References | 21 |
| A | Appendix | 23 |
| A.1 | Dataset | 23 |

List of Figures

| | | |
|---|--|----|
| 1 | Data preprocessing workflow | 12 |
| 2 | This is the caption for the figure. | 17 |
| 3 | This is the caption for the figure which is not even present. | 18 |
| 4 | Another caption | 18 |
| 5 | Raw dataset | 23 |
| 6 | Dataset with cleaned features (this is in addition to the fields of the raw dataset) | 24 |

List of Tables

1 Introduction

1.1 Clinical Overview

Hypoglycaemia (also known as a “hypoglycaemic episode” or a “hypo” for short) is the condition that occurs when the human body’s blood glucose (sugar) level drops below the normal healthy range of 4.0 to 6.0 mmol/L. While it can affect anyone, it is most common in diabetic individuals who are prescribed drugs like insulin or metformin to inhibit glucose. Hypoglycaemic events are relatively simple and straightforward to resolve, but they need to be treated immediately to avoid serious damage to the brain and heart as a result of loss of consciousness or arrhythmias. High-sugar consumables are generally effective in correcting mild cases and are commonly recommended for immediate treatment, but severe cases of hypoglycaemia such as when the person is unconscious or having a seizure can only be resolved with an urgent, immediate glucagon injection to prevent them from deteriorating into a coma (or in rare cases, even leading to death).

To underscore how and why this matters, diabetes is one of the most significant and expensive long-term health conditions faced by the NHS, with recent figures from Diabetes UK suggesting that over 5.8 million people in the UK are living with diabetes, regardless of a formal diagnosis. It is estimated to cost the NHS over £10.7 billion a year, approximately 10% of its entire annual budget, which could go up to £18 billion by 2035 [1]. A stark finding is that almost 60% of this cost (around £6.2 billion) is spent on treating the largely preventable complications of diabetes, such as heart attacks, strokes, blindness, and so on, including hypoglycaemia [2]. Hypoglycaemic instances make up a major component of these preventable costs, mainly accounting for the emergency, ambulance, and acute care expenses associated with diabetes. The Local Impact of Hypoglycaemia Tool (LIHT) suggests that hypoglycaemia can cost up to £2,195 per episode, possibly increasing substantially with a longer stay in hospital [3], and it is estimated that there are up to 100,000 ambulance callouts annually according to the Diabetes Research and Wellness Foundation (DRWF) [4]. DRWF’s study hinted that 1 in 10 individuals that experience a severe hypo (meaning requiring medical intervention or resuscitation) have considerable chances of another one within a fortnight.

1.2 Background

After introspective analysis supported by information from the National Diabetes Inpatient Safety Audit (NDISA) it has been recognized that severe hypoglycaemia and recurrent severe hypoglycaemia have been occurring relatively frequently across GSTT medical facilities. The NDISA forms part of the National Diabetes Audit (NDA), and it maintains that “The prevalence of diabetes continues to increase. In England

between 2017-18 and 2021-22 prevalence of type 1 diabetes went up from 248,240 to 270,935 and the prevalence of type 2 and other diabetes from 2,952,695 to 3,336,980”, as of 2022 [5].

GSTT administers upwards of 500,000 point-of-care glucose tests (POCT) annually, in addition to kidney function and glycated haemoglobin (HbA1c) tests as well. The Trust also possesses blood glucose / ketone data with additional linked data including demographics, dates of admission and discharge, patient as well as family history and current or previous medications. They have two major kinds of patient records, inpatient records for patients that have to stay over the course of one or multiple nights (for example, in case of surgeries or for long term care), and outpatient records where the patient doesn’t require overnight stay. The Trust manages all of this data through their electronic health record management system called Epic, and facilitates patient access to their own records through the MyChart web application.

Hypoglycaemia is a frequent complication amongst inpatients having complex health conditions, especially within those in intensive care settings that have been / are critically ill due to advanced diseases or comorbidities, or in patients following major surgical interventions. The Trust is undertaking proactive measures to identify and mitigate the risk of hypoglycaemic episodes at an early stage, to support better planning, reduce healthcare costs, efficiently allocate hospital resources and also schedule operations optimally. The ideal way to assess risk would along the lines of developing tools to predict individualized risk scores for inpatients after considering all relevant factors. However, this presents a herculean task due to the sheer volume and complexity of factors involved, compounded by the challenges of producing reliable results even within small populations — such as those in remote areas — while also adhering to legal and governmental regulations:

1. Weighing up the risk of hypoglycaemia depends upon numerous aspects such as lifestyle, renal function, recent food intake, blood glucose history and current medication to name a few, making this a highly complicated modelling problem. In addition to this, patients differ widely in age, comorbidities, ethnic factors and even insulin sensitivity. This variability makes it a formidable challenge to develop a model that is generalizable, dependable and unbiased.
2. Any such analytical tool in the vicinity of patient healthcare requires medical evaluation and approval, validation trials, governance oversight as well as ethical considerations. Even a good model may fail if it does not fit the clinical workflow. Initial skepticism towards AI, the effort required to train staff, defining clear responsibilities and limits of liability, and rehearsing procedures or plans of action for every possible scenario will all produce appreciable organizational inertia.

Successfully implementing even a small-scale solution, within GSTT to begin with, would be a significant strategic breakthrough that serves as a foundational model which other

NHS trusts or institutions could adapt and build upon. This positions this research initiative which is a Knowledge Exchange Project (KEP) with Guy's & St.Thomas' NHS Foundation Trust, an indispensable constituent of London's healthcare system, as a valuable and worthwhile research endeavour.

1.3 Aims and Objectives

This research project has the following objectives:

- **To extract insights from provided dataset for the given time period and population.** GSTT has expressed a strong interest towards gaining a deeper understanding of their inpatient population. The dataset they have provided includes demographic details, length of hospital stay, and ward information in addition to the main clinically relevant variables such as glycated haemoglobin levels, renal function measurements, patient age and so on. This enables a comprehensive, multifaceted analysis. The knowledge gained from this study, such as identifying which hospital wards have more vulnerable or at-risk patients, will be used to enhance staff training, in turn improving both future admissions routines as well as post-discharge support for patients. Every observation, regardless of scale, holds potential to refine hospital processes and operating procedures.
- **To identify the main influencing / contributing factors for hypoglycaemia and develop a weighted risk score to predict episodes (Variant 4 KEP).** The Trust is establishing and implementing measures to "pre-assess" inpatients to evaluate their risk of a hypoglycaemic episode, which will allow medical professionals to design protocols and policies to prevent episodes from occurring as well as take early remediative action as soon as possible to resolve an episode should it occur. I aim to find data-backed values for the key features responsible for hypoglycaemia, through statistical tests and machine learning algorithms, in order to formulate a risk score. This risk score can then be applied in hospital to determine the best course of early action or precautions to take based on the patient's reason for being admitted.

1.4 Report Structure

Section 2 contains a comprehensive, detailed review of similar research carried out by other universities, teaching hospitals and medical facilities including references to relevant medical literature. I have compared and contrasted datasets used, approaches taken and results obtained.

Section 3 delves deeper into the dataset provided by GSTT, elaborating on the raw features provided and those that were derived from them for analysis.

Section 4 (Methodology & Implementation) outlines the statistical and mathematical theory behind the concepts used for analysis, ranging from machine learning algorithms to hypothesis testing methods.

Sections 5 (Main Results) onwards discuss the main research executed within the project and deliberates on the results achieved

Section 6 (Ethical Professional Legal Social issues)

Section 7 (Conclusion and Applicability)

1.4.1 Dissertation Length

This dissertation comprises a total of **wordcount XXXX** words excluding references and appendices.

2 Literature Survey & Review

2.1 The Diabetes Management Tightrope

Managing diabetes often draws parallels with walking a metabolic tightrope. On one side lies the danger of hyperglycaemia and its associated *long-term complications* such as diabetic retinopathy (damage to blood vessels in the eye leading to blindness) or renal function impairment (diabetic nephropathy), while on the other lies the *immediate peril of hypoglycaemia*. For the longest time, clinicians have helped patients navigate this delicate balance, by utilising methods or practises that show where they are, but not necessarily where they are going, in terms of blood glucose measurements. Such a reactive approach with little account for anticipative elements has made hypoglycaemic episodes an unavoidable consequence of striving for tight glycaemic control.

This "tightrope" extends beyond just taking efforts to stay safe, requiring diabetics and patients to perform constant risk assessments in everyday life. UK driving law from the DVLA mandates a blood glucose reading of above 5.0mmol/L with a repeat test every two hours for longer journeys in order to be considered safe to drive. Patients are required to carry a "hypo kit" with fast-acting glucose at all times. The worry about having episodes in public and having to rely on the awareness of strangers or being a hindrance to social situations is a constant concern. As discussed here previously this also places notable financial strain on NHS resources through emergency or ambulance costs.

With scientific and technological progress that inevitably comes with time, comes the promise of a potential safety net: the ability to anticipate or foresee signs of an episode before they present. The dangerous tightrope walk can then be transformed into a manageable path, with the help of predictive systems built with advanced sensing technologies and computational power, that can assist patients to take pre-emptive action. This review will chart the progress in this field, understanding the methodologies and ideas that have been applied to forecast hypoglycaemic events, from early tracking methods to highly optimized modern algorithms.

2.2 Methods Of Monitoring Blood Glucose Over Time

The success rate of recognising patterns in blood glucose has been vastly upgraded through the years, spurred by both technological and procedural refinements in the monitoring of blood glucose, but it has been an arduous journey to get here. The earliest methods of testing blood glucose involved urine tests, where chemical reagents like Benedict's solution were used which changed colour in the presence of sugar. Such methods were only qualitative and retroactive - they offered no actionable information as they confirmed that blood glucose had been elevated at the time of the test or in the recent past. The first blood glucose meters did not appear until the 1960s, were large and cumbersome to work with, and were mainly found in clinics. Smaller, portable meters became available around the 1970s - 1980s yet still all such meters had to be used

repetitively throughout the day, offering only a snapshot of blood glucose level in time without any means of showing a trend, stability or lack thereof.

The introduction of Continuous Glucose Monitoring (CGM) solutions from 2005 onwards completely revolutionized diabetic healthcare, allowing measurements to be taken effortlessly through sensors attached to the body. Medtronic and Dexcom's devices released after 2015 with highly improved sensor accuracy and user-friendliness even allowed connecting to smartphones and automatic insulin pumps, which could automatically stop insulin delivery if the patient did not respond to an alarm. Today's CGM's are even more cutting-edge, in that they continuously transmit data to a receiver. They are smartphone-app based for ease of use, offering enhanced features to show the trend and speed of glucose changes, alarms and alerts to proactively warn users, as well as sharing data with family members or doctors for remote monitoring and emergency handling [6].

In modern times, there is huge amounts of data available to probe into, but the challenge lies in finding the correct relevant features, at the correct level of granularity as well as the right distribution, as medical data is exceptionally rarely obtained in balanced form. The availability of rich continuous streams of data from CGMs has stimulated additional research dedicated to developing and applying algorithms to both forecast hypoglycaemic events as well as identify outliers or patterns within the data, and I delve into this next.

2.3 Review Of Relevant Literature and Similar Research

A substantial body of literature now exists on the development of models for predicting hypoglycaemia in various settings. In many scenarios, a significant proportion of the research focuses on optimizing predictive accuracy of models. For instance H. Yang et al. in 2022 [7] have used electronic health records(EHR) of patients admitted to West China Hospitals to develop a predictive model based on laboratory derived biomarkers (like lipoproteins, creatinine, globulin etc.). A similar research to this was undertaken by S. Mantena et al. [8] but on the publicly-available eICU Collaborative Research v2.0 database (eICU-CRD) that holds de-identified data for 200,000+ admissions in 553 ICUs across the USA. In reinforcement to this, UK-based studies have also been executed by Y. Ruan et al. [9] on four years worth of EHRs provided by Oxford University Hospitals NHS Foundation Trust in which they compared the performance of eighteen different predictive models based on demographic, laboratory, vital signs and previous medication predictors.

A significant commonality was observed in all the three research works, which I have incorporated into my approach as well - the emergence of XGBoost as the best predictive model with highest Area Under Receiver Operating Characteristic Curve(AUROC). Even though all studies produce excellent results in terms of predictive performance, they are not primarily aimed at finding the critical values or "turning points" of the predictors at which predictions / classifications change, which is my objective for this project that also lines up with GSTT priorities.

It needs to be stressed that achieving high accuracy on a regulated dataset in controlled circumstances is different to creating a mechanism or system that is successful, effective and dependable in a real-world clinical setting. While also comparing the performance of ML models on a new dataset, I am exploring a completely new and unique patient population, which requires a holistic approach towards the data. There is scarce research that scouts the dataset to create insights about patient population, such as the spread or extent of hypoglycaemia across wards, as the major focus is primarily predictive or comparative modelling.

2.4 A Novel Contribution to ML-based Glycaemia management in the NHS

To my knowledge, the majority of published studies or models are developed using well maintained and curated datasets. Additionally, most analyses seen prior are based on a relatively homogeneous patient population, from a major location in countries like USA or China. These may not be generalisable across different geographies or demographics. This project uses an ethnically and socioeconomically diverse patient cohort from a leading NHS Trust in London, which has produced an analysis that is robust across various age and demographic groups. It also demonstrates that disparate data streams from different verticals of the NHS Epic EHR system can be integrated and harmonised to produce real world benefits after research, to create a working model specific to the NHS that other Trusts can follow. It emphasises on the "human insights" aspect of analysis more by having greater emphasis on patient features as opposed to laboratory measurements. Most analyses have a strictly methodical and statistical approach with little insights being generated around the actual groups of patients within the data.

3 Dataset

3.1 Raw Data

Our industry advisor from GSTT has graciously provided a year's worth of data in multiple .xlsx files, which have been combined into one for the purposes of analysis and research for this project. The **raw fields provided** within the data were:

- **UniqueID:** Unique identifier for the patient and test, which is just a number. Meant to identify same patients (not personally) when considered together with Order Time, Order Date and Age, as same patients can have multiple blood glucose tests during their stay.
- **Order Date:** The date when the glucose measurement was ordered or taken.
- **Order Time:** The timestamp at which the glucose measurement was ordered or taken.
- **Inpatient Admission Date:** The date at which the patient was admitted into the medical facility.
- **Discharge Date:** The date the patient was discharged from the medical facility.
- **Length of Stay:** The amount of time the patient has spent in the medical facility in days and hours (for eg. "5d 6h").
- **Ward:** The ward that the glucose measurement was taken in, usually matches the ward that the patient was admitted to.
- **Last Lab Test Results:** The result of the glucose measurement in mmol/L. Most values in this column are of the format "Manual blood glucose: 8.70 mmol/L" or "POCT Glucose Blood Manually: 2.7 mmol/L".
- **Age:** Age of the patient at the time of measuring blood glucose in years.
- **Ethnicity:** Specific ethnicity of the patient, values ranging from "South American - Columbian" to "Black or Black British - Nigerian" to "Other" or even missing.
- **Gender Identity:** Gender of the patient.
- **HbA1c:** Numerical value of HbA1c in mmol/mol, which is a measure of the average blood glucose over the past 2-3 months. Glucose in the body sticks to red blood cells to be transported around, and gets consumed to generate energy. If the body cannot use up sugar properly then more of it sticks to blood cells and builds up. Red blood cells are active for around 2-3 months, so the reading is generally taken quarterly, and is an indicator for blood sugar problems. [10].
- **HbA1c Date:** Date the HbA1c test was done for that patient.

- **eGFR:** Estimated Glomerular Filtration rate, which is a measurement of how well the kidneys are functioning. This is a percentage from 0 to 90, with anything 91 and over displayed by the NHS electronic health record system (Epic) as ">91" because an eGFR of 91 percent and above indicates healthy renal function.
- **eGFR Date:** The date the eGFR test was conducted.

3.2 Cleaned Dataset

The following variables were **derived from the raw features** and used for analysis:

- **Age_Range:** Categorical variable to store the age category of the patient based on their age to aid in visualisation. Possible values for this column are: "Young (1 to 25)", "Adult / Middle Aged (26-50)", "Older Adult / Old (51-75)" and "Elderly (76-100)".
- **Has_Hypoglycemia:** Binary variable to store whether the patient has hypoglycemia. A glucose measurement of 4mmol/L and below means the patient is hypoglycemic and has 1 in this column, 0 otherwise.
- **Glycemia_Type:** Categorical variable to store the type of glycemia based on the patient's glucose measurement. **For the purposes of this project, the classes we have been instructed to use are (all units in mmol/L):**
 1. "Severe Hypoglycemia" - for blood glucose values 2.2 and below
 2. "Hypoglycemia"- for blood glucose values from 2.3 to 4 both inclusive
 3. "Target Range"- for blood glucose values from 4.1 to 11 both inclusive
 4. "Hyperglycemia"- for blood glucose values above 11.
- **eGFR_Category:** Categorical variable that shows how serious the loss of kidney function is, based on the eGFR percentage. The possible values for this column are:
 1. "eGFR less than 20 - Kidney Failure" - for eGFR less than or equal to 20%
 2. "eGFR between 20 & 40 - Critical Loss of Kidney Function"- for eGFR above 20% but less than or equal to 40%
 3. "eGFR between 40 & 60 - Significant Loss of Kidney Function"- for eGFR above 40% but less than or equal to 60%
 4. "eGFR between 60 & 80 - Moderate Loss of Kidney Function"- for eGFR above 60% but less than or equal to 80%
 5. "eGFR between 80 & 90 - Minor Loss of Kidney Function"- for eGFR above 80% but less than or equal to 90%
 6. "eGFR above 90 - Normal kidney function"- for eGFR above 90% (data has been processed to only include ">91" for this class as healthy eGFR is 91% and above).

- **Wider_Ethnic_Group:** Categorical variable to store the overarching ethnic group based on the one specified in the ethnicity column, as that had a total of 57 unique values. Possible values are: “Unknown or Not Stated”, “White”, “Mixed”, “Asian or Asian British”, “Black or Black British” and “Other Ethnic Groups”.

*Note that columns obtained after cleaning the original data to extract a numerical value (such as blood glucose) **have been omitted for brevity.***

Please see Appendix A subsection A.1 for screenshots of the dataset(s).

4 Methodology and Implementation

4.1 Research Strategy and Approach

Before receiving the dataset, I have conducted an exhaustive investigation of the clinical landscape surrounding hypoglycaemia as a health condition, including studying the situations in which it commonly occurs, both in hospital settings as well as in public or everyday life. I have scrutinized a plethora of factors contributing to hypoglycaemia, including associated medicines (even conflicting medications), at-risk patient profiles, habits and lifestyles, dosing errors (both excessive as well as insufficient (*insulin*)), missed meals and even alcohol consumption. This has allowed me to better assess the quality of the incoming dataset and the relevance of its features. Upon requesting additional information regarding current patient medication and alcohol intake as it was not provided originally, GSTT advised that this data is unavailable because of its inconsistent self-reported nature and due to restrictions under their information governance policy that permits access to only the data deemed necessary for the project's scope.

After receiving the data, I have thoroughly **preprocessed** it to ensure it was suitable for meaningful analysis. This included addressing data type mismatches, deriving variables to aid in visualisation and understanding, and performing necessary imputations using appropriate methods. Duplicate and missing records were handled, categorical variables were encoded to make them compatible for predictive modelling, data validity and consistency checks were enacted to confirm that values were in expected ranges (for e.g. the glucose value field), normalization was carried out where necessary. These steps were necessary to lay a strong foundation for the subsequent application of statistical tests and machine learning models. The full preprocessing workflow is depicted in Figure 1 below.

Following this, my focus was on **exploratory data analysis** to spot any anomalies or patterns near the surface. After devising research questions around the dataset, I have generated a collection of plots through Python's widely used seaborn library that I describe in detail in the main results section, which shed light on the prevalence of hypoglycaemia across various different scenarios. Special attention was paid to drawing comparisons between hypoglycaemic and non hypoglycaemic patients, in alignment with GSTT's interests that they had clarified in the project's early stages.

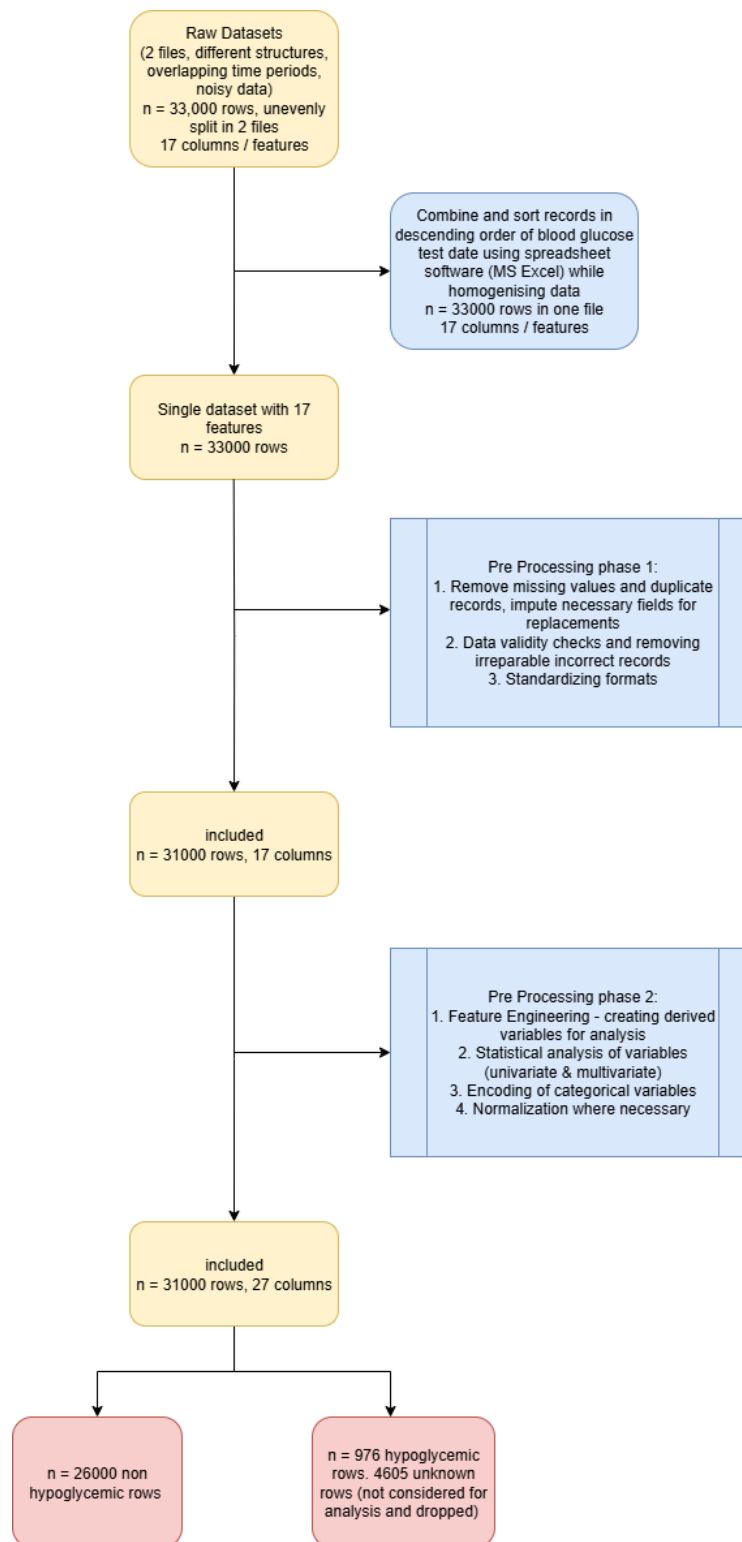


Figure 1: Data preprocessing workflow

4.2 Dealing With Imbalanced Data

It was found that there is a considerable imbalance between the number of hypoglycaemic patients (976) vs the non-hypoglycaemic patients (26022) through preprocessing. Considering this imbalance, no augmentation or sampling techniques have been carried out in the initial stages of this project to preserve the integrity and objectivity of the exploratory data analysis (EDA) phase. The rationale behind this decision was to ensure that any findings discovered from this effort remained reflective of the raw data as originally provided. Introducing synthetic data points at this stage would have led to skewed or misleading statistical and visual conclusions, especially when it comes to comparative sub-group analysis. Exploratory data analysis on a pure unadulterated dataset enabled me to locate and assess disparities between hypoglycaemic and non-hypoglycaemic patients with full clarity free from any distortion. I have visualised and documented these in the form of answers to research questions. This enabled me to meet the first objective of the project of generating beneficial, usable insights from the data without any ambiguity as to how they had been obtained.

In the predictive modelling stage of the project, some conditional sampling has been selectively employed, so as to maximise model performance. The rationale behind this was to ensure that the training process remained robust and was not biased towards the majority class. **Conditional Sampling** is a data manipulation technique where data points are drawn from a subset of a larger dataset that has been filtered on the basis of certain conditions to yield those data points that match a certain range or pattern. This yields better quality rows as compared to random sampling, and the exemplars so picked from conditional sampling were varied to not be identical to other ones. Care was taken in doing so to apply these techniques, so as to only produce sensible and valid values, in a manner so as to not undermine earlier findings. **A dual phased approach such as this one provided a healthy balance between data integrity in the early exploratory stages as well as enhanced performance in the later predictive modelling stages.**

4.3 Machine Learning Theories

With findings and observations from the exploratory data analysis process now established, I directed attention to the second major project objective. To devise a risk score for hypoglycaemia, I would first need to identify the principal components from the dataset, that is to say the features that most influence hypoglycaemia or contribute to its presence. Following this, I would need to find the values of those features which are "turning points" or at which a decision is made (for example, the "turning point" for the "age" feature would be in the higher range as old age invites complications, so this could be something from 70-90 years). This has been completed through both the EDA phase as well as some statistical and machine learning techniques that I have detailed further.

4.3.1 Decision Trees

The rationale behind using Decision Trees was that it is the most suitable algorithm to find the thresholds of the key predictors of hypoglycaemia (like HbA1c). A Decision Tree is a supervised learning algorithm that can be used for both classification and regression, matching our objective of predicting hypoglycaemia through a risk score. Its structure is hierarchical resembling a flowchart where each internal node represents a test on a feature, every branch represents the outcome of that test, and each leaf node represents the final decision or the class label. It functions by learning a set of interpretable rules from the dataset features.

The algorithm works by a process called recursive partitioning. The process begins at the root node, which contains the entire patient dataset in our case (both hypoglycaemic and non-hypoglycaemic patients). The algorithm then looks at each input feature one by one (HbA1c, eGFR etc.) and systematically test every possible value of that input feature for being a potential split point. It determines the best "split" by a metric known as **Gini's Impurity Index** which measures the probability of incorrectly classifying a randomly chosen patient from one of the 2 hypoglycaemic groups. The algorithm strives to minimise this value (A score of 0 is ideal meaning perfect purity), meaning that the resulting child nodes from the split are "pure" and contain no misclassified samples. Considering a dataset D with samples from k classes, Gini Impurity is calculated as follows:

$$Gini(D) = 1 - \sum_{i=1}^k p_i^2$$

Where:

- D is the dataset containing samples from k classes
- p_i is the probability of samples belonging to class i at a given node

The algorithm then compares purity scores from all the potential splits across all the features. It will select the single feature and the specific value of that feature that results in the most purity (and therefore creates the best possible split).

Applying this to our own dataset, to determine the decision point for a feature like HbA1c for example, the algorithm would work like so:

- Let's assume the observed values range from 40 to 100. The algorithm will test splits like $HbA1c > 40$, $HbA1c > 41$ and so on
- For each of these potential splits it calculates the impurity of the resulting 2 subsets of patients *i.e.* those above that current HbA1c threshold and those below it
- The specific HbA1c value that yields the lowest Gini Impurity value is then selected as the optimal decision rule for that node. We can then assert that HbA1c becomes

most discriminative (most influential) for predicting hypoglycaemia risk at this empirically derived value, and this is how we arrive at the most discriminative values for the main predictors of hypoglycaemia.

The feature and threshold selected at the root node split is considered to be the most powerful predictor in the dataset. However, with Decision Trees, it is vital to **avoid overfitting** which is where the tree ends up capturing or considering noise instead of the true behaviour / nature of the data. This can be managed by pruning (cutting short) the tree, or setting hyperparameters like *max_depth*, but as a step up I have employed Random Forest algorithm which works by aggregating multiple Decision Trees to get the best result.

4.3.2 Random Forest

Random Forest is an ensemble learning algorithm that works by building a large number of individual decision trees and then combining their outputs to make a final prediction. It is advantageous over regular decision trees as it is resistant to overfitting and as errors or biases in the model average out, it tends to produce a more stable collective result. Furthermore, by measuring how much a feature like HbA1c contributes to reducing impurity across all the trees that the model creates, the model can produce a feature importance score enabling us to rank the predictors in the dataset based on their significance or value.

- To start with, the algorithm conjures up many different training sets by repeatedly drawing random samples from the original dataset with replacement, meaning that some data points may be selected more than once and some may not be selected at all. Each individual decision tree is then trained on one of these unique randomly built datasets, hence the name "Random" for random sampling and "Forest" meaning multiple Decision Trees. This process is also called Bagging or Bootstrap Aggregation.
- As each tree is being built, it is not allowed to use all the available information to make its decisions. The tree is only given a random subset of the total features (e.g., only 3 out of 10 patient attributes) to choose from at every split point. This forces the forest to learn from a wider variety of predictors in addition to preventing any single strong feature from dominating all the trees.
- Once all the trees in the forest have been built, every tree in the forest makes a prediction when a new unseen data point comes along. The final prediction outcome from the Random Forest model is simply the one that received the most votes from all the trees, and this collective majority voting makes the final result more accurate as well as less error-prone.

Despite being more powerful than a single Decision Tree, it is still a formidable challenge to choose the optimal number and values of hyperparameters for a Random Forest, such

as the number of trees to build or the max depth of each tree. To tackle this problem we harness the power of GridSearchCV to find the best hyperparameters to build a Random Forest.

4.3.3 GridSearchCV

GridSearchCV stands for Grid Search Cross Validation. GridSearchCV is valuable and advantageous because it takes a methodical and exhaustive approach ("try everything out and find the best") to ascertain the optimal values of hyperparameters. Choosing the right values of hyperparameters to train a Random Forest is crucial because it can lead to a model that is either too simple and underfits or too complex and overfits, GridSearchCV enables us to address this issue. Its name breaks down its process:

- **Grid:** We define a "grid" of hyperparameters and their values to try. For example a grid could have *max_depth* values of [3,5,7] and *n_estimators* (number of trees to build) of [50,100,125] in the grid.
- **Cross Validation:** GridSearchCV splits the training data into "folds" and for each hyperparameter combination, it trains the model on some folds and tests it on the remaining folds, while continuously rotating around which fold is used for testing with every iteration. Finally it averages the performance scores from all the folds.

GridSearchCV is highly beneficial in that it automates what would have been a very tedious manual task while also producing the best possible reliable results and reducing the risk of overfitting. Although it does return the best possible hyperparameter values to train the model, there is a risk of making the "grid" too large, which could potentially lead to GridSearchCV running for ages, sometimes even days or weeks, so it is imperative to choose an accurately sized grid based on the data and objective at hand.

The content of “Main results” is in “\contents\introduction.tex”

5 Main Results and Findings

The chapter reports the contributions of your work. For example, it could contain the following sub-sections to summarise the contribution of the project such as Theoretical Development, Analysis and Design, Implementation and Experimental Work, Results, Observation and Discussion.

5.1 Interpretation Analysis and Evaluation

It summarises the results obtained from the proposed design and methodology. The way to obtain the results should be described in detail. Analysis and evaluation have to be performed. Comparisons should be made. It should justifies if the project aims, objectives, requirements and specifications have been achieved.

6 Math equations

This section is for demonstration of equations, figures, tables, which is not required for the report.

6.1 Maths

$$\frac{dS_t}{S_t} = rdt + \sigma dW_t, \quad S_0 > 0, \quad (6.1)$$

The equation $\sigma = ma$ follows easily [?].

6.2 Figures

Here is an example [?] of how to insert a picture:

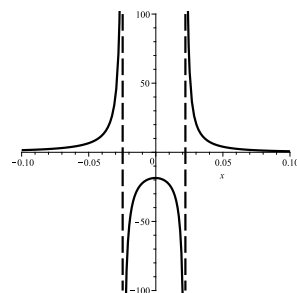


Figure 2: This is the caption for the figure.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero

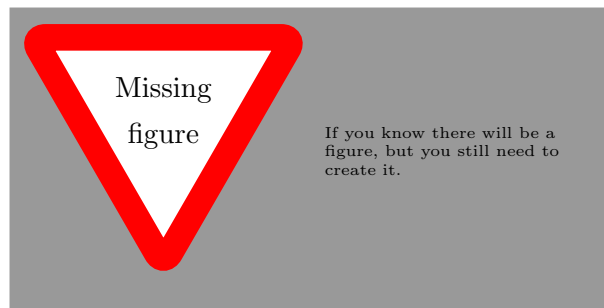


Figure 3: This is the caption for the figure which is not even present.

eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

or two side-by-side pictures:

This is a small Todo, please take care!

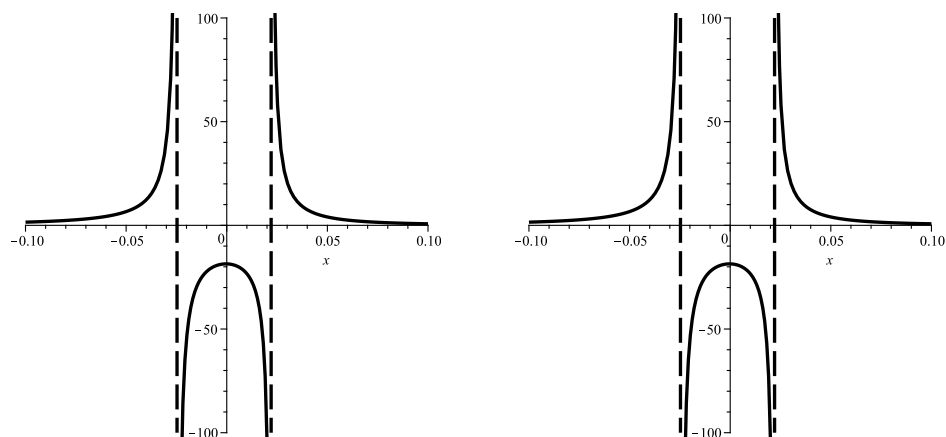


Figure 4: Another caption

7 Legal, Social, Ethical and Professional Issues

Research and projects within the medical domain are always inherently sensitive regardless of the kind of data involved or the presence of human participants. This sensitivity is amplified when a highly prominent industry stakeholder such as the NHS is interested, in view of the fact that it oversees public health across all of the UK. Right from the start, I have prioritized regular and transparent communication with our industry advisor through recurring meetings, while upholding implicit confidentiality agreements regarding the nature of the data and the project's specific objectives. All analysis was conducted within the agreed-upon scope. All deliverables were presented in a coherent and actionable format, thereby reflecting my commitment to their distinct requirements and towards fostering a trustworthy working relationship. Being cognizant of my social and ethical responsibility in this undertaking to advance public welfare, I have submitted an application in KCL's Research Ethics Management Application System (REMAS) which should supplement the agreements and principles established at the time of inception of the project, considering that the project is a KEP with industry (NHS England). According to KCL and REMAS guidelines, this project is classed as "Minimal Risk" [ref], in that it involves the study of pre-existing data that is not available to the general public, but is fully anonymous at the point which I as a researcher gain access to it. The industry advisor has kindly provided us the necessary data after complete anonymization, which removes any risk of personal identification. (Still submitted a Full Application Form instead of a minimal risk application) To further support this and in line with the guidelines listed in the General Data Protection Regulation (GDPR) as well as the Data Protection Act (DPA) 2018, the data was both shared with me and only accessed through secure organization / university credentials, meaning that it did not need to be fetched at all through any resource or API calls, eliminating the risk of interception. It was stored locally for on-machine data analysis and modelling through frequently used, open-source Python libraries, without the involvement of any online tools where the data has to be uploaded for research. Efforts have been taken to determine whether the project requires approval from any external entities, for example the Health Research Authority[ref] <https://www.hra.nhs.uk/planning-and-improving-research/research-planning/student-research/>. This was found to be not necessary. No recruitment of human participants was in the picture. Every care was taken to prevent any conflicts of interest from occurring, whether around other similar research, intellectual property, project objectives or any other sectors. I have also considered reliability measures to minimize the possibility of any kind of "reverse engineering" that may be carried out on my work. This substantiates that I have displayed special adherence to the British Computer Society (BCS) Code of Conduct and Code of Practise[ref], especially the directives regarding "Public Interest" and "Professional Competence and Integrity".

Socially, care has been taken to ensure that no adverse effects can occur as a result of this research

The content of “Conclusion” is in “\contents\conclusion.tex”

8 Conclusion

It is a chapter to sum up the main points and findings of the work; how you achieve the project aims and address the research questions; the contributions and results you have achieved. Future plan and development can be mentioned in this section as well. It is normally in one or two pages.

References

- [1] Diabetes UK, “Cost of devastating complications highlights urgent need to transform diabetes care in the UK.” [URL](#). Accessed: 2025-07-07.
- [2] Diabetes Professional Care, “Almost 60 costs are for ‘preventable’ complications.” [URL](#). Accessed: 2025-07-07.
- [3] Hospital Pharmacy Europe, “Data from health economic model shows cost of managing hypoglycaemia.” [URL](#). Accessed: 2025-07-07.
- [4] Diabetes Research and Wellness Foundation, “Emergency call-outs for diabetes-related condition reduced following hypos education campaign.” [URL](#). Accessed: 2025-07-07.
- [5] NHS Digital, “National diabetes audit 2021-22, report 1: Care processes and treatment targets, detailed analysis report.” [National Diabetes Audit 2021-22](#). Accessed: 2025-07-07.
- [6] C. Bender, P. Vestergaard, and S. L. Cichosz, “The history, evolution and future of continuous glucose monitoring (CGM),” *Diabetology*, vol. 6, no. 3, 2025.
- [7] Yang H, Li J, Liu S, Yang X, Liu J, “Predicting risk of hypoglycemia in patients with type 2 diabetes by electronic health record-based machine learning: Development and validation. *JMIR Medical Informatics* vol. 10,6 e36958. 16 Jun. 2022.” [URL](#). Accessed: 2025-07-08.
- [8] Mantena, S., Arévalo, A. R., Maley, J. H., da Silva Vieira, S. M., Mateo-Collado, R., da Costa Sousa, J. M., & Celi, L. A. (2022), “Predicting hypoglycemia in critically ill patients using machine learning and electronic health records. *Journal of Clinical Monitoring and Computing*, 36.” [URL](#). Accessed: 2025-07-08.
- [9] Y. Ruan, A. Bellot, Z. Moysova, G. D. Tan, A. Lumb, J. Davies, M. van der Schaar, R. Rea, “Predicting the risk of inpatient hypoglycemia with machine learning using electronic health records. *diabetes care* 1 july 2020.” [URL](#). Accessed: 2025-07-08.
- [10] Diabetes UK, “What is hba1c?.” [URL](#). Accessed: 2025-07-07.

A Appendix

A.1 Dataset

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|-----|----------|------------|-----------|-----------|----------------|---------------|--------------|------------------------|--------------------------------------|-----|--|------------------------|----------|-----------|-------------|------------|-------------|
| 1 | UniqueID | OrderDate | OrderTime | OrderID | InptAdmnsnDate | DischargeDate | LengthofStay | Ward | LastLabTestResults | Age | Ethnicity | Gender | Identity | LastHbA1c | LastHbA1cDt | LasteGFR | AdmitWeight |
| 266 | 285 | 05.06.2025 | 10:58:03 | 24856556 | 05.06.2025 | 06.06.2025 | 1d 9h | GH RICHARD BRIGHT WARD | Manual blood glucose: 6.20 mmol/L | 64 | Not stated/Undefined | Male | | 82 | 19.12.2024 | 73 | 27.06.2025 |
| 267 | 282 | 05.06.2025 | 11:02:08 | 248598252 | 05.06.2025 | 06.06.2025 | 1d 9h | STH GYNAECOLOGY WARD | Manual blood glucose: 10.80 mmol/L | 52 | Black or Black British - Unspecified | Female | | | | 27 | 06.06.2025 |
| 268 | 267 | 05.06.2025 | 11:02:22 | 248558363 | 05.06.2025 | 06.06.2025 | 1d 9h | STH GYNAECOLOGY WARD | Manual blood glucose: 12.00 mmol/L | 52 | Black or Black British - Unspecified | Female | | | | 77 | 06.06.2025 |
| 269 | 267 | 05.06.2025 | 11:21:12 | 248568251 | 05.06.2025 | 05.06.2025 | 3h | STH ADMISSIONS WARD | Manual blood glucose: 10.10 mmol/L | 56 | | | | | | 78 | 05.06.2025 |
| 270 | 269 | 05.06.2025 | 11:58:35 | 248598646 | 24.01.2025 | 157d 18h | | STH WILLIAM GULL WARD | Manual blood glucose: 5.20 mmol/L | 59 | Black or Black British - African | Female | | 43 | 20.01.2024 | 58 | 01.07.2025 |
| 271 | 3 | 05.06.2025 | 11:59:38 | 248598730 | 25.05.2025 | 18.06.2025 | 34d 2h | STH WILLIAM GULL WARD | Manual blood glucose: 10.40 mmol/L | 88 | White - Any other White background | Choose not to disclose | | 53 | 30.04.2024 | 59 | 18.06.2025 |
| 272 | 4 | 05.06.2025 | 12:06:47 | 248599054 | 05.06.2025 | 06.06.2025 | 1d 9h | GH RICHARD BRIGHT WARD | Manual blood glucose: 5.70 mmol/L | 64 | Not stated/Undefined | Male | | 82 | 19.12.2024 | 23 | 27.06.2025 |
| 273 | 262 | 05.06.2025 | 12:21:30 | 248598833 | 16.05.2025 | 27.06.2025 | 42d 1h | STH SARAH SWIFT WARD | Manual blood glucose: 11.90 mmol/L | 76 | White - British | | | 54 | 22.04.2024 | 50 | 30.06.2025 |
| 274 | 273 | 05.06.2025 | 12:54:34 | 249011417 | 17.05.2025 | | 45d 9h | STH HILLIERS WARD | Manual blood glucose: 7.40 mmol/L | 74 | Black or Black British - Caribbean | | | | >90 | 27.06.2025 | 55 kg |
| 275 | 11 | 05.06.2025 | 13:43:17 | 249031415 | 31.05.2025 | | 30d 18h | STH WILLIAM GULL WARD | POCT Glucose Blood Manually En...: 7 | 52 | White - English | Male | | | | 32 | 24.06.2025 |
| 276 | 275 | 05.06.2025 | 14:16:46 | 249045211 | 04.06.2025 | 07.06.2025 | 3d 3h | GH FLORENCE WARD | Manual blood glucose: 4.80 mmol/L | 74 | White - English | | | 61 | 13.11.2023 | 82 | 20.06.2025 |
| 277 | 276 | 05.06.2025 | 15:07:28 | 249067995 | 04.06.2025 | 20.06.2025 | 16d | STH GI PAGE WARD | Manual blood glucose: 7.20 mmol/L | 81 | Not stated/Undefined | Male | | | | 39 | 24.04.2025 |
| 278 | 277 | 05.06.2025 | 15:07:40 | 249067474 | 01.06.2025 | 06.06.2025 | 5d 16h | STH ALEXANDRA WARD | Manual blood glucose: 6.50 mmol/L | 78 | Black or Black British - Caribbean | Female | | | | 5 | 17.06.2025 |
| 279 | 278 | 05.06.2025 | 15:42:24 | 249082295 | 06.06.2025 | 06.06.2025 | 16h | STH ADMISSIONS WARD | Manual blood glucose: 9.60 mmol/L | 68 | Black or Black British - African | Male | | 47 | 03.05.2024 | 58 | 06.06.2025 |
| 280 | 279 | 05.06.2025 | 15:56:30 | 249087656 | 01.06.2025 | 09.06.2025 | 8d 17h | STH ALBERT WARD | Manual blood glucose: 5.70 mmol/L | 34 | Black or Black British - African | Male | | | >90 | 25.06.2025 | |
| 281 | 280 | 05.06.2025 | 16:07:02 | 249091566 | 05.06.2025 | 06.06.2025 | 1d 10h | GH ASTON KEY WARD | Manual blood glucose: 6.10 mmol/L | 72 | Black or Black British - Caribbean | Male | | 43 | 17.04.2024 | 60 | 06.06.2025 |
| 282 | 281 | 05.06.2025 | 16:59:47 | 249109429 | 05.06.2025 | 06.06.2025 | 1d 9h | STH GYNAECOLOGY WARD | Manual blood glucose: 7.50 mmol/L | 52 | Black or Black British - Unspecified | Female | | | 77 | 06.06.2025 | |
| 283 | 267 | 05.06.2025 | 17:56:36 | 249118534 | 05.06.2025 | 06.06.2025 | 1d 10h | GH ASTON KEY WARD | Manual blood glucose: 6.30 mmol/L | 72 | Black or Black British - Caribbean | Male | | 43 | 17.04.2024 | 60 | 06.06.2025 |
| 284 | 281 | 05.06.2025 | 17:55:00 | 249122117 | 05.06.2025 | 06.06.2025 | 1d 1h | GH SARAH WARD | Manual blood glucose: 6.50 mmol/L | 69 | White - English | | | | >90 | 05.06.2025 | |
| 285 | 284 | 05.06.2025 | 18:03:54 | 249123786 | 22.04.2025 | | 70d 9h | STH MARK WARD | Manual blood glucose: 5.10 mmol/L | 68 | White - British | | | 47 | 12.04.2024 | >90 | 30.06.2025 |
| 286 | 285 | 05.06.2025 | 19:02:34 | 249132405 | 03.06.2025 | 07.06.2025 | 4d 13h | GH ASTON KEY WARD | Manual blood glucose: 10.80 mmol/L | 50 | Black or Black British - Any other Black | Female | | 109 | 10.06.2025 | >90 | 10.06.2025 |
| 287 | 118 | 05.06.2025 | 19:25:07 | 249134589 | 17.05.2025 | | 45d 9h | STH HILLIERS WARD | Manual blood glucose: 6.00 mmol/L | 74 | Black or Black British - Caribbean | | | 54 | 22.04.2024 | 50 | 30.06.2025 |
| 288 | 11 | 05.06.2025 | 19:26:43 | 249134739 | 05.06.2025 | 12.06.2025 | 6d 21h | STH WILLIAM GULL WARD | Manual blood glucose: 11.10 mmol/L | 65 | White - English | Male | | | | 70 | 30.06.2025 |
| 289 | 288 | 05.06.2025 | 19:27:27 | 249134791 | 30.04.2025 | 61d 17h | | STH WILLIAM GULL WARD | Manual blood glucose: 13.00 mmol/L | 59 | White - Any other White background | Male | | | 72 | 26.06.2025 | 62.3 kg |
| 290 | 12 | 05.06.2025 | 19:56:11 | 75513886 | 05.06.2025 | 06.06.2025 | 14h | STH ADMISSIONS WARD | Manual blood glucose: 9.20 mmol/L | 41 | Black or Black British - Caribbean | | | | 63 | 06.06.2025 | |
| 291 | 290 | 05.06.2025 | 20:42:10 | 249140669 | 14.04.2025 | | 77d 17h | STH SOMERSET WARD | Manual blood glucose: 3.40 mmol/L | 35 | Any Other Ethnic Group | Male | | | 81 | 29.06.2025 | 77 kg |
| 292 | 148 | 05.06.2025 | 20:42:48 | 249140642 | 14.04.2025 | | 77d 17h | STH SOMERSET WARD | Manual blood glucose: 4.30 mmol/L | 35 | Any Other Ethnic Group | Male | | | 81 | 29.06.2025 | 77 kg |
| 293 | 148 | 05.06.2025 | 20:44:12 | 249140769 | 06.06.2025 | 07.06.2025 | 1d 13h | STH SOMERSET WARD | Manual blood glucose: 6.60 mmol/L | 53 | Black or Black British - Caribbean | Female | | 40 | 08.12.2023 | >90 | 05.06.2025 |
| 294 | 293 | 05.06.2025 | 21:07:45 | 249142353 | 06.06.2025 | 06.06.2025 | 16h | STH ADMISSIONS WARD | Manual blood glucose: 6.60 mmol/L | 68 | Black or Black British - African | Male | | 47 | 03.05.2024 | 58 | 06.06.2025 |
| 295 | 279 | 05.06.2025 | 21:07:45 | 249142357 | 06.06.2025 | 06.06.2025 | 16h | STH ADMISSIONS WARD | POCT Glucose Blood Manually E...: 11 | 69 | White - English | Male | | 47 | 03.05.2024 | 58 | 06.06.2025 |
| 296 | 279 | 05.06.2025 | 21:16:33 | 249143099 | 05.06.2025 | 06.06.2025 | 1d 1h | GH SARAH WARD | Manual blood glucose: 12.40 mmol/L | 69 | White - English | | | | >90 | 05.06.2025 | |
| 297 | 284 | 05.06.2025 | 21:43:01 | 249144918 | 21.05.2025 | | 40d 19h | STH ALEXANDRA WARD | Manual blood glucose: 6.20 mmol/L | 82 | Not stated/Undefined | | | 53 | 01.07.2025 | 67.2 kg | |
| 298 | 297 | 05.06.2025 | 22:01:57 | 249146383 | 25.05.2025 | 18.06.2025 | 24d 2h | STH WILLIAM GULL WARD | Manual blood glucose: 11.00 mmol/L | 88 | White - Any other White background | Choose not to disclose | | 53 | 30.04.2024 | 59 | 18.06.2025 |
| 299 | 4 | 05.06.2025 | 22:13:12 | 249147160 | 24.01.2025 | | 157d 18h | STH WILLIAM GULL WARD | Manual blood glucose: 6.10 mmol/L | 59 | Black or Black British - African | Female | | 43 | 20.01.2024 | 58 | 01.07.2025 |
| 300 | 3 | 05.06.2025 | 22:46:05 | 249149081 | 06.06.2025 | 10.06.2025 | 4d 9h | STH MARK WARD | Manual blood glucose: 9.20 mmol/L | 72 | Asian or Asian British - Arab | Male | | 64 | 13.05.2024 | 59 | 26.06.2025 |
| 301 | 300 | 05.06.2025 | 23:44:45 | 249152341 | 14.04.2025 | | 77d 17h | STH SOMERSET WARD | Manual blood glucose: 5.80 mmol/L | 35 | Any Other Ethnic Group | Male | | | 81 | 29.06.2025 | 77 kg |
| 302 | 148 | 06.06.2025 | 16:48:01 | 249838961 | 07.06.2025 | 14.06.2025 | 7d 18h | STH ALEXANDRA WARD | Manual blood glucose: 6.80 mmol/L | 99 | Black or Black British - Caribbean | Female | | | 26 | 16.06.2025 | |

Figure 5: Raw dataset

| | J | K | L | M | N | O | P | Q | R | U | V | W | Y | Z | AA | |
|-----|-----|--|------------------------|----------|------------|---------------|-----------|------------|--------------|---------------|---------------------|-----------------------------|-----------|-----------------|--|------------------------|
| 1 | Age | Ethnicity | Gender | Identity | Last HbA1c | Last HbA1c Dt | Last eGFR | eGFR Date | Admit Weight | Glucose Value | Length of Stay (Ti) | Age Range | Has_Hypog | Glycemia Type | eGFR Category | Wider_Ethnic_Group |
| 269 | 56 | 59 Black or Black British - African | Female | | 43 | 20.01.2024 | 78 | 05.06.2025 | | 10.1 | 0 days 03:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 60 & 80 - Moderate Loss of Kidney Function | Black or Black British |
| 270 | 58 | 88 White - Any other White background | Choose not to disclose | | 53 | 30.04.2024 | 58 | 01.07.2025 | 121 kg | 5.2 | 157 days 18:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 271 | 88 | White - Any other White background | Choose not to disclose | | 53 | 30.04.2024 | 59 | 18.06.2025 | | 10.4 | 24 days 02:00:00 | Elderly (76-100) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | White |
| 272 | 64 | Not stated/Undefined | Male | | 82 | 19.12.2024 | 23 | 27.06.2025 | | 5.7 | 1 days 09:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 20 & 40 - Critical Loss of Kidney Function | Unknown or Not Stated |
| 273 | 76 | White - British | | | | | 61 | 27.06.2025 | | 11.9 | 42 days 01:00:00 | Elderly (76-100) | | 0 Hyperglycemia | eGFR between 60 & 80 - Moderate Loss of Kidney Function | White |
| 274 | 74 | Black or Black British - Caribbean | | | 54 | 22.04.2024 | 50 | 30.06.2025 | 95 kg | 7.4 | 45 days 09:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 275 | 52 | White - English | | | | | >90 | | | 7.5 | 30 days 18:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR above 90 - Normal kidney function | White |
| 276 | 74 | White - English | Male | | | | 32 | 24.06.2025 | | 4.8 | 3 days 03:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 20 & 40 - Critical Loss of Kidney Function | White |
| 277 | 81 | Not stated/Undefined | | | 61 | 13.11.2023 | 82 | 20.06.2025 | | 7.2 | 16 days 00:00:00 | Elderly (76-100) | | 0 Target Range | eGFR between 80 & 90 - Minor Loss of Kidney Function | Unknown or Not Stated |
| 278 | 78 | Black or Black British - Caribbean | Female | | 39 | 24.04.2025 | 5 | 17.06.2025 | | 6.5 | 5 days 16:00:00 | Elderly (76-100) | | 0 Target Range | eGFR less than 20 - Kidney Failure | Black or Black British |
| 279 | 68 | Black or Black British - African | Male | | 47 | 03.05.2024 | 58 | 06.06.2025 | | 9.6 | 0 days 16:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 280 | 34 | Black or Black British - African | Male | | 280 | 17.04.2024 | >90 | 06.06.2025 | | 5.7 | 8 days 17:00:00 | Adult / Middle Aged (26-50) | | 0 Target Range | eGFR above 90 - Normal kidney function | Black or Black British |
| 281 | 72 | Black or Black British - Caribbean | Male | | 43 | 17.04.2024 | 60 | 06.06.2025 | | 6.1 | 1 days 10:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 282 | 52 | Black or Black British - Unspecified | Female | | 43 | 17.04.2024 | 77 | 06.06.2025 | | 7.5 | 1 days 09:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 60 & 80 - Moderate Loss of Kidney Function | Black or Black British |
| 283 | 72 | Black or Black British - Caribbean | Male | | 43 | 17.04.2024 | 60 | 06.06.2025 | | 6.3 | 1 days 10:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 284 | 69 | White - English | Male | | | | >90 | 05.06.2025 | | 6.5 | 1 days 01:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR above 90 - Normal kidney function | White |
| 285 | 68 | White - British | Female | | 47 | 12.04.2024 | >90 | 30.06.2025 | 75 kg | 5.1 | 70 days 09:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR above 90 - Normal kidney function | White |
| 286 | 74 | Black or Black British - Any other Black British | Female | | 109 | 10.06.2025 | >90 | 10.06.2025 | | 10.8 | 4 days 13:00:00 | Adult / Middle Aged (26-50) | | 0 Target Range | eGFR above 90 - Normal kidney function | White |
| 287 | 65 | White - English | Male | | 54 | 22.04.2024 | 50 | 30.06.2025 | 95 kg | 6.4 | 5 days 09:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 288 | 59 | White - Any other White background | Male | | | | 70 | 30.06.2025 | | 11.1 | 6 days 21:00:00 | Older Adult / Old (51-75) | | 0 Hyperglycemia | eGFR between 40 & 80 - Moderate Loss of Kidney Function | White |
| 289 | 59 | White - Any other White background | Male | | | | 72 | 26.06.2025 | 62.3 kg | 13 | 61 days 17:00:00 | Older Adult / Old (51-75) | | 0 Hyperglycemia | eGFR between 60 & 80 - Moderate Loss of Kidney Function | White |
| 290 | 41 | Black or Black British - Caribbean | Male | | 290 | 41 | 63 | 06.06.2025 | | 9.2 | 0 days 14:00:00 | Adult / Middle Aged (26-50) | | 0 Target Range | eGFR between 60 & 80 - Moderate Loss of Kidney Function | Black or Black British |
| 291 | 35 | Any Other Ethnic Group | Male | | 291 | 35 | 81 | 29.06.2025 | 77 kg | 3.4 | 77 days 17:00:00 | Adult / Middle Aged (26-50) | | 1 Hypoglycemia | eGFR between 80 & 90 - Minor Loss of Kidney Function | Other Ethnic Groups |
| 292 | 35 | Any Other Ethnic Group | Male | | | | 81 | 29.06.2025 | 77 kg | 4.3 | 77 days 17:00:00 | Adult / Middle Aged (26-50) | | 0 Target Range | eGFR between 80 & 90 - Minor Loss of Kidney Function | Other Ethnic Groups |
| 293 | 53 | Black or Black British - Caribbean | Female | | 40 | 08.12.2023 | >90 | 05.06.2025 | | 6.6 | 1 days 13:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR above 90 - Normal kidney function | Black or Black British |
| 294 | 68 | Black or Black British - African | Male | | 47 | 03.05.2024 | 58 | 06.06.2025 | | 0 | 0 days 16:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 295 | 68 | Black or Black British - African | Male | | 47 | 03.05.2024 | 58 | 06.06.2025 | | 11.4 | 0 days 16:00:00 | Older Adult / Old (51-75) | | 0 Hyperglycemia | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 296 | 69 | White - English | | | | | 58 | 06.06.2025 | | 12.4 | 1 days 01:00:00 | Older Adult / Old (51-75) | | 0 Hyperglycemia | eGFR above 90 - Normal kidney function | White |
| 297 | 82 | Not stated/Undefined | | | 53 | 01.07.2025 | 53 | 01.07.2025 | 67.2 kg | 6.2 | 40 days 19:00:00 | Elderly (76-100) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Unknown or Not Stated |
| 298 | 88 | White - Any other White background | Choose not to disclose | | 53 | 30.04.2024 | 59 | 18.06.2025 | | 11 | 24 days 02:00:00 | Elderly (76-100) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | White |
| 299 | 59 | Black or Black British - African | Female | | 43 | 20.01.2024 | 58 | 01.07.2025 | 121 kg | 6.1 | 157 days 18:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Black or Black British |
| 300 | 72 | Asian or Asian British - Arab | Male | | 64 | 13.05.2024 | 59 | 26.06.2025 | | 9.2 | 4 days 09:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | Asian or Asian British |
| 301 | 35 | Any Other Ethnic Group | Male | | | | 81 | 29.06.2025 | 77 kg | 5.8 | 77 days 17:00:00 | Adult / Middle Aged (26-50) | | 0 Target Range | eGFR between 80 & 90 - Minor Loss of Kidney Function | Other Ethnic Groups |
| 302 | 99 | Black or Black British - Caribbean | Female | | | | 26 | 16.06.2025 | | 6.8 | 7 days 18:00:00 | Elderly (76-100) | | 0 Target Range | eGFR between 20 & 40 - Critical Loss of Kidney Function | Black or Black British |
| 303 | 76 | White - British | Female | | 61 | 27.06.2025 | 61 | 27.06.2025 | | 7.3 | 42 days 01:00:00 | Elderly (76-100) | | 0 Target Range | eGFR between 60 & 80 - Moderate Loss of Kidney Function | White |
| 304 | 88 | White - British | Female | | | | 49 | 10.06.2025 | | 4.8 | 6 days 21:00:00 | Elderly (76-100) | | 0 Target Range | eGFR between 40 & 60 - Significant Loss of Kidney Function | White |
| 305 | 71 | White - British | | | | | >90 | 07.06.2025 | | 6.2 | 3 days 08:00:00 | Older Adult / Old (51-75) | | 0 Target Range | eGFR above 90 - Normal kidney function | White |

Figure 6: Dataset with cleaned features (this is in addition to the fields of the raw dataset)