1. **Project Title:** Evaluating the Effectiveness of SMOTE and SMOTE-ENC in Class Imbalance Medical Data

2. **A Short Summary of Project Topic and Background:** Recent data shows that Chronic Obstructive Pulmonary Disease (COPD) poses major health challenges internationally because it impacts millions of patients while driving both high mortality rates and escalating healthcare expenditures. Although early and accurate diagnosis serves to improve patient health results, current diagnostic methods remain invasive and resource-demanding with frequent incorrect results (Yang, 2024). The ExaSens dataset develops a non-invasive diagnostic solution through saliva sample analysis which enables the classification of respiratory diseases like COPD together with asthma and infection. The dataset contains unbalanced class distributions which create major difficulties when developing precise classification techniques. Machine learning models face unreliable predictions of minor classes due to class imbalance which favors the majority class outcomes in sensitivity tests with conditions such as specific COPD subtypes being misclassified (Glyde, 2024). To develop medical diagnostic machine learning applications and enhance clinical decision-making healthcare professionals must overcome data class imbalance challenges. The main focus of this research is to evaluate oversampling techniques named SMOTE and SMOTE-ENC (SMOTE-Encoded Nominal and Continuous), in the context of classification tasks using the ExaSens dataset, which contains saliva samples of COPD (Chronic Obstructive Pulmonary Disease) patients. Drawing from the methodology proposed by Mukherjee and Khushi (2021), who introduced SMOTE-ENC for dataset with mixed (nominal and continuous) features, it is extended to improve prediction accuracy for COPD classification. The ExaSens dataset, which combines both continuous and categorical data, represents a suitable test case for the applicability of SMOTE as well as SMOTE-ENC. By applying these oversampling techniques to this dataset, the goal is to balance the minority class and improve the accuracy of ML models, particularly in the prediction of underrepresented COPD conditions.

3. **Research Question:** How does SMOTE or SMOTE-ENC technique perform in balancing datasets with both nominal and continuous features for COPD classification using the ExaSens dataset?

4. **Objectives:**
    - To conduct a comprehensive literature review on class imbalance in datasets with mixed nominal and continuous features and identify existing gaps.
    - To gather and pre-process the ExaSens dataset, which contains saliva samples from COPD patients, ensuring that both continuous and nominal features are properly formatted for machine learning analysis.
    - To apply SMOTE as well as SMOTE-ENC separately in addressing class imbalance in the pre-processed data for COPD classification task.
    - To implement ML models, such as Decision Trees, SVM, and Random Forest, on both SMOTE and SMOTE-ENC balanced data.
    - To compare the performance of various classifiers on both SMOTE and SMOTE-ENC balanced data by performance metrics such as accuracy, precision, recall, F1-Score, Area Under Precision-Recall Curve (AUC-PR), Matthews Correlation Coefficient (MCC) and balanced accuracy.

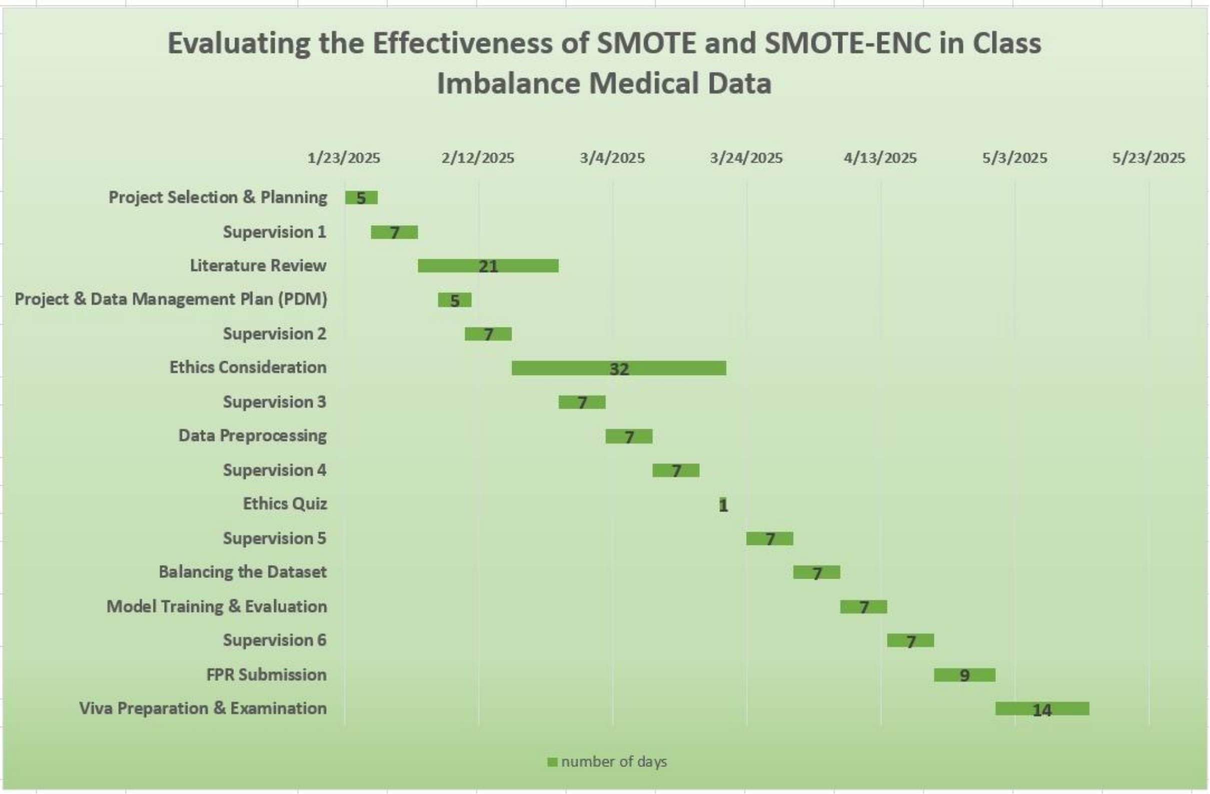
5. **Data Management Plan:**
    - **Overview of the Dataset:** The ExaSens dataset contains saliva samples from patients with COPD, asthma, respiratory infections, and healthy controls, providing valuable data for classification tasks in medical research.

- **Data Collection:** The dataset will be downloaded from IEEE's Dataport repository (https://ieee-dataport.org/open-access/exasens-novel-dataset-classification-saliva-samples-copd-patients).
- **Metadata:** The dataset contains approximately 399 records with 8 attributes, including demographic details and saliva permittivity measurements. The file size is around 30.81 KB.
- **Document Control:** A GitHub repository (https://github.com/Nandana-vijayan/Evaluating-the-Effectiveness-of-SMOTE-ENC-in-Class-Imbalance-Medical-Data.git) will serve as the code storage location while weekly code changes will be committed. Each file across the repository must follow a standard naming method while each commit needs to include detailed descriptions about the implemented changes.
- **ReadMe File:** The ReadMe document provides instructions about project description, installation process, data usage steps and dependency requirements. Future users will find guidance in the written instructions to understand and make use of the provided code.
- **Security and Storage:** Each week all data as well as code will be uploaded to GitHub platform while also being backed up to an online cloud platform. The project staff and markers will receive the exclusive access to view the GitHub repository.
- **Ethical Requirements:** The GDPR does not apply to this dataset because it doesn't involve both personal information and identifying attributes about any individual. The research follows University of Hertfordshire's ethical principles while also avoiding ethical problems. The required permission to use the dataset from IEEE Dataport has been granted since there are no research restrictions. The research did not require ethical considerations since the utilized fictional dataset excluded any traces of personal information.

6. Project Plan:

Task to be done	Description of assigned tasks	Start Date of the task	End Date of the task
Choice of Project & Planning	Choose research topic, finalize dataset, and submit project selection form.	23/01/2025	27/01/2025
Supervision 1	Discuss project scope, outline initial plan with supervisor.	27/01/2025	02/02/2025
Literature Review	Conduct literature search on COPD classification.	03/02/2025	23/02/2025
Project & Data Management Plan (PDM)	Prepare PDM document and presentation.	06/02/2025	10/02/2025
Supervision 2	Present PDM plan and receive feedback.	10/02/2025	16/02/2025
Ethics Consideration	Study ethical requirements and prepare for Ethics Quiz.	17/02/2025	20/03/2025
Supervision 3	Submit draft literature review.	24/02/2025	02/03/2025
Data Preprocessing stage	Handling of missing values as well as scaling of numerical data.	03/03/2025	09/03/2025
Supervision 4	Present draft methodology.	10/03/2025	16/03/2025
Ethics Quiz	Complete and submit ethics assessment.	20/03/2025	20/03/2025
Supervision 5	Mock viva preparation, review code.	24/03/2025	30/03/2025
Balancing the Dataset	Implement SMOTE and SMOTE-ENC on the datasets.	31/03/2025	06/04/2025

Model Training & Evaluation	Train models DT, SVM, RF with SMOTE and SMOTE-ENC datasets and compare performance.	07/04/2025	13/04/2025
Supervision 6	Submit draft FPR and discuss analysis with supervisor.	14/04/2025	20/04/2025
FPR Submission	Submit final project report for assessment.	21/04/2025	29/04/2025
Viva Preparation & Examination	Prepare for viva and attend scheduled examination session.	30/04/2025	13/05/2025



7. Reference List and Bibliography:

Glyde, H.M.M., 2024. Predictive modelling acute exacerbations of chronic obstructive pulmonary disease. Doctoral dissertation, University of Bristol.

Mukherjee, M. and Khushi, M., 2021. SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. Applied System Innovation, 4(1), p.18.

Yang, X., 2024. Application and prospects of artificial intelligence technology in early screening of chronic obstructive pulmonary disease at primary healthcare institutions in China. International Journal of Chronic Obstructive Pulmonary Disease, pp.1061-1067.