Final Project Proposal

Healthcare Predictive Analytics Project

(Oral Cancer Data Prediction)

1. Project Description:

Oral cancer remains a significant public health concern worldwide, influenced by multiple lifestyle and genetic factors such as tobacco use, alcohol consumption, HPV infection, poor oral hygiene, and family medical history. This project focuses on Healthcare Predictive Analytics through exploratory data analysis (EDA) and predictive modeling on a large dataset of 84,922 patient records from various countries. The aim is to uncover the relationships among demographic, behavioral, and clinical factors that contribute to oral cancer risk, and to design predictive models for early detection and recovery estimation. By deriving additional features (e.g., predicted recovery time, hospital stay, diagnosis year, and age group), the project also seeks to support healthcare planning and improve decision-making in preventive medicine.

2. Group Members & Roles

No.	Team Member Name	Role
1	**Mohamed Gamal Eldeen Ismail Gamal Eldeen**	Team Leader
2	**Shimaa Mostafa Soliman Eid**	Data Preprocessing & Cleaning
3	**Habiba Hany Farouk Hakki**	Exploratory Data Analysis (EDA) & Visualization
4	**Zeiad Saher Salama Mahmoud**	Feature Engineering & Model Preparation
5	**Nada Mohammed Abdelaleem Mohamed**	Model Development & Evaluation
6	**Rowida Khaled Mohamed Redawe**	Documentation & Presentation Design

3. Team Leader

Name: Mohamed Gamal Eldeen Ismail Gamal Eldeen

Responsibilities: Overall project coordination, task distribution, version control management, and final report consolidation.

4. Objectives

1. Perform in-depth exploratory data analysis (EDA) to identify key risk factors influencing oral cancer.
2. Engineer derived features (Predicted LOS, Recovery Days, Diagnosis Date/Year, Age Group) to enhance predictive insights.
3. Develop and compare machine learning models for predicting oral cancer diagnosis and patient recovery outcomes.
4. Visualize trends and correlations through interactive dashboards for medical and policy insights.
5. Ensure reproducibility and clarity for both technical and non-technical audiences.

5. Tools & Technologies

Category	Tools / Libraries
Programming Language	Python
Development Environment	Jupyter Notebook
Data Analysis & Preprocessing	Pandas, NumPy
Visualization	Matplotlib, Seaborn, Plotly, Dash
Modeling (Planned)	Scikit-learn, XGBoost, RandomForest
Version Control	GitHub
Documentation & Reporting	MS Word, PDF, PowerPoint

6. Milestones & Deadlines

Milestone	Description	Due Date
Milestone 1	Data Collection, Exploration & Preprocessing	27 September 2025
Milestone 2	Advanced Data Analysis, Visualization & Feature Engineering	18 October 2025
Milestone 3	Model Development & Optimization	8 November 2025
Milestone 4	ML Ops, Deployment & Monitoring	25 November 2025
Milestone 5	Final Documentation & Presentation	29 December 2025

7. KPIs (Key Performance Indicators)

Data Quality

- Percentage of missing values handled: 2 %
- Data accuracy after preprocessing: 99 %
- Dataset diversity (representation of different categories): 85 %

Model Performance (Not Prepared)

- Model accuracy (Accuracy/F1-Score): ....%
- Model prediction speed (Latency): .... milliseconds
- Error rate (False Positive/False Negative Rate): ....%

Deployment & Scalability (Not Prepared)

- API uptime: ....%
- Response time per request: .... milliseconds
- (If applicable) Real-time processing speed: ....

Business Impact & Practical Use

- Reduction in manual effort: 40 %
- Expected cost savings: 25 %
- User satisfaction: 85 %

Notes for Submission:

- **The proposal aligns Milestone-1 findings and prepares the foundation for model development in Milestone-3.**