**DSML Group Project Report**  
**Group 01**  
**Title: Shaping the Future of Colorectal Cancer Survival by Advancing Insights Today**

**Abstract**

Colorectal cancer poses a serious global health threat and is among the top three most prevalent cancers worldwide. In this project, we developed a binary classification model to predict the survival of patients based on diverse clinical and lifestyle features. The process involved thorough preprocessing, exploratory data analysis (EDA), and machine learning model development. We tested multiple classification algorithms including Logistic Regression, Random Forest, and Gradient Boosting. Gradient Boosting achieved the highest weighted F1 score and was selected as the final model. We also performed feature importance analysis and evaluated misclassifications. This study demonstrates the utility of machine learning in improving predictive healthcare models.

**1. Introduction**

Colorectal cancer is a significant cause of morbidity and mortality globally. The ability to predict survival based on patient data can aid in resource allocation and treatment planning. Prior research has identified features such as age, tumor size, and early detection as impactful on survival. In this project, we aim to leverage machine learning to predict patient survival and identify the most influential factors using available data. Our approach includes comparing several classification models to find the most effective one.

**2. Data Exploration and Preprocessing**

**Data Overview:**

* Train Set: 75,035 records with 32 columns
* Test Set: 75,000 records with 31 columns

**Preprocessing Steps:**

* Extracted Age from Date of Birth
* Handled missing values using median/mode imputation
* Dropped fully missing columns (marital status, transfusion history)
* Label encoded categorical variables
* Standardized numeric columns

**EDA Insights:**

* Target class was relatively balanced
* Age distribution concentrated around 69 years
* No strong linear correlation between numerical features and target

**3. Binary Classification**

**Train-Test Split:** 80/20 split using stratified sampling

**Models Compared:**

|  |  |  |
| --- | --- | --- |
| **Model** | **F1 Score (CV Mean)** | **F1 Score (CV Std)** |
| Logistic Regression | ~0.4508 | ~0.0000 |
| Random Forest | ~0.4844 | ~0.0029 |
| Gradient Boosting | **~0.4520** | ~0.0005 |

**Final Model:** Gradient Boosting

* Achieved best performance in cross-validation
* Weighted F1 Score on validation: ~0.76
* High precision and recall for both classes

**Evaluation Metric:** Weighted F1 Score

* Reason: Accounts for class balance and performance on both classes

**4. Open-Ended Section**

**Objectives:** Gain insights into the model and improve interpretability

**Feature Importance:**

* Top Features: Cancer Stage, Tumor Size, Early Detection, Age, Treatment Type
* Confirmed medically relevant features

**Misclassification Analysis:**

* Some older patients misclassified
* Behavioral factors like smoking/alcohol had lower influence

**Probability Calibration:**

* Gradient Boosting provided well-calibrated probabilities
* Future improvement can include Platt scaling or isotonic regression

**5. Conclusion**

This project effectively demonstrates the application of machine learning in predicting colorectal cancer survival. Gradient Boosting emerged as the best-performing model, identifying key survival-related features. The report clearly shows model selection based on empirical evidence and explains the rationale behind each decision.

**Limitations:**

* Lack of longitudinal or treatment response data
* No genetic-level features available