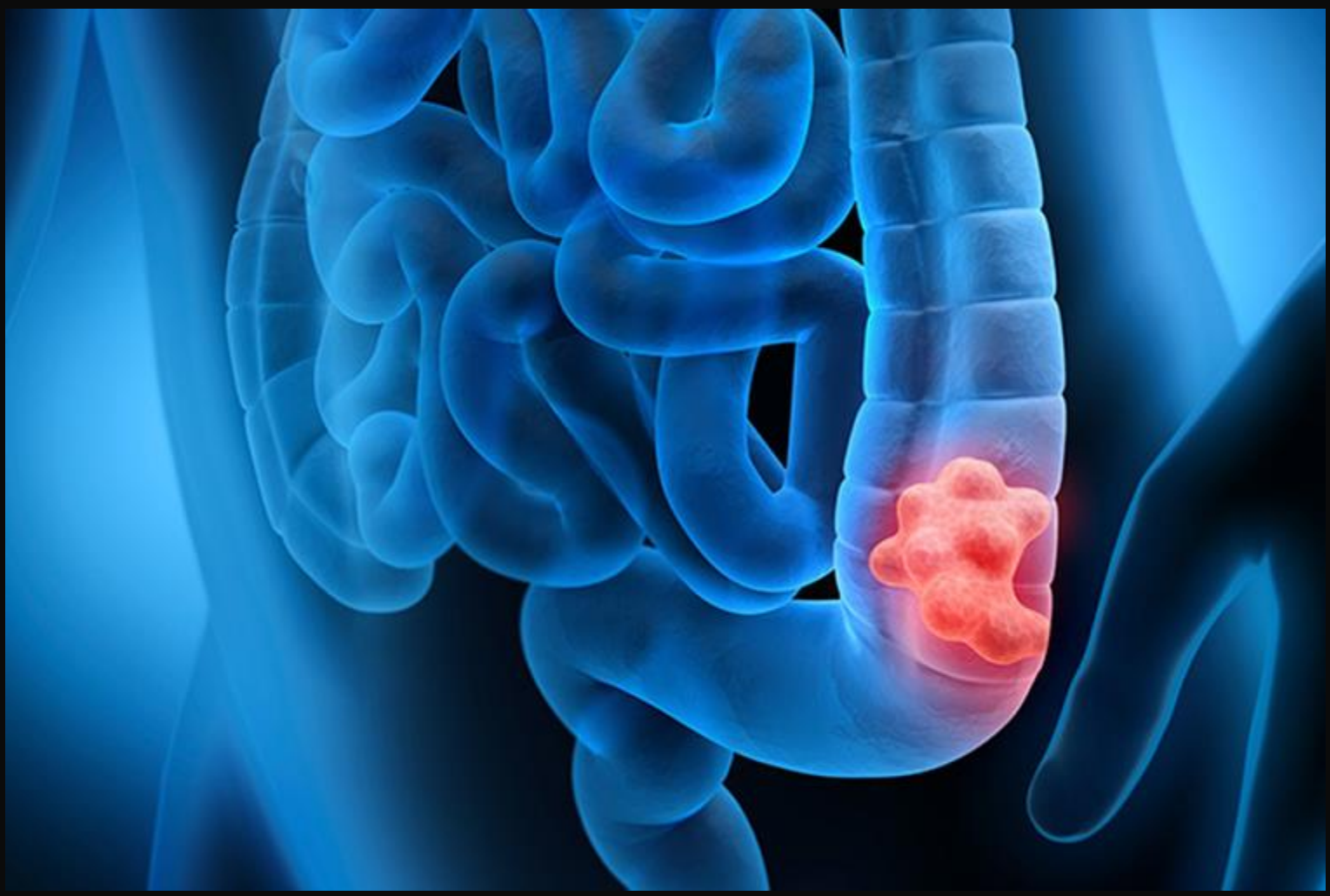
MARCH, 2025

SHAPING THE FUTURE OF COLORECTAL CANCER SURVIVAL BY ADVANCING INSIGHTS TODAY

GROUP PROJECT
DATA MINING II 2024/2025



01

I. INTRODUCTION

Colorectal cancer is a major global health problem, affecting millions and causing significant morbidity and mortality. Studying it is crucial because it is the **third most common cancer worldwide**, and understanding its survival outcomes can save lives. By developing predictive models for colorectal cancer survival, we can identify key risk factors and improve early detection and treatment strategies. **This research can have a profound impact on patient care**, leading to enhanced survival rates and optimized resource allocation. The ultimate benefit is to reduce the burden of colorectal cancer and improve the quality of life for patients globally.

II. PROJECT GOALS

The primary components and goals of the project are:

1. **Preprocessing and Exploratory Data Analysis (EDA):** Clean and preprocess your dataset, then perform an exploratory analysis to uncover patterns and relationships.
2. **Binary Classification:** Create a classification model to accurately predict if a certain patient is able or not to survive. To do that, you will need to **develop a consistent model assessment** strategy to compare different candidate models, optimize them, and find the most generalizable one. Note: Don't forget to report all models and discuss the findings.
3. **Kaggle Competition:** Teams can submit multiple predictions on Kaggle, with the scoreboard ranking submissions based on **F1 Score 'weighted'**. Before the competition ends, you must choose one submission to compete.
4. **Additional Insights:** This component is open-ended, allowing you to explore a wide range of ideas of your preference, as long as it relates with the topic/dataset, and you can clearly communicate them. Here are some suggestions:
    a. Analyze and discuss the importance of the features.
    b. Analyze and discuss the incorrect predictions.
    c. Analyze the probability of your predictions and the calibration of the model.
    d. Interpret the model's predictions.
    e. Integrate your dataset with data from other sources to generate new insights.

02

III. DATASET

You have access to two different datasets:

In the **training set**, you will find features that give detailed information about each patient. Use this training data and its features to build and validate your machine-learning models. The goal is to apply the models you have created to make predictions on unseen data (i.e., the test set). **Important note: You should not consider the target variable as feature for any of the predictive models.**

In the **test set**, you will still have access to the same attributes. However, the target variable, which you are trying to predict, will not be available.

The available data contains the following attributes:

ATTRIBUTE	DESCRIPTION
ID	Patient unique identifier
Country	Country of origin
Date of Birth	Patient date of birth
Gender	Patient gender (M / F)
Cancer Stage	Cancer stage (Localized, Regional, Metastatic)
Tumor Size (mm)	Tumor size in mm
Family History	Family history of similar cancer (Yes / No)
Smoking History	Smoker (Yes / No)
Alcohol Consumption	Regular consumer of alcohol (Yes / No)
Obesity BMI	BMI classification (Normal / Overweight / Obese)
Diet Risk	Dietary risk classification (Low / Moderate / High)
Physical Activity	Physical activity level (Low / Moderate / High)
Diabetes	Has diabetes (Yes / No)
Diabetes History	Has a past history of diabetes (Yes / No)

03

ATTRIBUTE	DESCRIPTION
Inflammatory Bowel Disease	Has Inflammatory Bowel Disease (Yes / No)
Genetic Mutation	Observed genetic mutation (Yes / No)
Hypertension	Has Hypertension (Yes / No)
Screening History	Cancer screening frequency (Regular / Irregular / Never)
Transfusion History	Patient receives blood transfusions
Heart Disease History	Has an history of heart diseases (Yes / No)
Non Smoker	Patient doesn't smoke (Yes / No)
Early Detection	Early detection (Yes / No)
Treatment Type	Treatment (Surgery / Chemotherapy / Radiotherapy / Combination)
Healthcare Costs	Healthcare costs
Incidence Rate per 100K	Country incident rate at the time (new colorectal cancer cases diagnosed per 100,000 people)
Mortality Rate per 100K	Country mortality rate ate the time (number of deaths caused by colorectal cancer per 100,000 people)
Urban or Rural	Residence area (Urban / Rural)
Marital Status	Marital status of the patient
Healthcare Access	Healthcare availability (Low / Moderate / High)
Insurance Costs	Cost of insurance (No insurance / Basic / Extended)
Insurance Status	Insurance status (Insured / Uninsured)
Survival Prediction	Survival prediction of the patient **(Target Variable)**

04

IV. OUTLINE

Your project report, written in English, should **respect the following outline and format:**

Abstract
Provide a small overview of your work (200 to 300 words): What is the context? What are your goals? What did you do? What were your main results, and what conclusions did you draw from them?

I. Introduction
- Overview of the project
- Main goals of the project
- Are there any similar works? What has been done? What did other researchers find? What would you expect your results to be based on their previous findings?

II. Data Exploration and Preprocessing
- Description of data received -> key insights
- Steps taken to clean and prepare the data

III. Binary Classification
- Additional preprocessing steps adopted
- Feature selection strategy
- Explanation of model assessment strategy and metrics used
- Comparison of performance between candidate models
- Optimization efforts, results and discussion

IV. Open-Ended Section
- Objectives for the section
- Description of the actions taken
- Results and discussion of main findings

V. Conclusion
- Summary of objectives and findings
- Do the findings match what you initially expected? How?
- Discussion of limitations of your work
- Suggestions for possible work to follow on your work

Report settings:
- Heading: Calibri, size 14 pt, in bold
- Text: Calibri, size 11 pt, line spacing 1.15 pt and paragraph spacing of 6 pt

V. DELIVERABLES

Upon the project's deadline, you will be required to submit:

- A Jupyter **notebook (or zip file) containing all the code** used throughout the project **and any extra data you may have used**. Make sure to include detailed markdown cells and comments within the notebook to guide readers through the code. Clearly explain the purpose of each code segment, the insights gained, and the decisions made.
  - Name your file in the format ***XT_DM2_GroupXX_Notebook***, where ***XT* should be *DT* or *NT***, depending if you are from Daytime or Nighttime classes, and ***GroupXX*** should be your group number.

- A **report** that describes the analytical processes as outlined in chapter IV. It can't be longer than **15 pages** (every page counts). The body of text should only include figures and tables that are essential to understand the work. Supporting figures and Tables can be added to annexes but must be referenced in the text.
  - Name your file in the format ***XT_DM2_GroupXX_Report.pdf***, like the notebook.

VI. EVALUATION

Your work will be evaluated according to the following criteria:

CRITERIA	PERCENTAGE (%)	MAXIMUM GRADE (OUT OF 20)
Notebook Explanation	10	2
Preprocessing and EDA	10	2
Modeling and Optimization	30	6
Model performance on Kaggle	10	2
Open-Ended Section	20	4
Report Quality and Storytelling	10	2
Discussion (In-Person)	10	2

06

Your grade will reflect our assessment of the quality of your work in terms of quality of writing, clarity, conciseness, correctness and efficiency. Please find below more details about what is taken for each topic:

- **Notebook Explanation:** Assess notebook clarity and readability, including whether the goals are clearly stated, the code contains comments, the techniques are appropriate, the insights and conclusions are highlighted, and the implications for further steps.
- **Preprocessing and EDA:** Describe the dataset and extract meaningful insights that you consider relevant to the problem. Avoid adding unnecessary visualizations or elements. This section also addresses the initial preprocessing of the data, providing a clear explanation of each step taken and the rationale for your choice.
- **Modeling and Optimization:** Describe your strategy for the classification objective. This section is separated into different components:
  - Additional preprocessing
  - Feature selection
  - Modelling approach – assessment strategy (holdout, cross-validation, etc.) and the algorithms explored
  - Performance assessment – rationale for choice of evaluation metric(s) and interpretation of results
  - Model optimization
- **Model performance on Kaggle:** Grading based on groups ranking.
- **Open-Ended Section:** Describe your strategy for the additional insights objective. This section is separated into different components:
  - Formulation and adequacy
  - Difficulty
  - Correctness/efficiency of implementation
  - Explanation and discussion of results
- **Report Quality and Storytelling:** A good report provides a clear understanding of the problem, the steps taken and its rationale, linked to the main results and insights. When referencing a figure, clearly highlight the point you want to convey. This section also evaluates the overall quality of your introduction and conclusions.
- **Discussion:** This section evaluates the effectiveness of each individual's ability to explain the project and answer confidently to questions regarding data, methods, insights, conclusions, decisions, etc.

07

VII. PARTING NOTES

- Deliveries after the deadline will be penalized at 1 point per day.
- Deliveries made before the deadline will receive a bonus of 0.15 points per day of delivery in advance (up to a maximum of 1 point).
- For modelling purposes, any **algorithm implementation outside the vanilla scikit-learn is explicitly off-limits** and will result in a 1-point penalty.
- The report will be the primary method of evaluating your work. When preparing it, remember that a **reader should be able to understand your work without needing to check your notebook**. We won't be able to consider any steps or results not mentioned in your report.
- Ensure your report is concise, focused and based on reliable sources. You should look to source information provided from peer-reviewed journals (thus, avoid citing Medium, TowardsDataScience and similar sources). Avoid irrelevant, unimportant, or redundant information. Don't provide theoretical explanations of topics covered in class.
- Before submitting, run your notebook from the start one last time (if you used a GridSearch, you can comment this cell, but you should run the final model with the GS parameters in a different cell).
- **Your submitted notebook should include all the unneeded code you used to obtain your final solution, but it should also be commented**.
- We will run your Jupyter Notebooks if we have any doubts. So, please **make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.**
- The report and code will pass through a process of plagiarism and AI generation checking.
- You must submit your predictions on the Kaggle competition to get points for that component.
- When determining the grade for your work, there will be a comparative component between it and the work presented by your peers.

- **Friendly Reminders:**
  - Attendance at the **defense is mandatory for approval** in the project. The defense has a group component and an individual component.
  - **Do not include techniques/algorithms/steps you cannot explain** in your report: we will ask about them in the defense.
  - If a team member is not contributing to the project, you must report it at least one month before submission date.