# **Predicting 5-year survival among Colorectal Cancer Patients**

# **Project Proposal**

# **Submitted to**

Hamed Zolbanin  
Assistant Professor  
University of Dayton  
 hzolbanin1@udayton.edu

# **By**

# The University of Dayton BAN 710 Project Team:

Ayush Shrivastava  
Ashutosh Gunvantrao Masulkar  
Rishikesh Rajesh Mankar  
Shantanu Chandrakant Dahiphale  
Rahul Kumar Swami

Table of Contents

[1. Background 3](#_Toc196226415)

[1.1 Business Unit 3](#_Toc196226416)

[2. Problem Description 3](#_Toc196226417)

[3. Approach 4](#_Toc196226418)

[Model Selection 4](#_Toc196226419)

[Planned Steps: 4](#_Toc196226420)

[3.1 Scope 4](#_Toc196226421)

[4. (Anticipated) Results 5](#_Toc196226422)

[5. Benefits 5](#_Toc196226423)

[6. Risks 5](#_Toc196226424)

[7. Statement of Work 5](#_Toc196226425)

[8. Resources and Costs 6](#_Toc196226426)

[9. Contact Information 6](#_Toc196226427)

[10. References 6](#_Toc196226428)

# **Predicting 5-year survival among Colorectal Cancer Patients**

***Executive Summary***

*This project proposes developing a binary classification machine learning model to predict whether colorectal cancer patients will survive more than 60 months (5 years), using data from the SEER database. The goal is to accurately classify patients as either surviving beyond 5 years (=1) or not (=0), providing clinically meaningful outcomes to aid decision-making. We ensure the use of predictive evaluation metrics such as AUC, accuracy, precision, and recall and also b*y offering transparent feature importance outputs, the project will assist clinicians, researchers, and healthcare administrators in better resource allocation and care personalization. This proposal outlines the business background, the clinical problem, proposed approach, anticipated results, risks, and next steps.

## **1. Background**

### One of the leading causes of cancer death globally is colorectal cancer. Precisely forecasting long-term survival enhances health resource management, follow-up planning, and patient counseling. The SEER database is an effective tool for predictive modeling since it offers comprehensive clinical, pathological, and demographic data. It introduces a data-driven approach to assess the likelihood of 5-year survival in colorectal cancer patients, aiming to improve clinical decision-making through personalized risk profiling.

### **1.1 Business Unit**

This capstone project, part of the University of Dayton MBAN Course, intends to demonstrate the convergence of healthcare and data science by developing a clinically effective model that supports individualized care decisions. The goal is to deliver a machine learning model that enhances healthcare decision-making through prediction, showcasing the integration of data science and clinical applications.

## **2. Problem Description**

Staging approaches for colorectal cancer sometimes ignore patient-specific characteristics, leading erroneous projections of overall survival. Despite significant variations in age, tumor activity, and response to treatment, several algorithms might classify patients similarly. By creating a machine learning classification model that uses unique clinical, demographic, and treatment data to forecast if a patient with colorectal cancer will live past 60 months (5 years), this study fills that knowledge gap. The approach will provide patient-specific risk classification, allowing for more precise and individualized oncology care decisions.

## **3. Approach**

### We will apply machine learning classification models to predict the binary outcome of 5-year survival based on a derived label from the "Survival Months" column (1 if ≥ 60 months, 0 otherwise).

### **Model Selection:**

|  |  |
| --- | --- |
| **Model** | **Reason for Selection** |
| XGBoost Regressor | High accuracy, robust missing data handling, widely trusted in healthcare ML. |
| LightGBM Regressor | Fast training, efficient handling of large tabular datasets. |
| Random Forest Regressor | Baseline model, robust to overfitting, interpretable feature importance. |
| CatBoost Classifier | Handles categorical features natively. |
| Logistic Regression | Simple and strong for benchmarking. |

### 

### **Planned Steps:**

* **Data Cleaning**: Handle missing values, encode categorical variables, and filter colorectal cancer patients.
* **Exploratory Data Analysis (EDA)**: Explore distributions, identify outliers, and visualize relationships with binary survival outcome.
* **Feature Preparation**: Select important clinical, demographic, and treatment variables.
* **Model Training**: Train multiple classification models using cross-validation and tune hyperparameters.
* **Model Evaluation**: It will be assessed using classification metrics such as AUC, Accuracy, precision, recall and confusion matrix.
* **Feature Importance Analysis**: To interpret model predictions and identify key survival predictors.
* **Model Selection**: Compare models and select the best based on evaluation metrics and clinical relevance.

***3.1 Scope***

**Included**:

* Binary classification of 5-year survival outcome.
* SEER colorectal cancer cases.
* Utilizing clinical, demographic, and treatment-related features for model training and feature engineering.

**Excluded**:

* Time series forecasting methods are outside the scope of this project.
* No external datasets beyond the SEER database will be used during the initial modeling phase.

## **4. (Anticipated) Results**

* Achieve a **AUC** > 0.80 1, for a substantial class separation.
* Achieve **Accuracy > 0.75 ,** indicating the overall correctness of the prediction.
* **Recall** > 0.85 to ensure high-risk patients (non-survivors) are correctly identified.
* Target an **R-squared (R²)** value above 0.60, indicating a strong proportion of variance in binary outcome explained by the model.
* Confusion Matrix to show correct and incorrect predictions for both survival classes.

## **5. Benefits**

* Enables early identification of high-risk patients.
* Supports more personalized treatment decision-making.
* Optimizes hospital resource management.
* Increases the long-term prognosis's accuracy.

## **6. Risks and Strategies**

* Incomplete or missing clinical data – Filtering irrelevant features through imputation techniques (mean/mode or KNN).
* Risk of model overfitting – Applying L1/L2 regularization, cross-validation and model comparison.
* Potential bias if specific subgroups (e.g., minorities) are underrepresented - Conduct subgroup performance analysis and apply technique such as stratified sampling.

## **7. Statement of Work**

|  |  |
| --- | --- |
| **Phase** | **Activity** |
| Phase 1 | Data Cleaning and Preprocessing |
| Phase 2 | Exploratory Data Analysis (EDA) and Feature Engineering |
| Phase 3 | Model Training, Hyperparameter Tuning, and Validation |
| Phase 4 | Results Interpretation and Analysis |
| Phase 5 | Final Report Preparation and Client Presentation |

## 

## **8. Resources and Costs**

* Dataset: SEER Colorectal Cancer Data
* Tools: Python (Pandas, Scikit-Learn, XGBoost, LightGBM, etc.)
* Computing: Google Collab / Local Machine
* Estimated Additional Costs: None anticipated

## **9. Contact Information**

* Ayush Shrivastava [shrivastavaa5@udayton.edu](mailto:shrivastavaa5@udayton.edu)
* Ashutosh Gunvantrao Masulkar  [masulkara1@udayton.edu](mailto:masulkara1@udayton.edu)
* Rishikesh Rajesh Mankar [mankarr1@udayton.edu](mailto:mankarr1@udayton.edu)
* Shantanu Chandrakant Dahiphale [dahiphales1@udayton.edu](mailto:dahiphales1@udayton.edu)
* Rahul Kumar Swami   [swamir2@udayton.edu](mailto:swamir2@udayton.edu)

## **10. References**

* Surveillance, Epidemiology, and End Results (SEER) Program. [SEER Website](https://seer.cancer.gov/)
* Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. [Link](https://arxiv.org/abs/1603.02754)
* Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. [Link](https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)