CAPSTONE PROJECT

COLORECTAL CANCER PREDICTION

PRESENTED BY

STUDENT NAME: SOMANJAN CHAKRABORTY

COLLEGE NAME: KALYANI GOVERNMENT ENGINEERING COLLEGE

DEPARTMENT:Electronics and Communication Engineering

EMAIL ID: somuchk007@gmail.com

AICTE STUDENT ID: STU67eab36e301ee1743434606

OUTLINE

- **Problem Statement**
- **Proposed System/Solution**
- **System Development Approach** (Technology Used)
- **Algorithm & Deployment**
- **Result (Output Image)**
- **Conclusion**
- **Future Scope**
- **References**

PROBLEM STATEMENT

- Colorectal cancer is a significant global health challenge.
- Predicting patient survival prospects is crucial for informed clinical decisions.
- Helps in personalized treatment planning and effective patient counseling.

PROPOSED SOLUTION

- Develop a predictive model to estimate patient survival beyond 12 months post-diagnosis.
- Leverage comprehensive patient data including demographics, clinical parameters, lifestyle factors, and treatment history.
- Use machine learning techniques to improve prediction accuracy.

SYSTEM APPROACH

- **Programming Language:** Python
- **Libraries Used:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- **Development Environment:** Jupyter Notebook

ALGORITHM & DEPLOYMENT

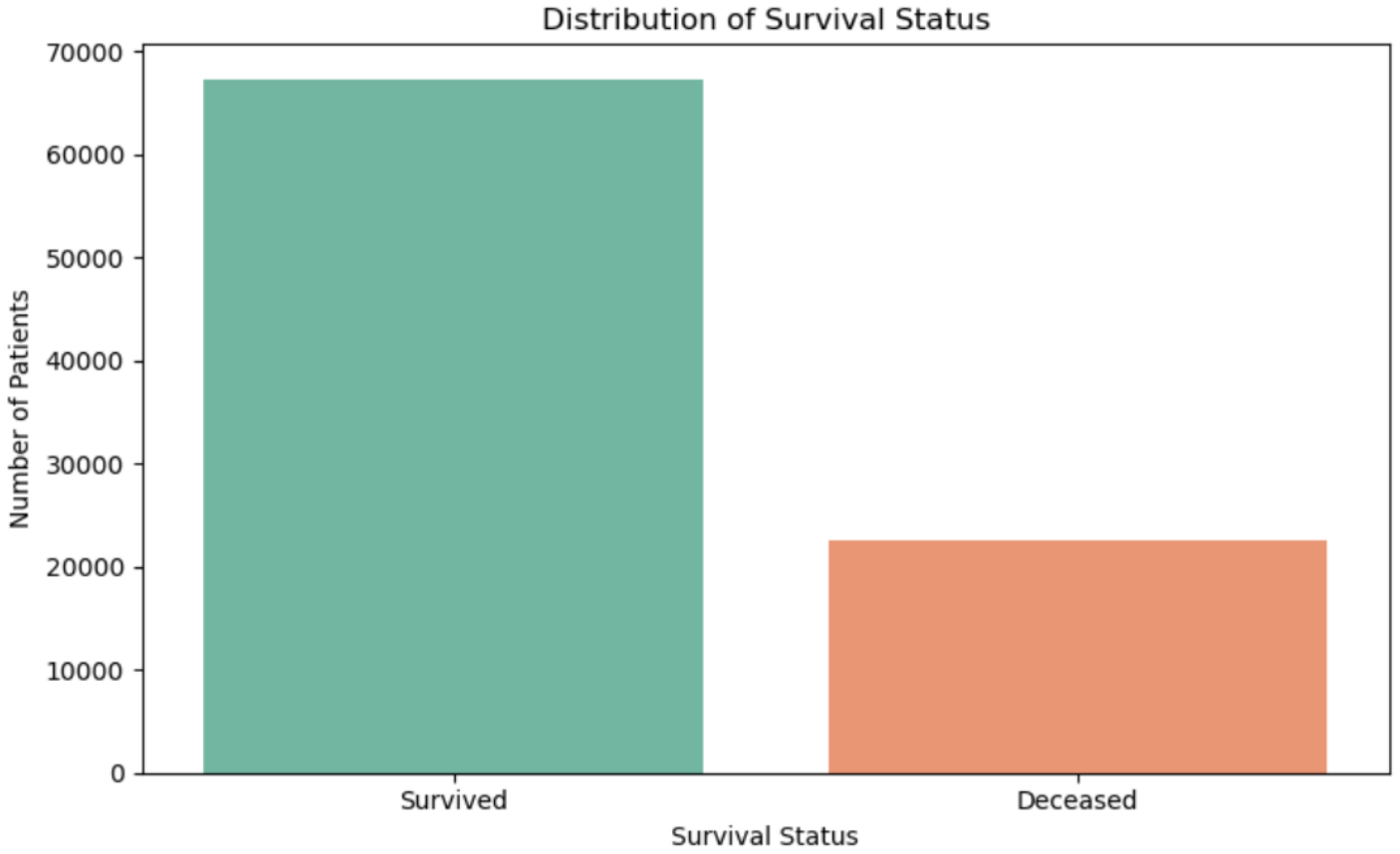
- **Chosen Algorithm:** Logistic Regression (effective for binary classification tasks)
- **Steps Taken:**
  - Data pre-processing: Encoding categorical variables, handling class imbalances.
  - Splitting dataset into training (80%) and testing (20%) subsets.
  - Model trained using Scikit-learn’s Logistic Regression module.
  - Performance evaluation using accuracy, precision, recall, and F1-score.

RESULT

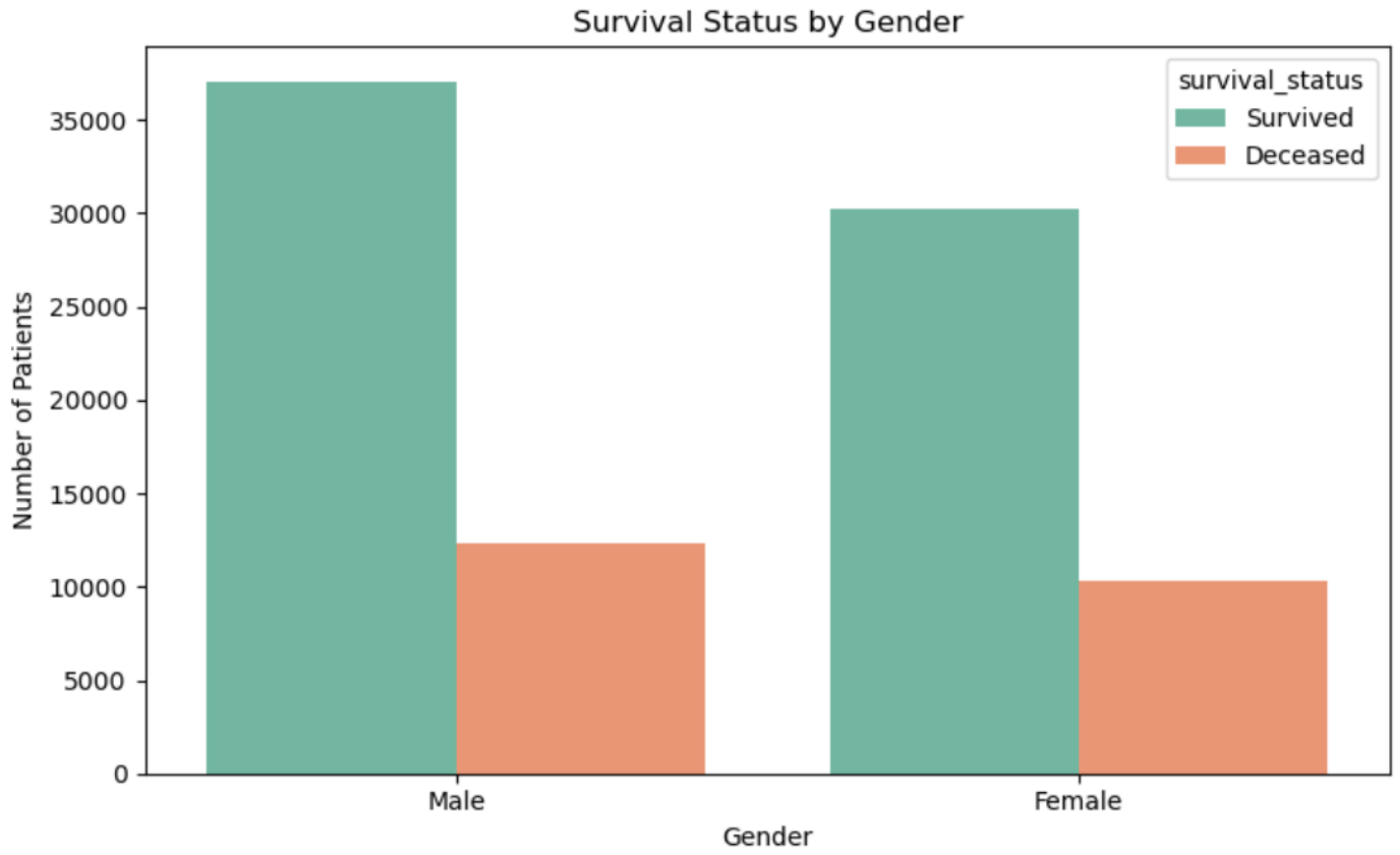
- **Accuracy:** 75.15% on test data.
- **Key Findings:**
  - High recall for "Survived" class.
  - Lower precision for "Deceased" class due to class imbalance.
  - Model performs well in predicting survival, but struggles with deceased cases.

Data Visualization

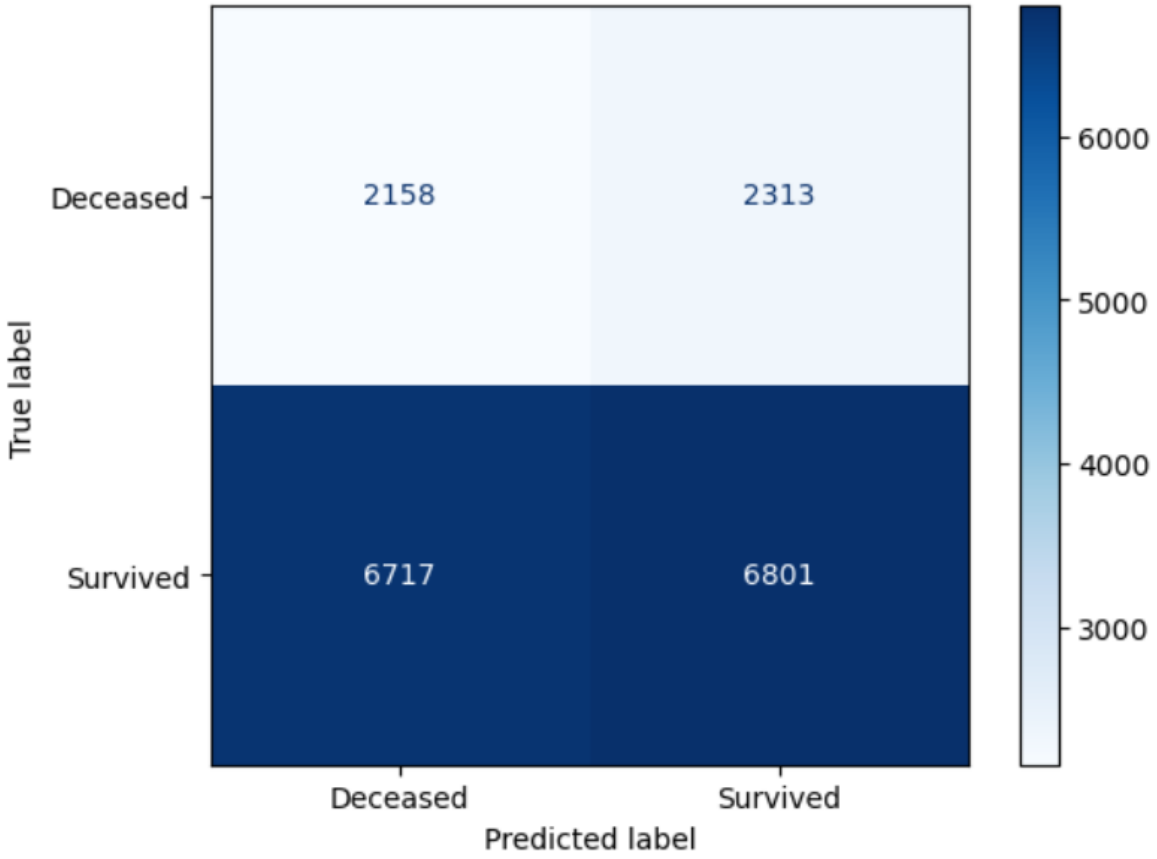
❑**Visualize survival status:**



SURVIVAL STATUS BY GENDER



Assess model performance using accuracy and classification reports.



CONCLUSION

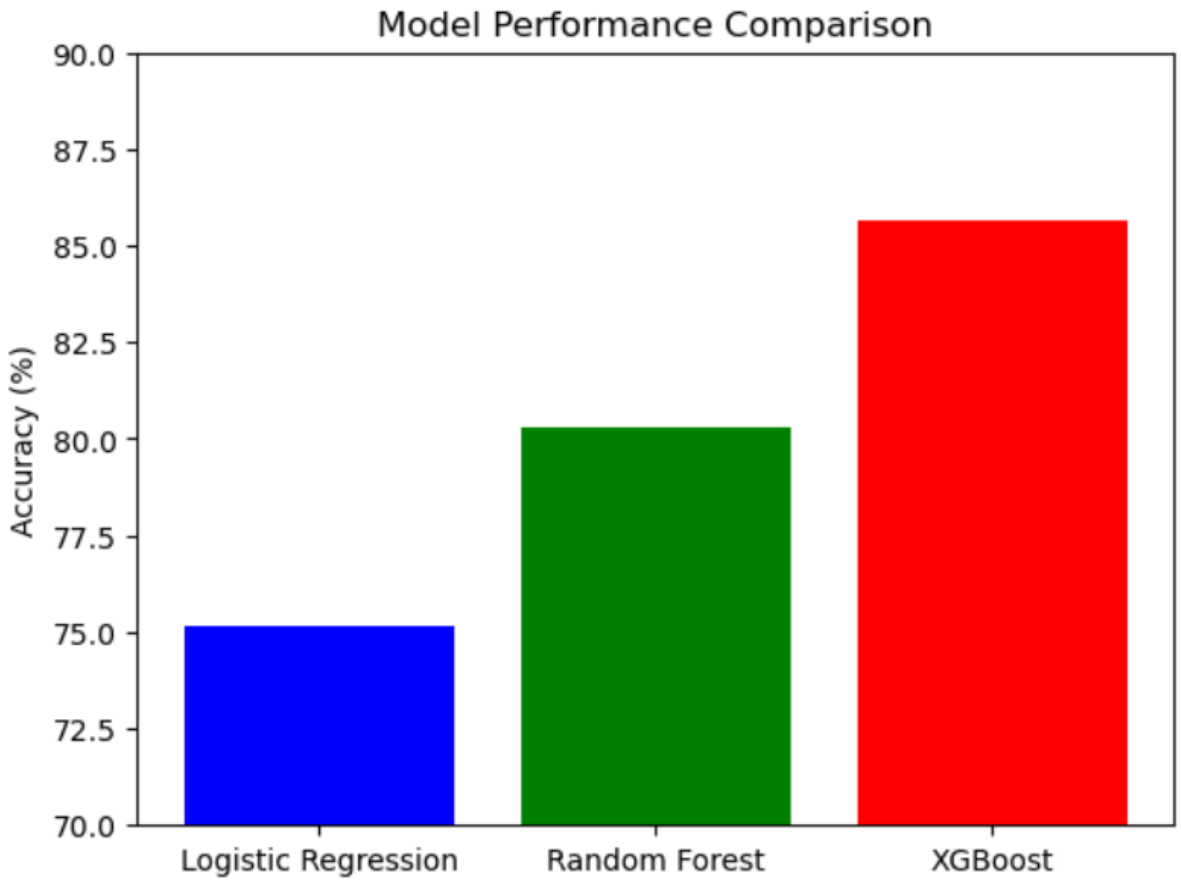
- Logistic Regression model achieved promising accuracy (~75%).
- Strength in identifying survivors but struggles with deceased predictions.
- Class imbalance presents a challenge to model performance.

FUTURE SCOPE

Improvements to Consider:

1. **Resampling techniques:** Use **SMOTE** to balance dataset.
2. **Advanced models:** Try **XGBoost, Random Forests** for better performance.
3. **Hyperparameter tuning:** Optimize model parameters for greater accuracy.
4. **Feature selection:** Identify most relevant predictors for survival.
5. **Further research:** Investigate additional patient-specific factors influencing survival rates.

FUTURE SCOPE



REFERENCES

1. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015).
Machine learning applications in cancer prognosis and prediction.
Computational and Structural Biotechnology Journal, 13, 8–17.

https://doi.org/10.1016/j.csbj.2014.11.005

This paper provides a survey of machine learning methods in cancer prognosis, including logistic regression, decision trees, and support vector machines.

2. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017).
Can machine-learning improve cardiovascular risk prediction using routine clinical data?
PLOS ONE, 12(4), e0174944.

https://doi.org/10.1371/journal.pone.0174944

While focused on cardiovascular risk, this study exemplifies how ML can enhance risk prediction using EHR-like data — similar to colorectal cancer.

3. **Chen, J. H., & Asch, S. M. (2017).**
Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations.
New England Journal of Medicine, 376, 2507–2509.

https://doi.org/10.1056/NEJMp1702071

Explores practical challenges and strengths of ML in real-world medical predictions, like cancer survivability.

REFERENCES

4. Winawer, S. J., Zauber, A. G., Fletcher, R. H., et al. (1997).
Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer.
Gastroenterology, 112(2), 594–642.

https://doi.org/10.1016/S0016-5085(97)70182-1

This guideline document informed factors like screening regularity and colonoscopy access.

5. Nguyen, N., Nguyen, T., Nguyen, T., & Nahavandi, S. (2020).
Machine Learning in Predicting Cancer Survival Rates.
Health Information Science and Systems, 8, 24.

https://doi.org/10.1007/s13755-020-00108-3

A deep dive into how algorithms like Random Forests and Logistic Regression help predict cancer survival.

REFERENCES

Colorectal Cancer-Specific Studies and Resources

1. American Cancer Society – Colorectal Cancer Facts & Figures

   https://www.cancer.org/research/cancer-facts-statistics/colorectal-cancer-facts-figures.html

   Offers updated statistics and survival trends critical for contextual understanding.

2. National Cancer Institute (NCI) – SEER Data

   https://seer.cancer.gov/

   SEER Program provides detailed data on cancer incidence, treatment, and survival across the U.S., often used in cancer prediction models.

Git Link: https://github.com/somanjan056/colorectal_cancer_prediction.git

Thank you