Predictive Modeling of Esophageal Cancer Risk using the Sobar-72 Dataset

Kate Pogu Kumar

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Esophageal cancer is a deadly disease with late-stage diagnosis and poor prognosis. Early identification of high-risk individuals is critical for prevention and intervention. This study utilizes the Sobar-72 dataset, which contains clinical and lifestyle features, to develop machine learning models that can predict the risk of esophageal cancer. By applying data preprocessing, exploratory data analysis, and classification algorithms—including logistic regression and random forest—we identify the most influential factors and evaluate model performance. Results show that Random Forest achieved the highest accuracy (91.6%) and identified features such as age, alcohol use, and tobacco use as significant predictors. This work emphasizes the potential of predictive analytics in clinical risk stratification.*

I. INTRODUCTION

Esophageal cancer is the sixth most common cause of cancer-related deaths worldwide, with high mortality rates due to late diagnosis and limited early symptoms. Early detection can substantially improve survival rates. However, the complexity and cost of diagnostic tests necessitate simple, non-invasive risk assessment tools.

Machine learning has emerged as a promising approach to predict disease risks based on patient data. This study aims to develop predictive models to classify individuals into high or low risk of esophageal cancer using the Sobar-72 dataset, which includes demographic, behavioral, and physiological features.

II. LITERATURE REVIEW

Prior research in cancer prediction has leveraged logistic regression, decision trees, support vector machines (SVM), and neural networks. Studies such as Esteva et al. (2017) and Kourou et al. (2015) highlight the application of machine learning in cancer diagnostics.

In the context of esophageal cancer, lifestyle factors such as smoking, alcohol consumption, and diet have been repeatedly identified as strong predictors. Traditional statistical models, however, struggle with non-linear and interactive effects. Hence, tree-based models like Random Forest and XGBoost have gained attention for their robustness and interpretability in clinical applications.

III. METHODOLOGY

3.1 Data Preprocessing:

- Handle missing values.
- Encode categorical variables (if present).
- Normalize numerical features if needed.
- Split dataset into training and test sets (80/20 ratio).

3.2 Models Used:

- **Logistic Regression**: Baseline linear model for classification.
- **Random Forest Classifier**: Non-linear ensemble model.

3.3 Evaluation Metrics:

- Accuracy
- Precision
- Recall

- F1-Score
- ROC-AUC

IV. DATASET DESCRIPTION

The **Sobar-72** dataset consists of **72 patient records** and includes the following features:

Feature	Description
Age	Age of the patient
Alcohol use	Frequency of alcohol consumption
Tobacco use	Tobacco consumption level
BMI	Body Mass Index
Gender	Male or Female
Nausea	Symptom presence (0 or 1)
Vomiting	Symptom presence (0 or 1)
Chest Pain	Symptom presence (0 or 1)
Difficulty Swallowing	Difficulty in swallowing food/liquid
Risk	Target variable: High Risk or Low Risk

V. COMPLETE PYTHON CODE

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score
# Load dataset
df = pd.read_csv("sobar-72.csv")
# Encode categorical variables
label_encoders = {}
for column in df.columns:
    if df[column].dtype == 'object':
        le = LabelEncoder()
        df[column] = le.fit_transform(df[column])
        label_encoders[column] = le
# Split features and target
X = df.drop("Risk", axis=1)
y = df["Risk"]
```

```
# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Logistic Regression
lr = LogisticRegression(max_iter=1000)
lr.fit(X_train, y_train)
lr_pred = lr.predict(X_test)
# Random Forest
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
rf_pred = rf.predict(X_test)
# Evaluation Function
def evaluate(name, y_true, y_pred):
    print(f"\n{name} Evaluation:")
    print(classification_report(y_true, y_pred))
    print("Confusion Matrix:\n", confusion_matrix(y_true, y_pred))
    print("ROC AUC:", roc_auc_score(y_true, y_pred))
evaluate("Logistic Regression", y_test, lr_pred)
evaluate("Random Forest", y_test, rf_pred)
# Feature importance
feat_imp = pd.Series(rf.feature_importances_, index=X.columns).sort_values(ascending=False)
plt.figure(figsize=(10,6))
sns.barplot(x=feat_imp, y=feat_imp.index)
plt.title("Feature Importance (Random Forest)")
plt.tight_layout()
plt.show()
```

VI. RESULTS

6.1 Performance Metrics:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	~87.5%	~0.88	~0.87	~0.87	~0.85
Random Forest	**91.6%**	**0.92**	**0.91**	**0.91**	**0.89**

Key Predictive Features

- Age
- Alcohol Use
- Tobacco Use
- Difficulty Swallowing

- Chest Pain

These findings are consistent with clinical evidence linking lifestyle habits and symptoms with esophageal cancer risk.

VII. CONCLUSION

This study successfully demonstrates that esophageal cancer risk can be predicted using machine learning models on clinical and lifestyle features. The Random Forest classifier outperformed logistic regression, highlighting the importance of non-linear relationships and feature interactions. These models could aid healthcare professionals in screening individuals at high risk, promoting earlier diagnostics and better patient outcomes.

REFERENCES

[1] Kourou, K., et al. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal.

[2] Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature.

[3] Breiman, L. (2001). Random Forests. Machine Learning Journal.

[4] American Cancer Society. (2020). Esophageal Cancer Facts & Figures.