Machine Learning Model Report

# • Feature Selection and Handling Missing Data:

Selected features were chosen based on their correlation with the target label 'Survival\_Status'. The correlation analysis was done using Spark's DataFrame statistical functions, and the top features selected were:

Age, Smoking\_History, Obesity, Diabetes, Stage\_at\_Diagnosis, Physical\_Activity\_Level, Gender

These features were selected because they showed higher absolute correlation values with 'Survival\_Status'.

# • Assembling Features:

VectorAssembler was used to combine the selected features into a single feature vector column named 'features'. This column was then used for training the model.

# • Model Training and Evaluation:

A Logistic Regression model was trained using the assembled feature vectors. The dataset was split into training and testing sets in a 70:30 ratio.

The model's performance was evaluated using several metrics:

* • Accuracy: 0.874
* • F1-Score: 0.815
* • ROC-AUC: 0.498
* • RMSE: 0.354
* • R²: -0.143

# • Findings

The logistic regression model was built to predict patient survival status using selected features with the highest correlation. The RMSE of 0.354 indicates that the model's predictions deviate slightly from actual values, while the R-squared value of -0.143 suggests that the model doesn't generalize well on this test set. Despite a high accuracy (0.874), the low ROC-AUC indicates class imbalance or poor distinction capability. These results imply that while the model may correctly classify many instances, it lacks strong predictive power for both classes equally. Additional feature engineering or trying more complex models may enhance performance.