Project Description Document

1- Introduction

This project explores two types of datasets, image-based and numeric, to train machine learning models for specific predictive tasks. The goal is to compare the performance of different models in terms of accuracy, precision, recall, and AUC. Key visualizations like loss curves, ROC curves, and confusion matrices are presented to evaluate the models.

2- Datasets

2.1- Image Dataset

- **Dataset Name:** `Cell Images for Detecting Malaria`
- **Number of Classes:** `2 (Parasitized, Uninfected)`
- **Total Number of Samples**: 27,558
- **Image Size**: 64x64 pixels
- **Data Split**:
  - Training: 22046 samples
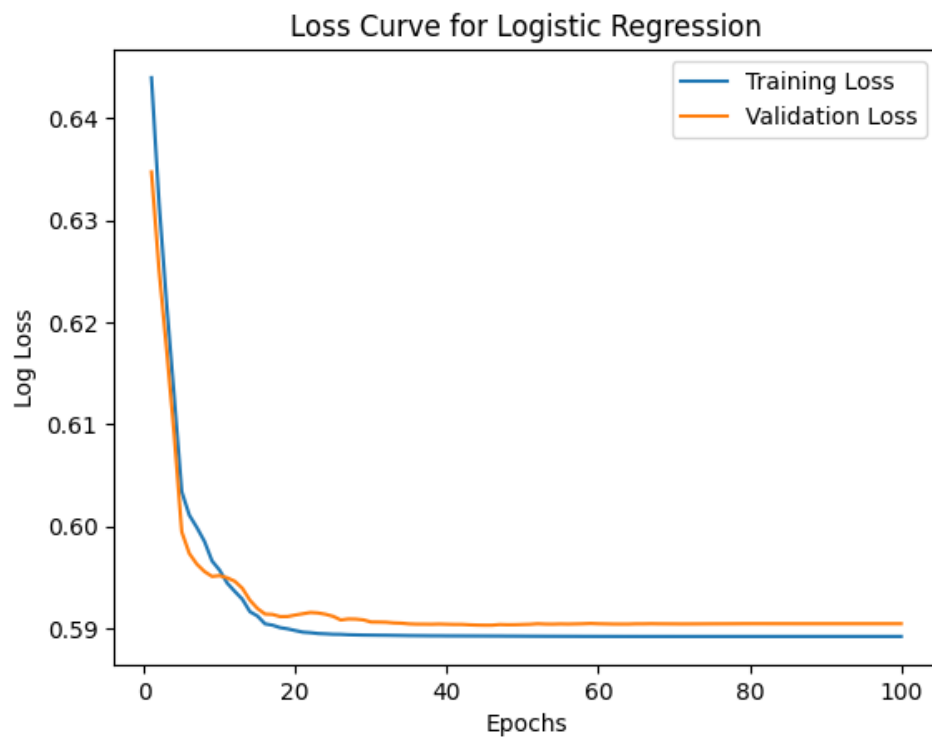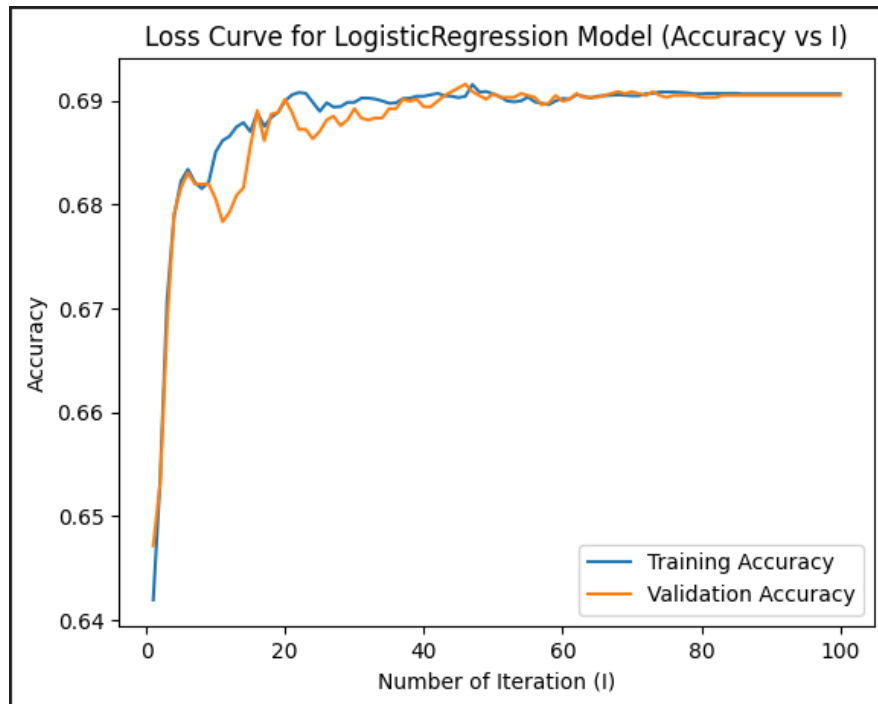 - Testing: 5512 samples

2.2- Numeric Dataset
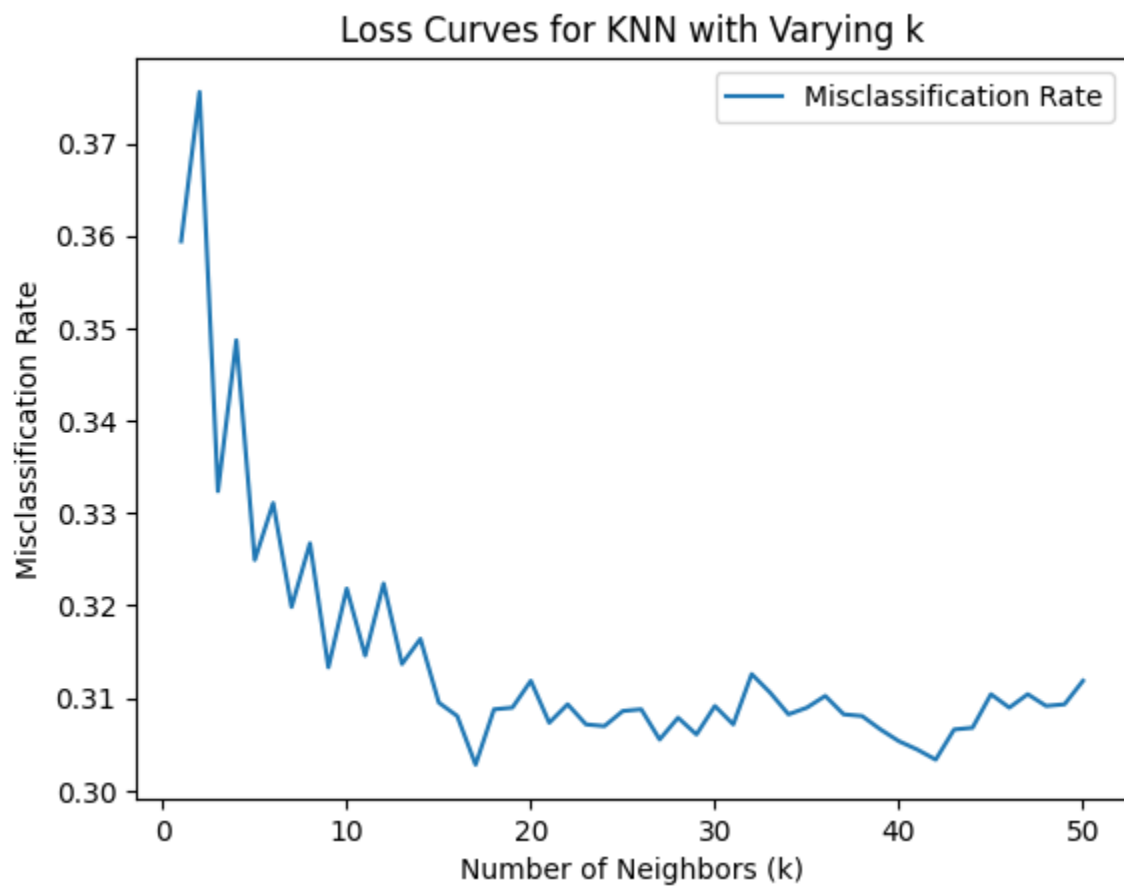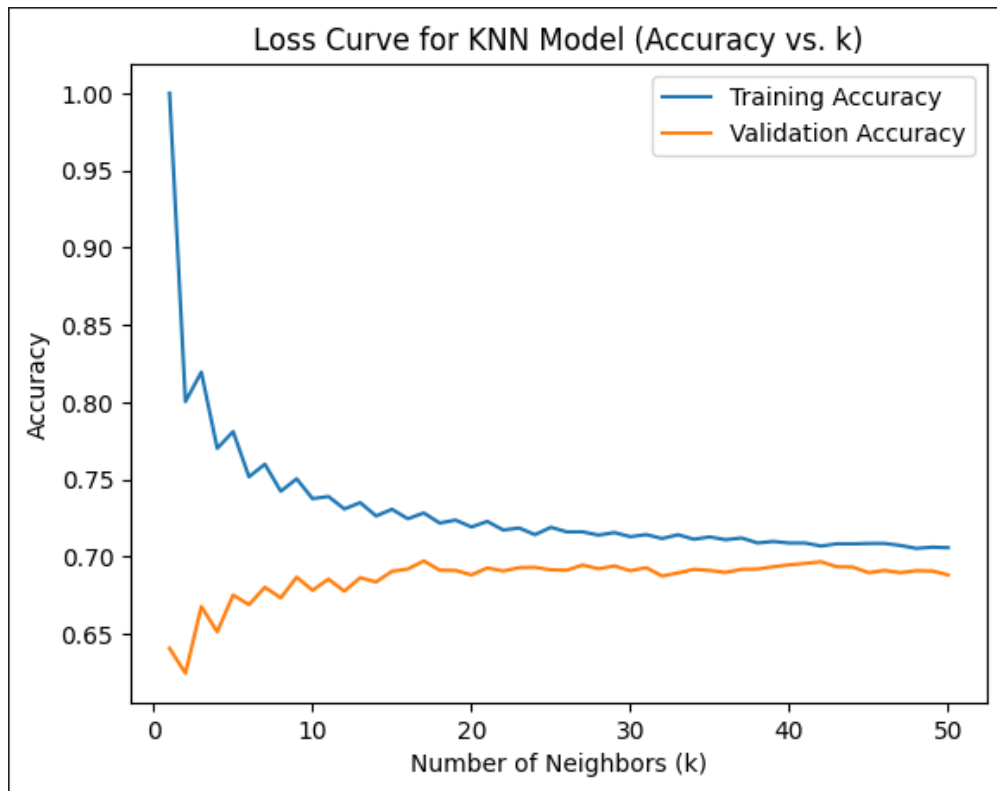
- **Dataset Name**: `Used Cars for Sale in Egypt`
- **Total Number of Samples**: 33559
- **Data Split**:
  - Training: 26848 samples
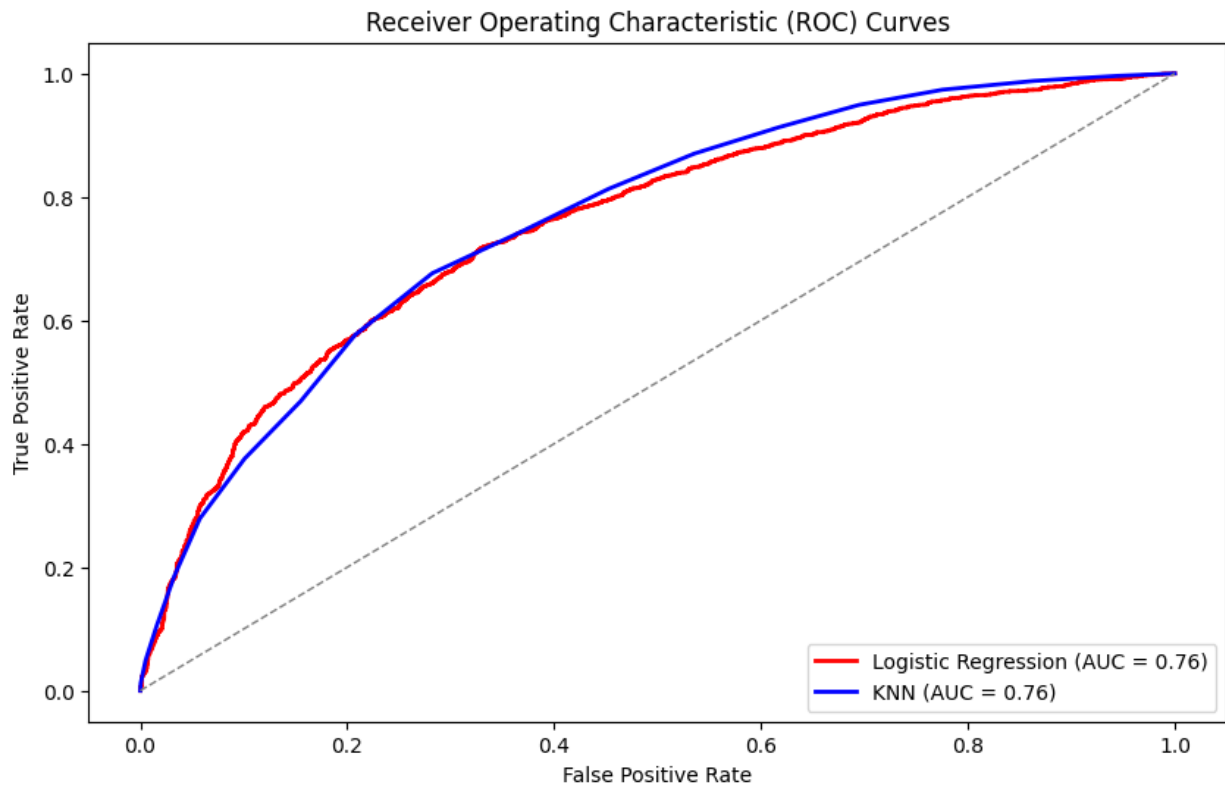 - Testing: 6711 samples

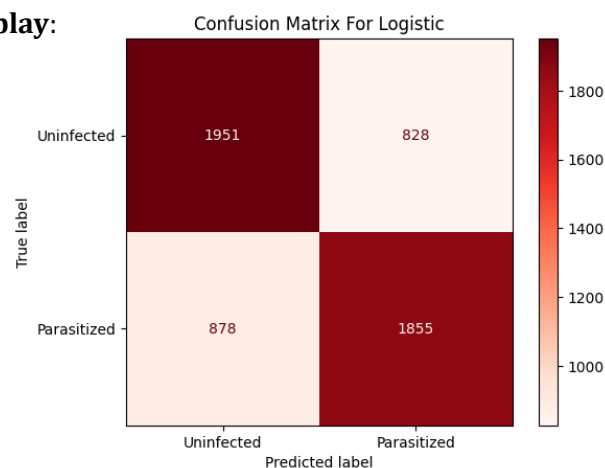# 3- Models

3.1- Image Dataset Models

- **Loss Curve:**

Loss Curve for KNN Model (Accuracy vs. k)

Loss Curves for KNN with Varying k

- **ROC Curve and AUC Plot:**

Receiver Operating Characteristic (ROC) Curves

3.1.1- Logistic Regression Model

- Uses logistic function for probabilities

- **Accuracy**: 0.6904934687953556

- **Precision**: 0.6913902348117779

- **Recall**: 0.6787413099158434

- **Confusion Matrix**: [[1951, 828], [878, 1855]]

- **Confusion Matrix Display**:

Confusion Matrix For Logistic
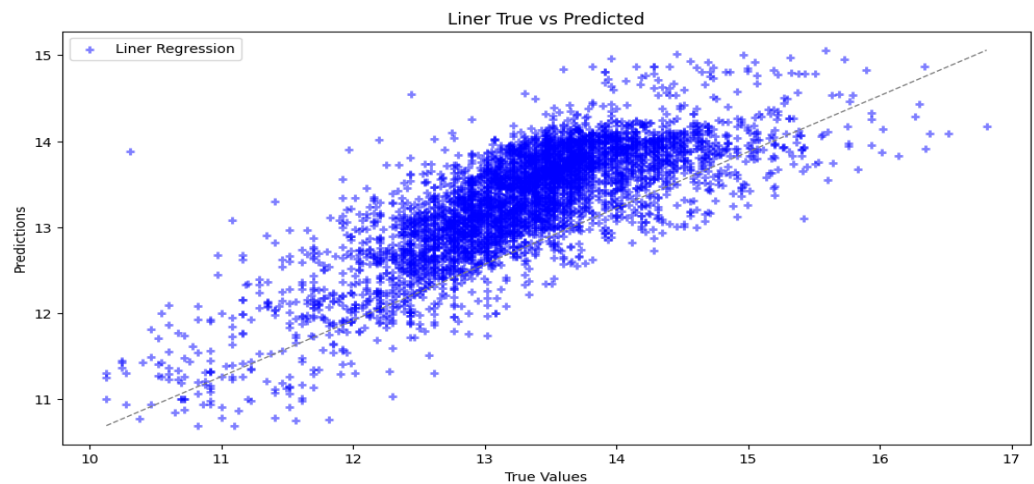
3.1.2- KNN Model

- Instance-based, distance-based, k neighbors

- **Accuracy**: 0.6972060957910015

- **Precision**: 0.7021276595744681

- **Recall**: 0.6761800219538968

- **Confusion Matrix**: [[1995 , 784], [885, 1848]]

- **Confusion Matrix Display**:
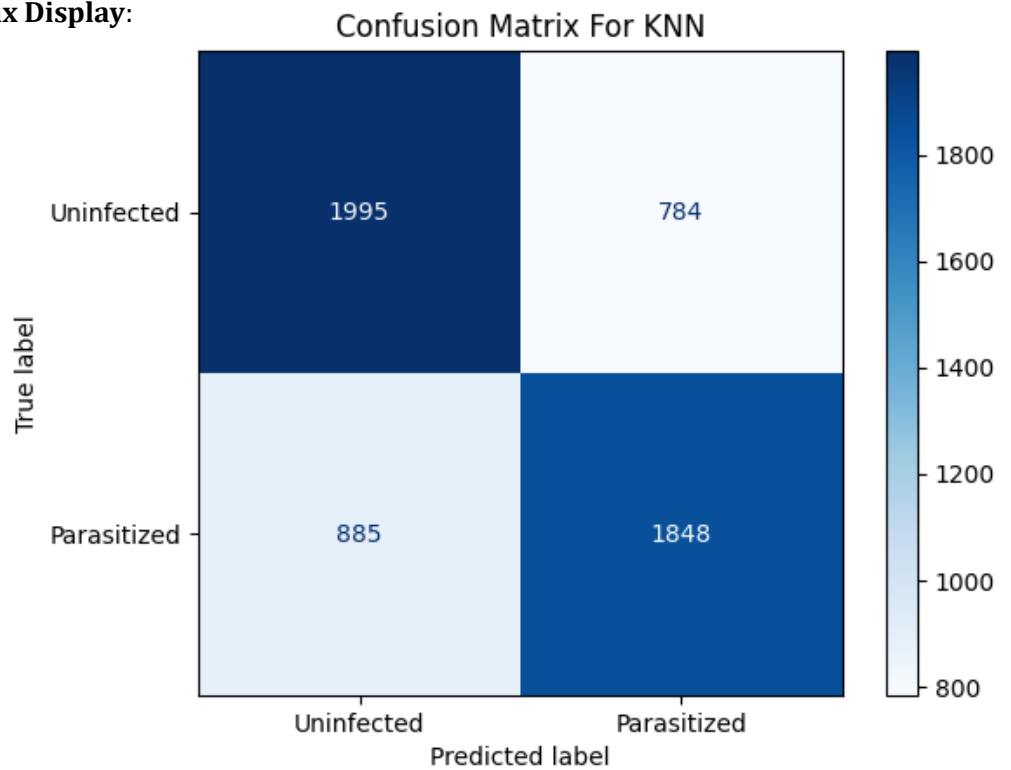
3.2- Numeric Dataset Models

3.2.1- Linear Regression Model

- Fits a linear equation to data

- **MAE**: 0.4043

- **MSE:** 0.2842

- **R2 Score**: 0.6118

- **Scatter Plot:**

3.2.2- KNN Model

- Instance-based, distance-based, k neighbors

- **MAE**: 0.1871

- **MSE:** 0.0877

- **R2 Score:** 0.8802

- **Scatter Plot:**