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`
- **URL**: https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria
- **Number of Classes :** `2 (Parasitized, Uninfected)`
- **Image Size**: varied
- **Total Number of Samples**: 27,558
- **Data Split**:
    - Training: 16535 samples
    - Testing: 11023 samples

2.2- Numerical Dataset

- **Dataset Name**: `Healthcare Insurance Expenses`
- **URL:** https://www.kaggle.com/datasets/arunjangir245/healthcare-insurance-expenses/data
- **Total Number of Samples**: 1338
- **Data Split**:
    - Training: 1071 samples
    - Testing: 267 samples

3- Models

Linear Regression Model:

- Dataset used: Healthcare Insurance Expenses

- Implementation details:

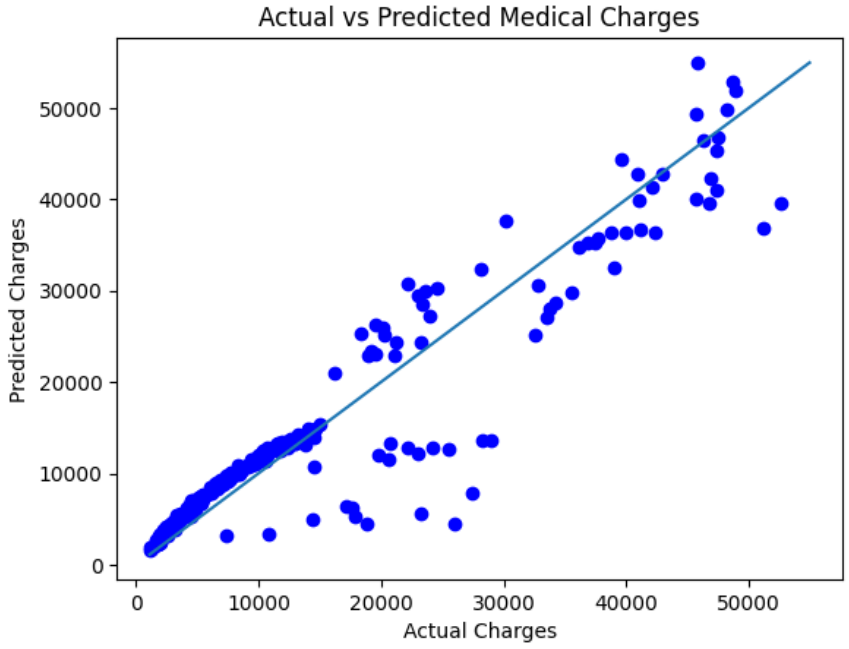
  Feature extraction: There were total of 7 features extracted, 4 numerical features and 3 non numerical. The numerical fields were: age, bmi, no. of children, charges. The non numerical fields were: sex, smoker, region. We also applied data augmentation to create a new feature: **smoker_bmi**. This is because charges more likely to be high for someone who has a high bmi and is a smoker.

  Dimensions of resulted features: We used label encoder for the sex, smoker fields and one hot encoder for the region field. Therefore the dimensions are: 6 numerical fields(including sex and smoker) + 4 categorical fields(for region; because it has 4 unique values) = 10 features.

  Cross Validation: Cross validation was not used in training the model, only a train test split.

  Hyperparameters: no hyperparameters were tuned.

- Results:
  Scatter Plot:



- Mean Squared Error: 0.13
- R^2 Score: 0.88

KNN Model:

- Dataset used: Healthcare Insurance Expenses

- Implementation details:

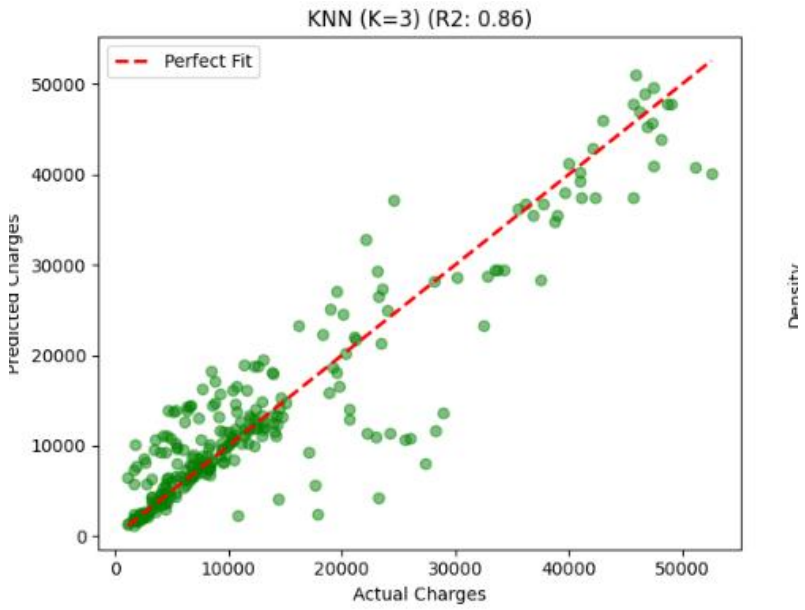
  Feature extraction: There were total of 7 features extracted, 4 numerical features and 3 non numerical. The numerical fields were: age, bmi, no. of children, charges. The non numerical fields were: sex, smoker, region. We also applied data augmentation to create a new feature: **smoker_bmi**. This is because charges more likely to be high for someone who has a high bmi and is a smoker.

  Dimensions of resulted features: We used label encoder for the sex, smoker fields and one hot encoder for the region field. Therefore the dimensions are: 6 numerical fields(including sex and smoker) + 4 categorical fields(for region; because it has 4 unique values) = 10 features.

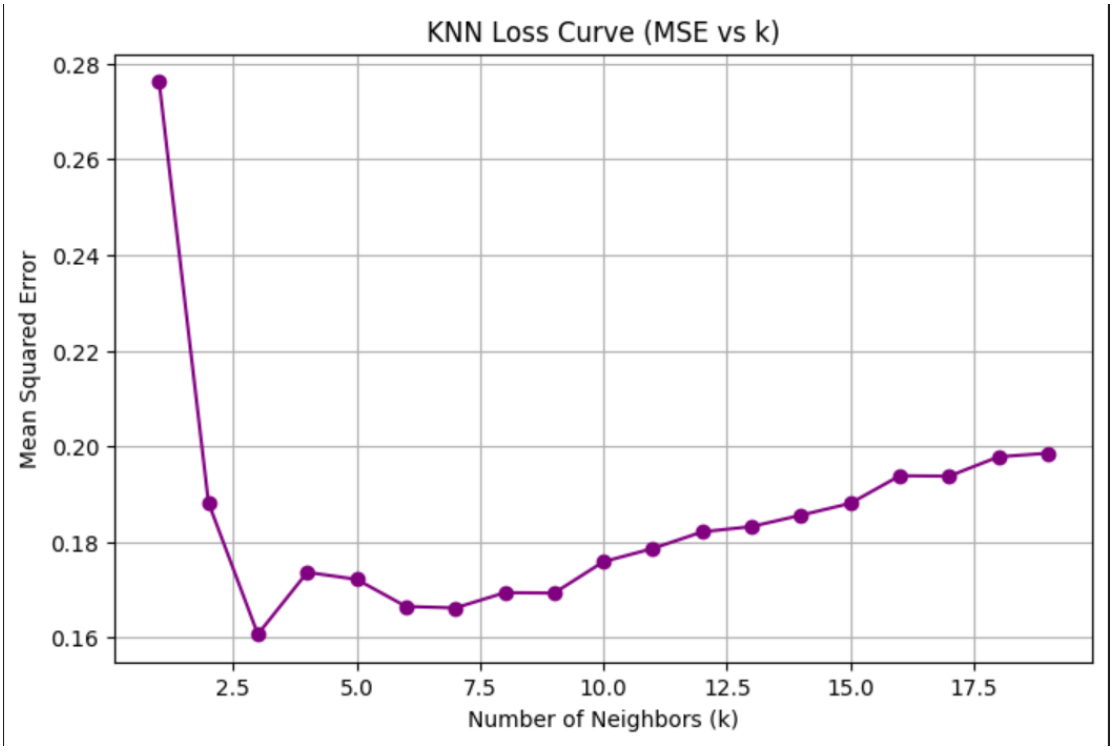
  Cross Validation: Cross validation was used in hyperparameter K.

  Hyperparameters: the only hyperparameter tuned was the value of **K** which is the number of neighbors used in the model. It was calculated dynamically by a for loop. For k=1 to k=20, we calculated the R^2 score for each k, then picked the k corresponding to the best score.

- Results:
  Scatter Plot:



Loss Curve: MSE vs K



- KNN Mean Squared Error: 0.16
- KNN R^2 Score: 0.86

Logistic Regression Model:

- Dataset used: Cell images for detecting malaria

- Implementation details:

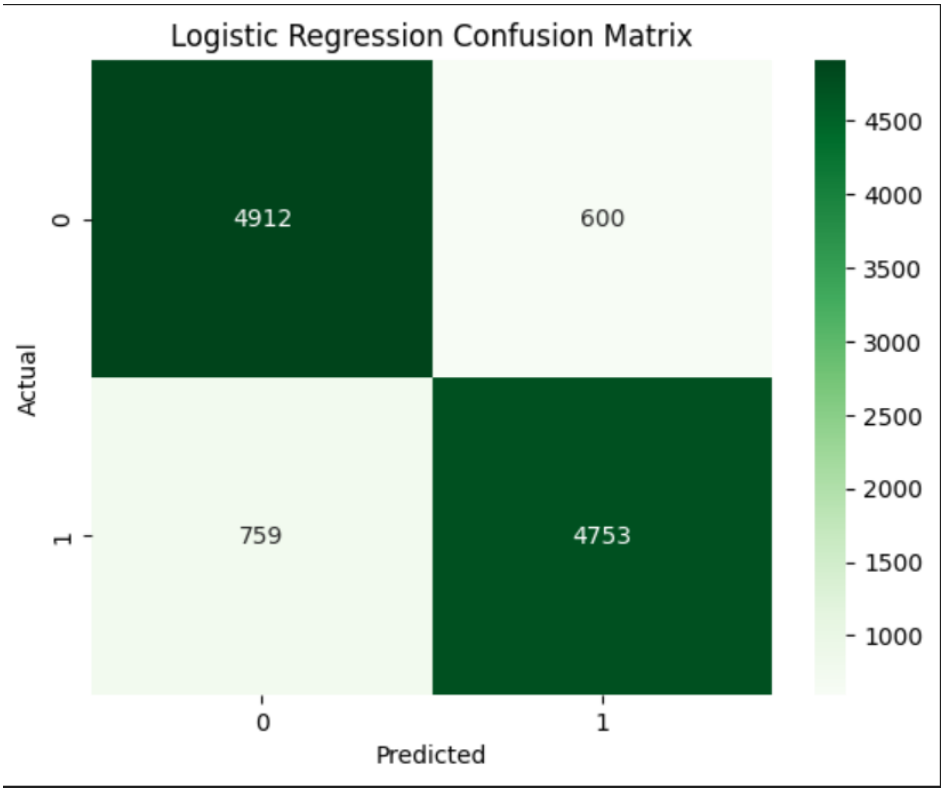
  Feature extraction:

- The mean and variation of Hue, Saturation, Value (HSV) color of each image
- Edge density of each image
- Blue, green, red (BGR) color ratio of each image
- Histogram representing intensity distribution across each image

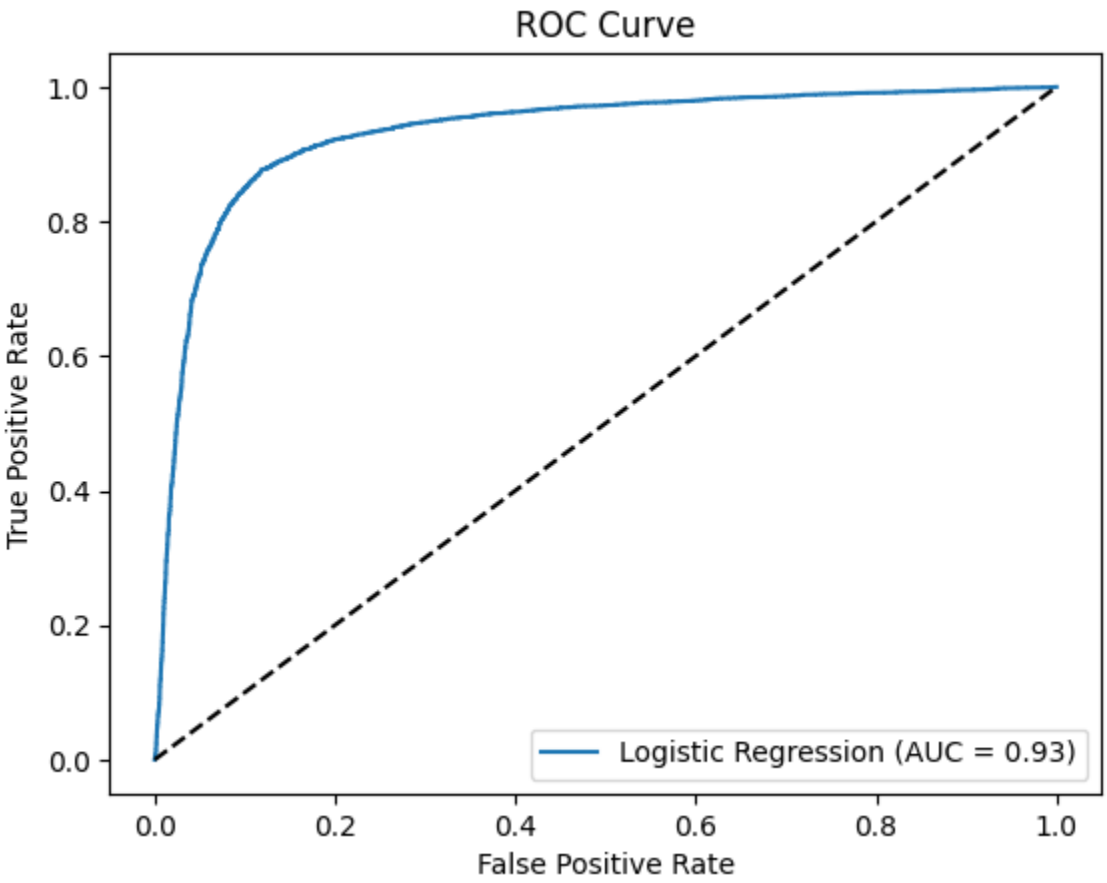
  Dimensions of resulted features: Each feature vector has 12 features total.

  Hyperparameters: maximum iterations=1000

- Results:
  Confusion Matrix:



- ROC Curve



Logistic Regression Accuracy: 0.88

K-Means Model:

- Dataset used: Cell images for detecting malaria

- Implementation details:

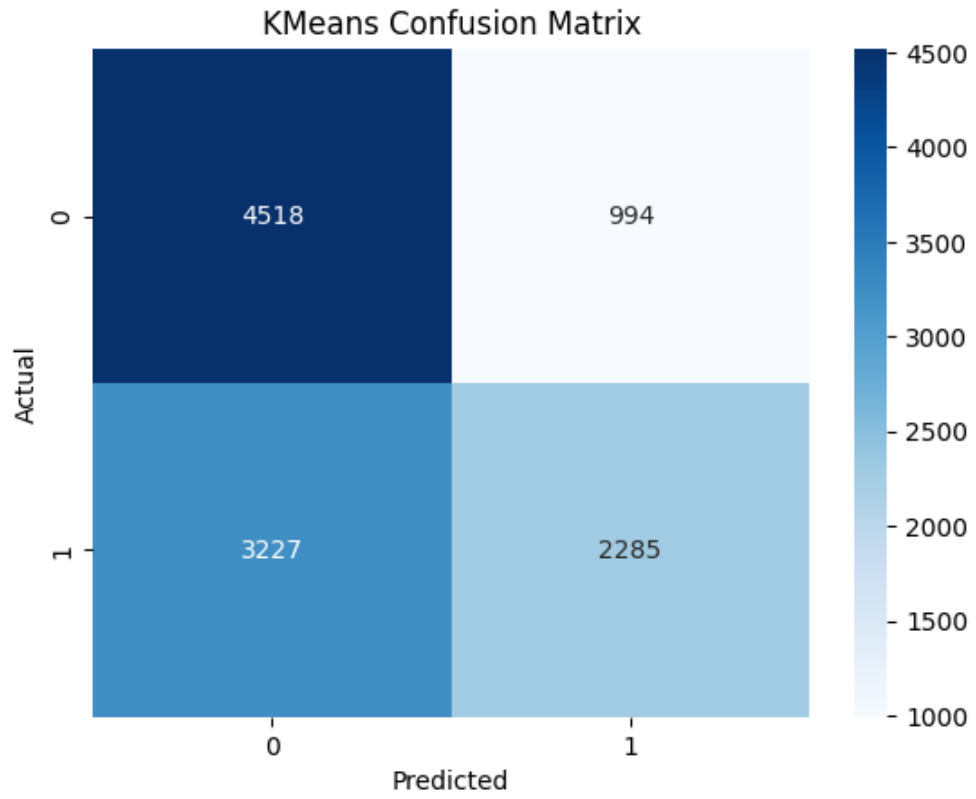
  Feature extraction:

- The mean and variation of Hue, Saturation, Value (HSV) color of each image
- Edge density of each image
- Blue, green, red (BGR) color ratio of each image
- Histogram representing intensity distribution across each image

  Dimensions of resulted features: Each feature vector has 12 features total.

  Hyperparameters: PCA n_components: set to 2.

- Results:
  Confusion Matrix:



- KMeans Accuracy: 0.62