12/1/2020

PROJECT-V2

**Sagunesh Grover 100800447**

Contents

[1. Problem Statement and Motivation 2](#_Toc54272707)

[2. The Data 2](#_Toc54272708)

[2.1 Data Acquisition 2](#_Toc54272709)

[2.2 Data Characterstics 3](#_Toc54272710)

[2.3 Data Assumption, Limitations and Constraints 3](#_Toc54272710)

[3. Performance Measure 3](#_Toc54272711)

[4. EDA(Exploratory Data Analysis) 4](#_Toc54272708)

[4.1 Data Sampling 5](#_Toc54272709)

[4.2 Pair Plots 6](#_Toc54272710)

[4.3 Correlation Heat Map 7](#_Toc54272710)

## Problem Statement and Motivation

According to World Health Organization (WHO), Malaria is a preventable and curable life-threatening disease caused by parasites that is transmitted to people through the bites of infected female Anopheles mosquitoes. In 2018, there were an estimated 228 million cases of malaria, aggregating to an estimated death count of 4,05,000 worldwide. Malaria control and elimination alone estimated to US$ 2.7 billion in funding, of which 30% (US$ 900 million) amounted to contributions from governments of endemic countries. As years pass by, these numbers continue to rise.

Moreover the process of blood smear workflow for Malaria detection involves intensive examination of the blood smear at a 100X magnification, where people manually count red blood cells that contain parasites out of 5000 cells which alone signifies the motivation for the development of a robust tool which eases and affectively automate this cumbersome process of blood smear workflow for Malaria detection enabling researchers and on-site infectious disease physicians to arrive at a quick and accurate diagnosis.

## The Data

## 2.1 Data Acquisition

## The following dataset ([Malaria Datasets](ftp://lhcftp.nlm.nih.gov/Open-Access-Datasets/Malaria/cell_images.zip)) was acquired from [National Library of Medicine (NLM)](https://lhncbc.nlm.nih.gov/publication/pub9932).

The dataset contains 2 folders

* Infected / Parasitized
* Uninfected
* And a total of 27,558 images.

## 2.2 Data Characteristics

Giemsa-stained thin blood smear slides from 150 Parasitized, falciparum-infected and 50 healthy patients were collected and photographed at Chittagong Medical College Hospital, Bangladesh. The images were manually annotated by an expert slide reader at the Mahidol-Oxford Tropical Medicine Research Unit in Bangkok, Thailand. The dataset contains a total of 27,558 cell images with equal instances of parasitized and uninfected cells. An instance of how the patient-ID is encoded into the cell name is shown herewith: “P1” denotes the patient-ID for the cell labeled “C33P1thinF\_IMG\_20150619\_114756a\_cell\_179.png”.

## 2.2 Data Assumption, Limitations and Constraints

Since the dataset I’m working on consists of 27,558 images of blood smear slides that were collected and photographed at Chittagong Medical College Hospital in Bangladesh, I have assumed the working assumption that this dataset would be sufficient for training the deployable model to effectively detect malarial instances (at least in Bangladesh). I have also assumed that the expert slide readers at the Mahidol-Oxford Tropical Medicine Research Unit in Bangkok, Thailand correctly annotated these samples.

## 3 Performance Measure

In order to guarantee the quality of our work, I will be using metrics like accuracy, ROC-AUC Score and F1 score metrics to compute the performance of the model. Our main goal is to reduce False Positive rate in-order to achieve state of the art performance.

## 4 EDA (Exploratory Data Analysis)

We will preprocess image in-order to generate data first so there are various steps for the same.



Step1: - On the Image Gaussian Blur is Applied



Step2: - After Gaussian Blur image is converted to gray scale as it will easy for the algorithm to process the same as compared to RGB color.



Step3: - Threshold of image is changed in-order to identify contours



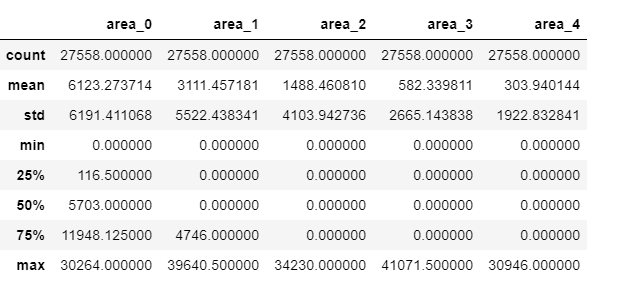
Step4: - After Identifying, contour area is highlighted

Area is calculated: - **['221.0', '9086.5', '0', '0', '0']**

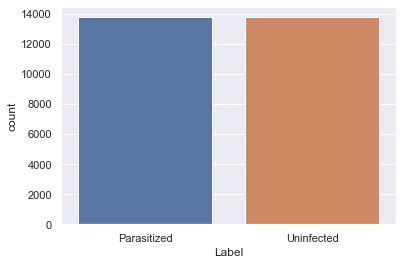
* We will apply the above steps to each image and Area for each Image will be calculated as shown in the below image for both the classes.

## 

## 4.1 Data Sampling

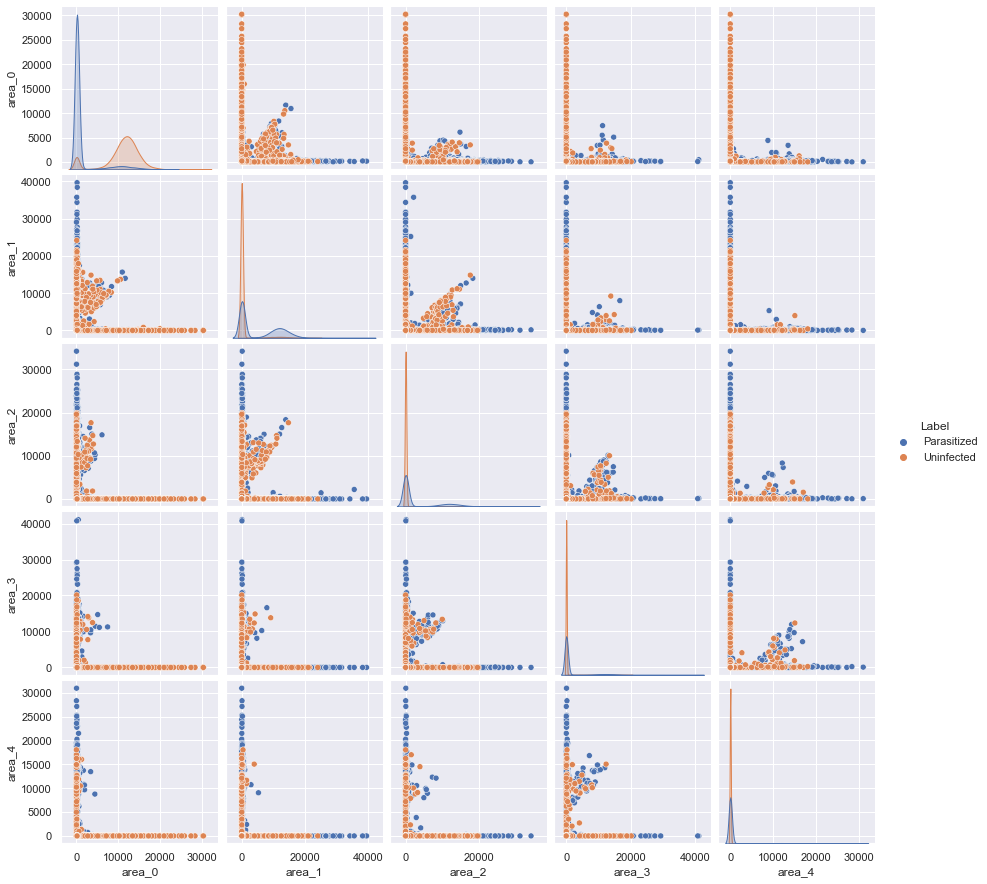


We can see in the below plot both the classes are balanced, so we have a balanced dataset



4.2 Pair Plots

Below pairplot shows relationship of different contour areas with each other and also we can see data distribution through histograms that it is not normally distributed



4.3 Correlation Heatmap

Below correlation heatmap shows that variables are not correlated with each other.

