Malaria Prediction Using SVM

Gaurav Khatri (A25318673)
The University of Alabama in Huntsville
CS617 : Machine Learning

Final Report

Date *Nov 28, 2022*

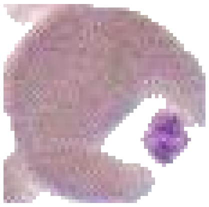
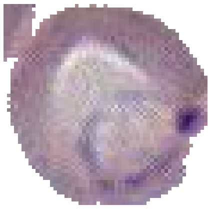
1. Introduction

This report is prepared as a part of the final submission requirement for CS617, Homework 4. In this project, we focus on a prediction of Malaria using RBC images for over 2500 different samples. We focus on classifying the data of each sample across two categories i.e. Class 0 (Negative Cases) and Class 1 (Positive Cases).

2. Methodologies

2.1 Dataset

The UAH Curated Dataset has been used for this project, which consists of 2565 samples.The samples belong to 2 different classes i.e. 0 and 1. Each data sample has RGB components, with varying resolution. Hence it is reduced to a uniform size in later phases.

 Fig 1 : Class 0	 Fig 2 : Class 1

2.2 Data Processing

In this step we combine the Data Extraction, Cleaning ,Aggregation and Storage steps of the Big Data Analytics Lifecycle. Initially, the data is compiled from the source into a single repository. Then each image is resampled to (60*60*3) image size, in order to maintain uniformity and reduce the data complexity. Then each image is converted to grayscale and flattened to give us a singular representation of 10,800 features per image. Finally all the images are stored in a single Pandas DataFrame in order to make the analysis easier.

2.3 Exploratory Data Analysis

In this step, we aim to understand the data a bit more before moving on towards modeling. Initially we looked at the overall class distribution and found that around 60% of the samples belonged to Class 0. The distribution seems balanced, hence we can proceed further with balanced assumptions.

With over 10000 feature values, it is difficult to understand the overall features for this dataset. Hence we now move towards dimensionality reduction.

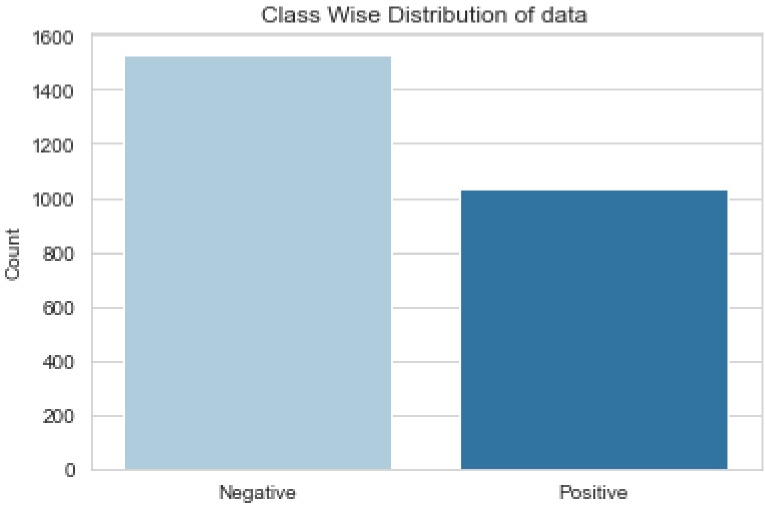


Fig 2.3.1 : Class Wise Distribution of Data Samples

2.4 Dimensionality Reduction

In this step, we construct a sklearn data pipeline in order to Scale the data and then apply dimensionality reduction using PCA. We design the pipeline in such a way that 20% of data samples will be reserved for actual testing at the end and 80% of the data is then used for modeling and cross validation (10 fold cross validations). We tuned our PCA model to attain only 90% of the variance, which resulted in reduced feature size to : 135 Principal Components, which is a reduction in dimensions by 80 times.

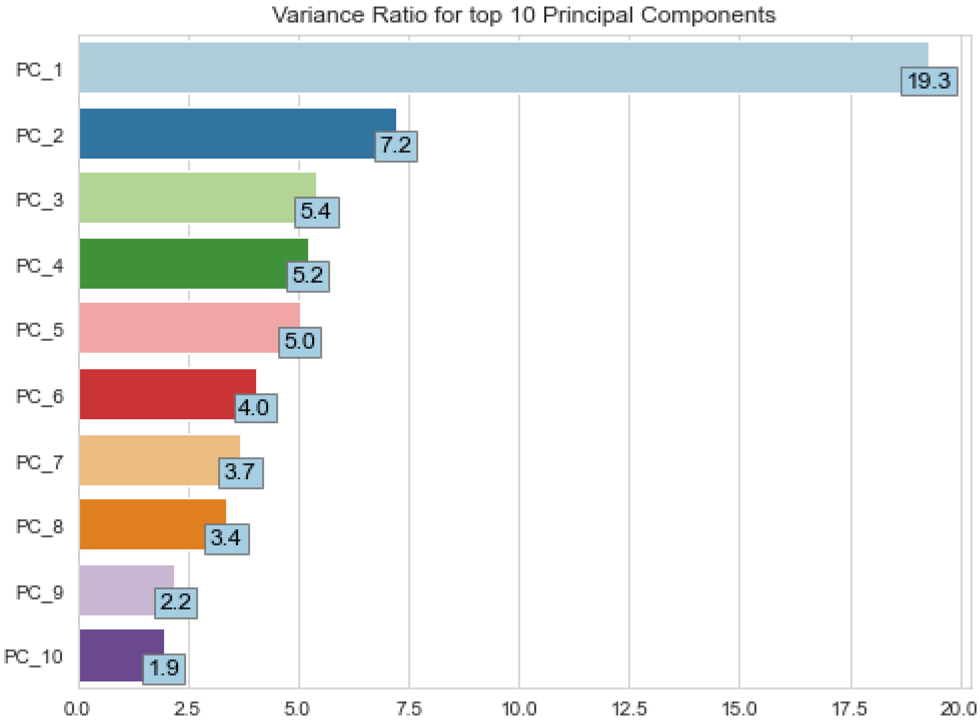


Fig 2.4.1 : Explained Variance Ratio for PCA

2.5 Exploratory Data Analysis after PCA

Since our data was reduced to a much smaller number of dimensions with a fairly better understanding of relative feature importances, different analyses were performed to see the overall data distribution via pairplots.

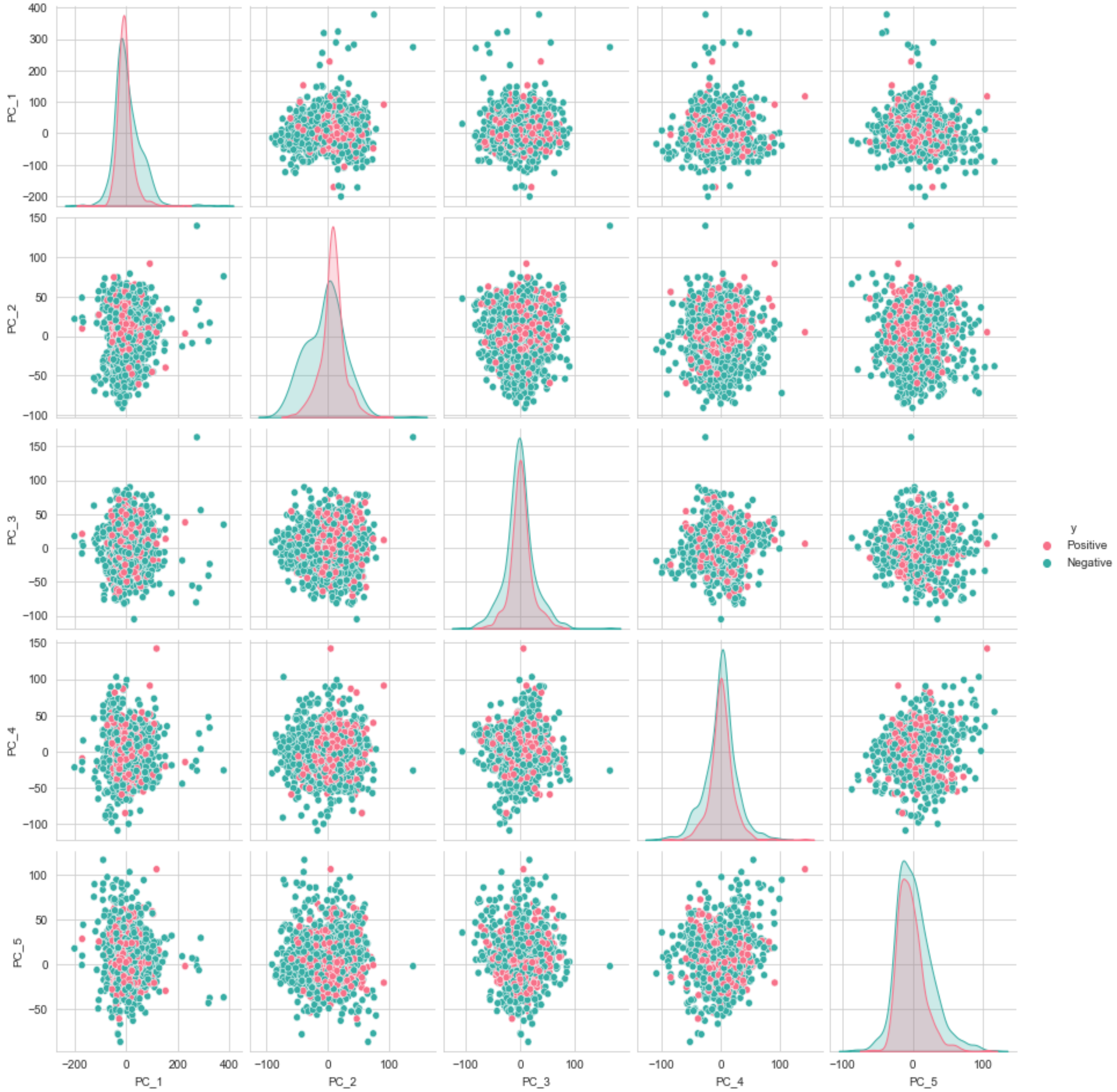


Fig 2.5.1 : Pairplot Distribution of Top 5 PCA components

Upon running the pairplot of top 5 PCA components, we could see that each of the 2d scatter plots were much convoluted. Hence this led to the assumption that the data might not be linearly separable, thus leading to the belief that non-linear classification methods such as SVM (Support Vector Machines) could lead to higher accuracy.

2.6 Data Modelling

Based on the findings of earlier steps, SVM was chosen to be a good algorithm for this classification task. A training pipeline to was designed to develop a robust machine learning model with following steps:

1. Parameter space definition for varying regularization parameter C and kernel functions
2. Grid Search Space to search for best solution in the parameter space
3. K fold Cross Validation (steps = 10) to find the robust model within the parameter space

The overall search space specification can be tabled as follows:

	Parameters

SVM	c_ = [0.01, 0.1, 0.5,1,2,5,10]
	kernel = ['rbf','poly']

3. Results and Analysis

This method resulted in highest test accuracy of 96.68%, subjected to C = 5 and kernel = rbf.

Classification Report for Test Data:

	Precision	Recall	F1 Score	Accuracy
Average	0.97	0.96	0.97	0.968

As expected, the overall accuracy for SVM was really high, given the nature of data.

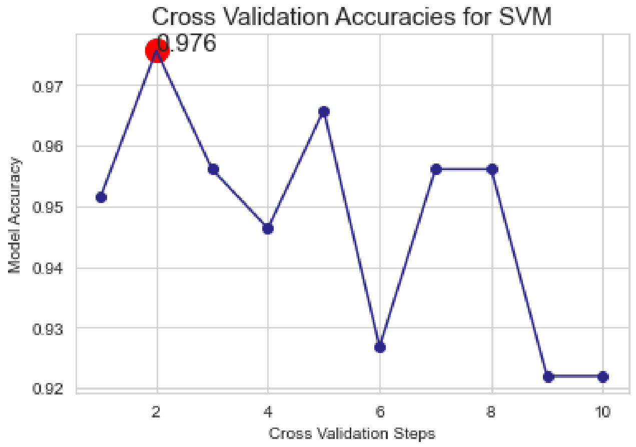


Fig 3.1.1 : Cross validation accuracies for SVM

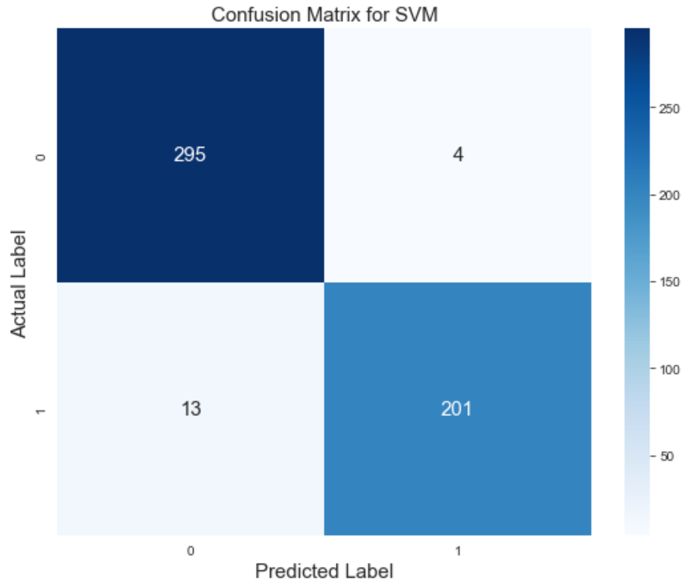


Fig 3.1.2 : Confusion Matrix for SVM

4. Conclusion

In this way we concluded a very successful data science project on predicting the prevalence of Malaria on the test dataset using cell images as input. We finalized on the SVM model, for which hyperparameter tuning was done with corresponding KFold Cross Validation , resulting in a model with 96.7% accuracy in the test dataset.

5. References

1. Khattak, Wajid, et al. *Big Data Fundamentals: Concepts, Drivers & Techniques*. Edited by Thomas Erl, Prentice Hall, 2016.
2. “Support vector machine.” *Wikipedia*, Wikipedia, 2022, https://en.wikipedia.org/wiki/Support_vector_machine. Accessed 26 November 2022.