Assignment 2 ML
Aarzoo (2022008)

Section A

Q1

(a) $\sigma^2 = 36$

$\mu = 10$

~~P(Profit increase ≥ 4)~~

$P(\text{Dividend}) = 0.8$

$P(\text{Not dividend}) = 0.2$

$P(\text{Dividend} \mid \text{Profit} = 4\%) \propto P(\text{Profit} \mid \text{Dividend}) \times P(\text{Dividend})$

$$= \frac{P(\text{Profit} \mid \text{Dividend}) \times P(\text{Dividend})}{P(\text{Profit} = 4\%)}$$

$P(\text{Profit} \mid \text{Dividend})$

$= P(\text{Profit} \mid \mu, \sigma) = ?$

since it follows Normal/Gaussian Distribution

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{(X-\mu)^2}{2\sigma^2}\right)}$$

$$= \frac{1}{\sqrt{2\pi \times 36}} e^{-\left(\frac{(4-10)^2}{2 \times 36}\right)}$$

$$= \frac{1}{6\sqrt{2\pi}} e^{-\left(\frac{1}{2}\right)} = 0.0403$$

$(\mu = 0)$ $P(\text{Profit} \mid \sim\text{Dividend}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{(4-0)^2}{2 \times 36}\right)}$

$= 0.0532$

$\therefore$ $P(\text{Profit} = 4\%) = P(\text{Dividend}) \times P(\text{Profit} = 4\% \mid \text{Dividend}) + P(\text{Not Dividend}) \times P(\text{Profit} = 4\% \mid \text{Not Dividend})$

$$= 0.8 \times 0.0403 + 0.2 \times 0.0532$$
$$= 0.0429$$

$\therefore$ P (Dividend | Profit = 4%)
$$\propto (0.0403 \times 0.8) \propto 0.032$$

(which can also be stated as)

$$= \frac{0.0403 \times 0.8}{0.0429} = 0.7518$$

$\therefore$ 75.18% chance is there that the company issues the dividend given the profit 4%.

Q2

(b) $IG = H(Y) - H(Y|X)$

(i) $H(Y) = -\frac{7}{12}\log(7/12) - 5/12\log(5/12)$

$P(\cancel{Yes}) = 7/12, \quad P(No) = 5/12$

Part I $\boxed{H(Y) = 0.9798}$

(ii) $\cancel{IG}$ $H(Y|\text{Class Time}) = ?$

Morning $\Rightarrow -4/12\left(3/4\log 3/4 + \frac{1}{4}\log 1/4\right)$
$+$
Noon $\Rightarrow -4/12\left(1/2\log(1/2) + 1/2\log 1/2\right)$
$+$
Afternoon $-4/12\left(1/2\log(1/2) + 1/2\log(1/2)\right)$

$\Rightarrow \boxed{0.937}$

$\therefore \boxed{IG(\text{Class Time}) = 0.0428}$

(iii) $H(Y|\text{sleep}) = ?$

$= -\frac{1}{2}(1\log 1 + 0\log 0) +$

$= -\frac{1}{2}\left(1/6\log(1/6) + 5/6\log(5/6)\right)$

$= \boxed{0.3249}$

$\therefore \boxed{IG(\text{sleep}) = 0.6549}$

(iii) $H(Y \mid \text{Weather}) = ?$

Cool $= -\frac{5}{12}\left(\frac{4}{5}\log\frac{4}{5} + \frac{1}{5}\log\left(\frac{1}{5}\right)\right) +$

Rainy $-\frac{2}{12}\left(1\log 1 + 0\log 0\right) +$

Hot $-\frac{5}{12}\left(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right)$

$= \boxed{0.704}$

$\therefore \boxed{IG(\text{Weather}) = 0.275}$

$\therefore$ Sleep has highest information thus would be placed first. (Root node)

Part II

$H(\text{Sleep} = \text{Yes}) = 0$

$H(\text{Sleep} = \text{No}) = 0.6549$

* Only considering sleep = No case now:

(i) ~~H(~~ $IG(\text{Time}) = ?$

$H(\text{Morning}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$
$= 1$

$H(\text{Noon}) = -0\log 0 - 1\log(1) = 0$

$H(\text{Afternoon}) = -0\log 0 - 1\log(1) = 0$

$\therefore IG = 0.654 - \frac{1}{3}(1) = \boxed{0.321}$

(ii) $IG(Weather) = ?$

$H(Cool) = -\frac{1}{2}\log(\frac{1}{2}) - \frac{1}{2}\log(\frac{1}{2}) = 1$

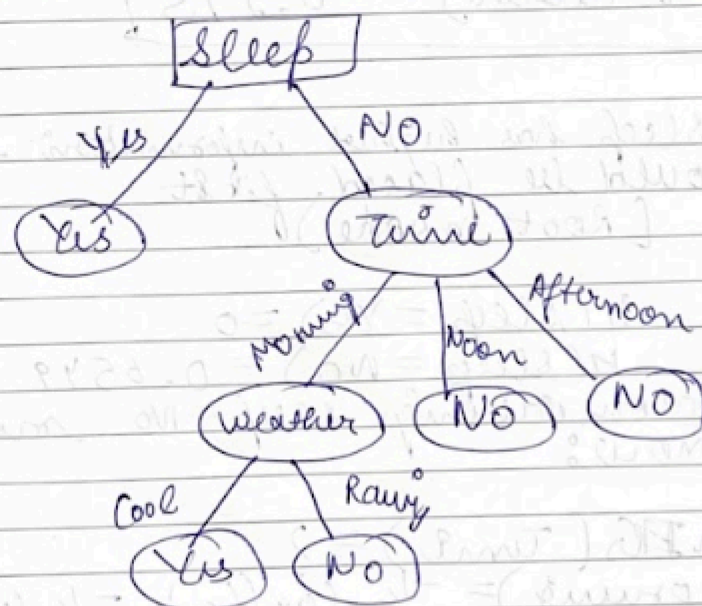
$H(Rainy) = -0\log(0) - 1\log(1) = 0$

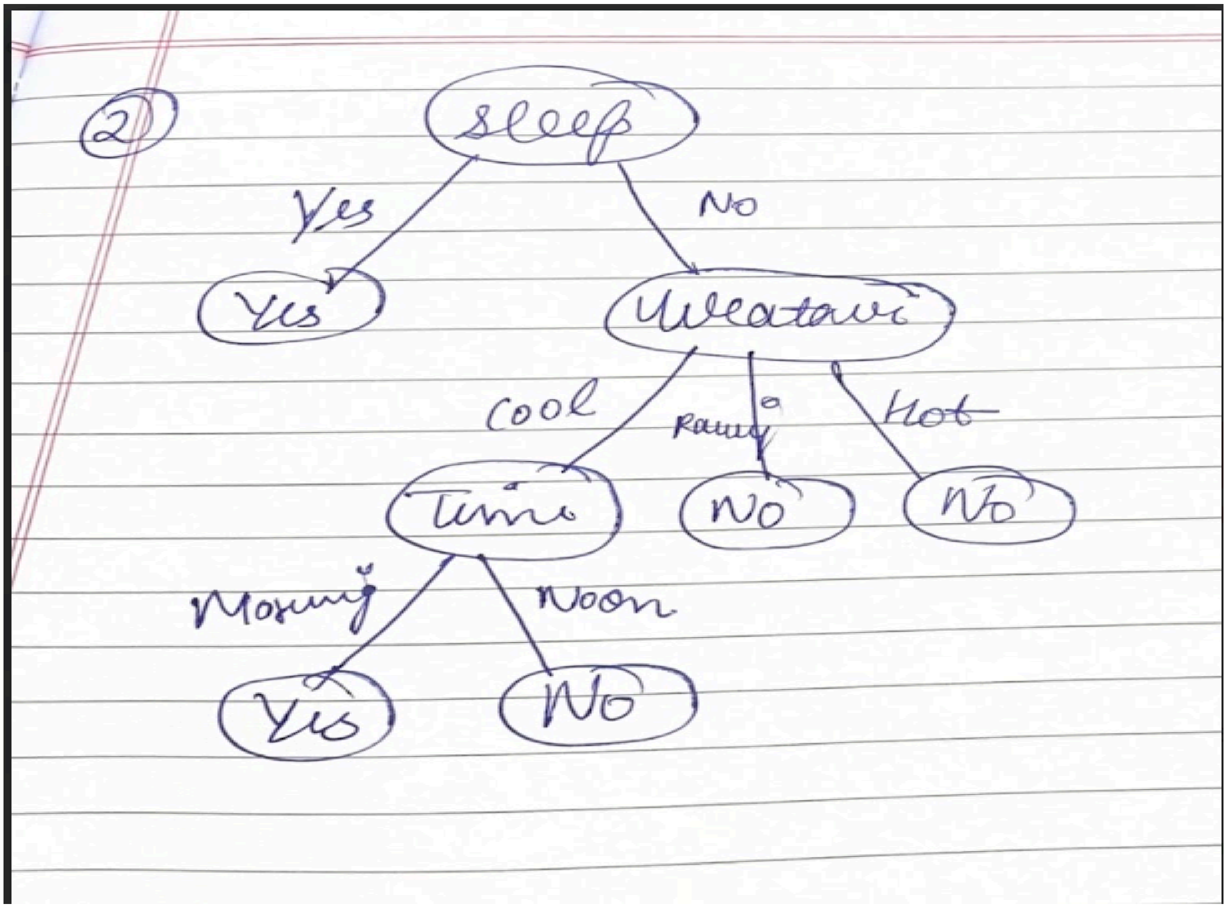
$H(Hot) = -0\log 0 - 1\log 1 = 0$

$\therefore \boxed{IG = 0.321}$

$\therefore$ can choose any $\therefore$ 2 Decision trees.

(1)





Q3 Reference: https://www.cs.cmu.edu/~avrim/ML10/lect0125.pdf

c) Proving: The number of mistakes M on S (sequence of labelled examples)

Part I: by the mentioned Perceptron Algorithm is at most $(1/\gamma)^2$, where

$$\gamma = \min_{x \in S} \frac{|w^* \cdot x|}{||x||}$$

where w^* is a unit vector and also given that $w^* \cdot x > 0$.

The parameter "γ" stated is called "margin" of w^*.

- Claim 1: $w_{t+1} \cdot w^* \geq w_t \cdot w^* + \gamma$, that is every time we made a mistake, the dot product of our weight vector with the target increases by at least γ.

$\Rightarrow$ If $x > 0$, $w_{t+1} \cdot w^* = (w_t + x) \cdot w^* = w_t \cdot w^* + x \cdot w^* \geq w_t \cdot w^* + \gamma$

$\Rightarrow$ If $x < 0$, $(w_t - x) \cdot w^* = w_t \cdot w^* - x \cdot w^* \geq w_t \cdot w^* + \gamma$.

- Claim 2: $||w_{t+1}||^2 \leq ||w_t||^2 + 1$. I.e. every time we make mistake, the length squared of our weight vector increases by atmost 1.

$\Rightarrow$ If $x > 0$, $||w_t + x||^2 = ||w_t||^2 + 2 w_t \cdot x + ||x||^2$. This is less than $||w_t||^2 + 1$ since $w_t \cdot x < 0$ and vice versa.

this implies, after M mistakes, $w_{M+1} \cdot w^* \geq \gamma M$
Claim 2 implies that after M mistakes
$\|w_{M+1}\| \leq \sqrt{M} \Rightarrow w_t \cdot w^* \leq \|w_t\|$,
since w^* is a unit vector.

$$\Rightarrow \quad \gamma M \leq \sqrt{M} \Rightarrow \boxed{M \leq \frac{1}{\gamma^2}}$$

which is a polynomial of $\frac{1}{\gamma}$.

Part II:

If we can halt ~~the~~ in the number of rounds that is polynomial in $\frac{1}{\gamma}$ if we replace $\frac{\gamma}{2}$ with $(\cdot\gamma \ (1-\epsilon)\gamma$ then

the polynomial ~~m~~ is $1/(\epsilon\gamma)$.

Part III:

The number of mistakes (including margin mistakes) made by margin perceptron S is at most $8/\gamma^2$.

Proof: each update increases $w_t \cdot w^*$ by at least γ. since here we are taking the angle is more than 90°. $\Rightarrow \|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$,

$$\|w_{t+1}\| \leq \|w_t\| + \frac{1}{2\|w_t\|}$$

$$\|w_{t+1}\| \leq \|w_t\| + \frac{1}{2\|w_t\|} + \frac{\gamma}{2}$$

$\Rightarrow$ If $\|w_t\| \geq \frac{2}{\gamma}$ then $\|w_{t+1}\| \leq \|w_t\| + \frac{3\gamma}{4}$

So, after M updates we have $\|w_{M+1}\| \leq \frac{2}{\gamma} + 3M\gamma/4 \Rightarrow \boxed{M \leq 8/\gamma^2}$ as desired.

Q4

(a)

Email	Word "buy"	Word "cheap"	Spam (Label)
1	1	0	1
2	1	1	1
3	0	1	0
4	1	0	0

(a) $P(\text{spam}) = \frac{1}{2}$, $P(\text{Not spam}) = \frac{1}{2}$

① $P(\text{cheap} \mid \text{Spam}) = 1/2$
② $P(\text{buy} \mid \text{spam}) = 1$
③ $P(\text{cheap} \mid \text{Not spam}) = 1/2$
④ $P(\text{buy} \mid \text{not spam}) = 1/2$

(b) ① $P(\text{Spam} \mid \text{Cheap, Not buy}) \propto P(\text{Cheap} \mid \text{Spam}) \times P(\text{Not buy} \mid \text{spam}) \times P(\text{Spam})$

$\propto \frac{1}{2} \times 0 \times 1/2$

$P(\text{Not buy} \mid \text{spam}) = 0$

$= \boxed{0}$

② $P(\text{Not spam} \mid \text{Cheap, Not buy}) \propto P(\text{cheap} \mid \text{Not Spam}) \times P(\text{Not buy} \mid \text{Not Spam}) \times P(\text{Not Spam})$

$\propto \frac{1}{2} \times 1/2 \times 1/2 = \frac{1}{8} = \boxed{0.125}$

$P(\text{Not buy} \mid \text{Not Spam}) = 1/2$

3) If one of the conditional probability is zero, then the entire probability is zero,
so to address it Laplace and m-estimate methods are used.

Laplace: $P(A_i | C) = \frac{N_{ic}+1}{N_c+c}$

c: no. of classes
p: prior probability
m: parameter

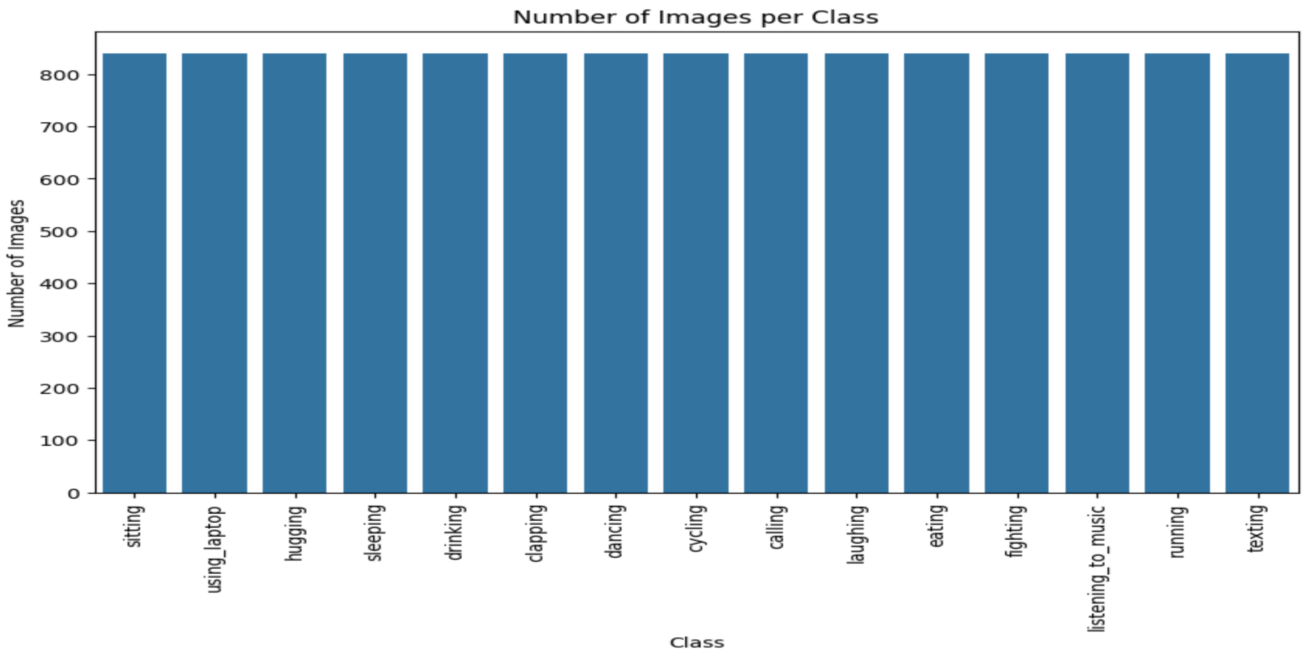
m-estimate: $P(A_i | C) = \frac{N_{ic}+mp}{N_c+m}$

Section C

Part A)

a) Observations:

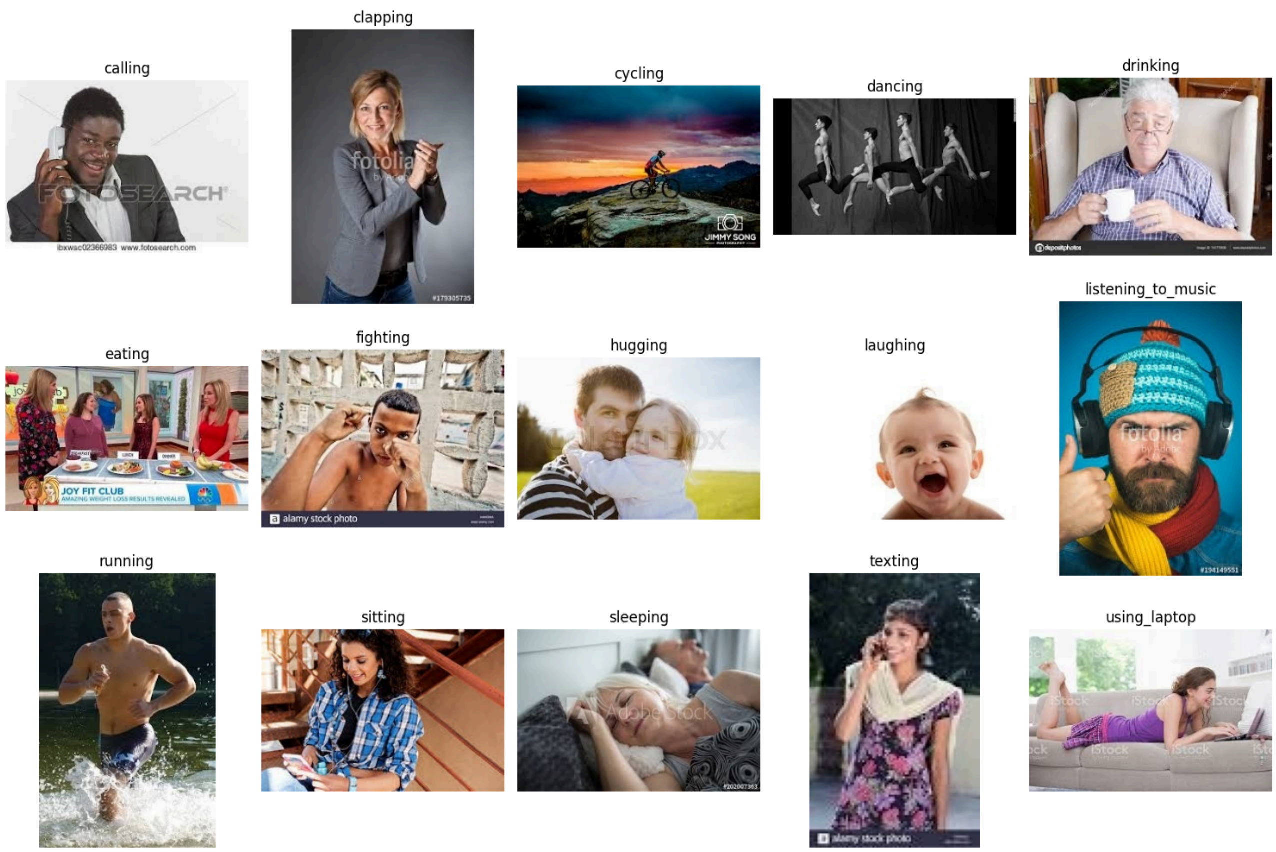
- The dataset is highly evenly distributed and contains 15 classes each having 12600 images.
- The description of how the image sizes are distributed is illustrated below that: the mean height and width are 195.57*260.38
- The image sizes are mostly deviated around (39.9, 35.28)
- The max and min sizing and some other features have also been described below.
- The pie chart, bar graphs etc, are illustrated below.



Images Sizes Description :

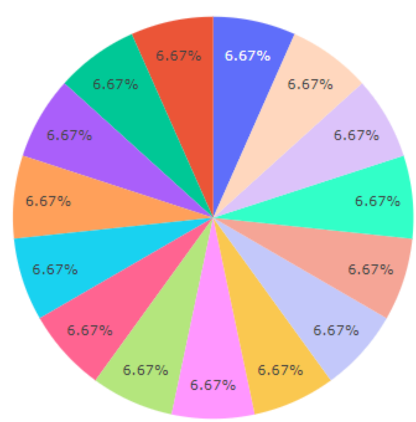
	width	height
count	12600.000000	12600.000000
mean	260.381032	196.573571
std	39.919281	35.281402
min	84.000000	84.000000
25%	254.000000	181.000000
50%	275.000000	183.000000
75%	276.000000	194.000000
max	478.000000	318.000000

b)



Random image generated from each label

Distribution of Human Activity Classes



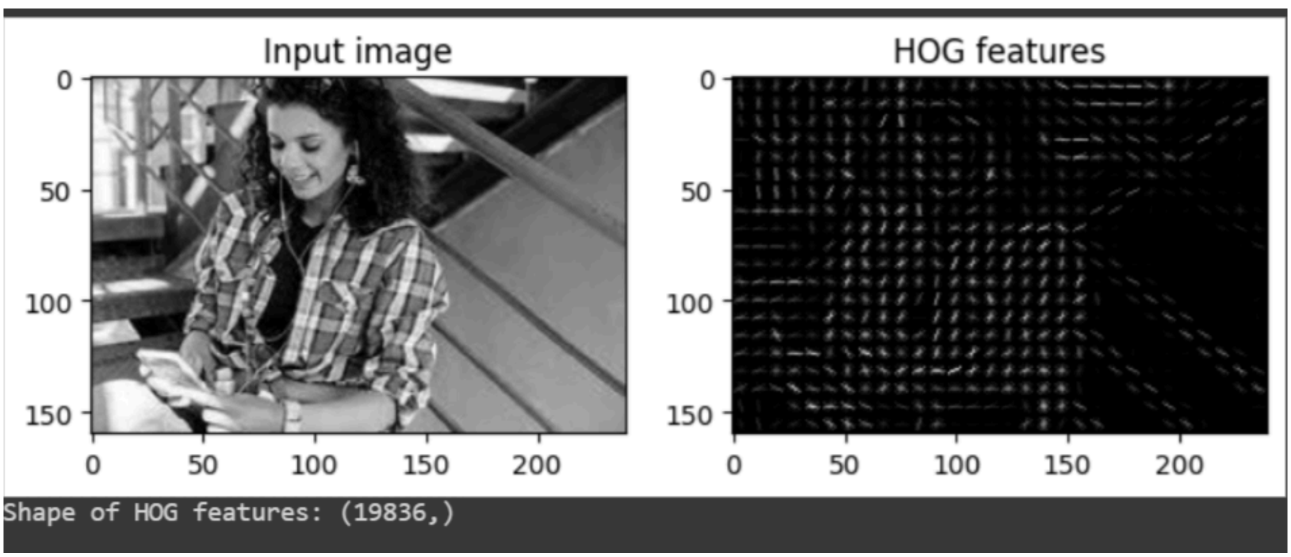
Generating 1 random image from each class and the distribution in class is represented above by the pie chart

c) There are no such class distributions except of the fact there is 1 image in the dataset which doesn't has a corresponding label. Thus, there is no need for data augmentation or resampling.

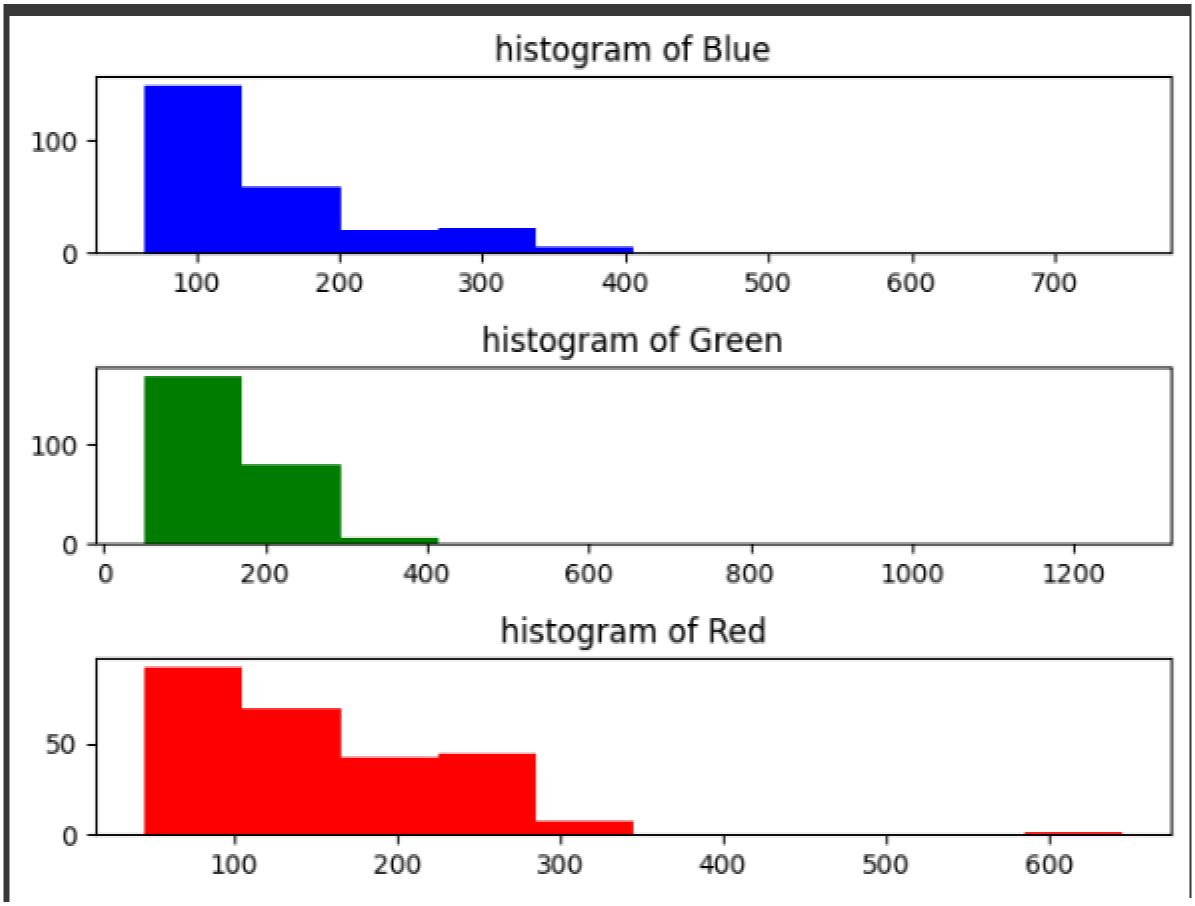
Part B) Features extracting methods tried: HOGS, LBP , Color Histogram

Observations and Evaluations:

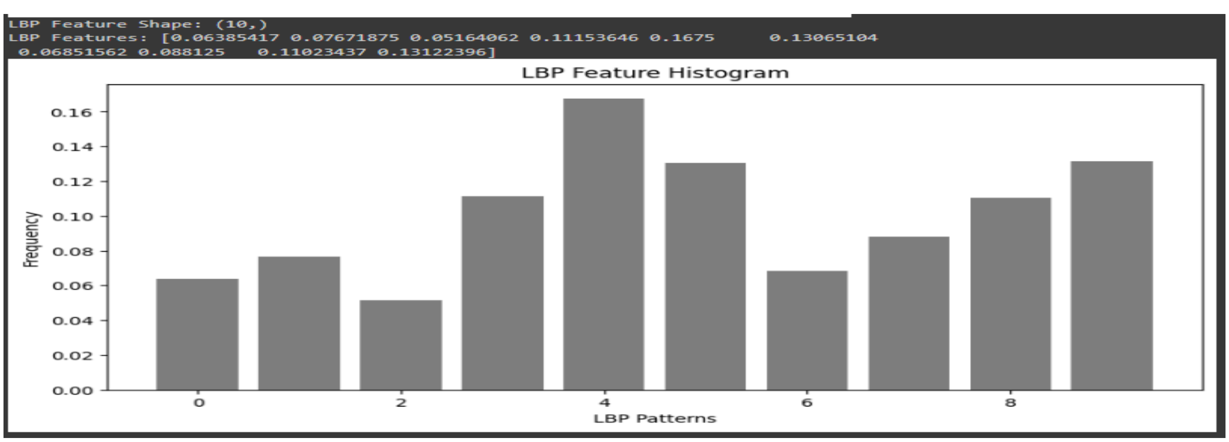
HOGs: The HOG features extraction method computes histograms of oriented gradients in localized regions of the image. The code implemented takes pixels per cell 8*8 and cells per block 2*2 for it. The method extracted 19836 features from an image. It does not proved to be a fit for my model and gave the lowest accuracy.



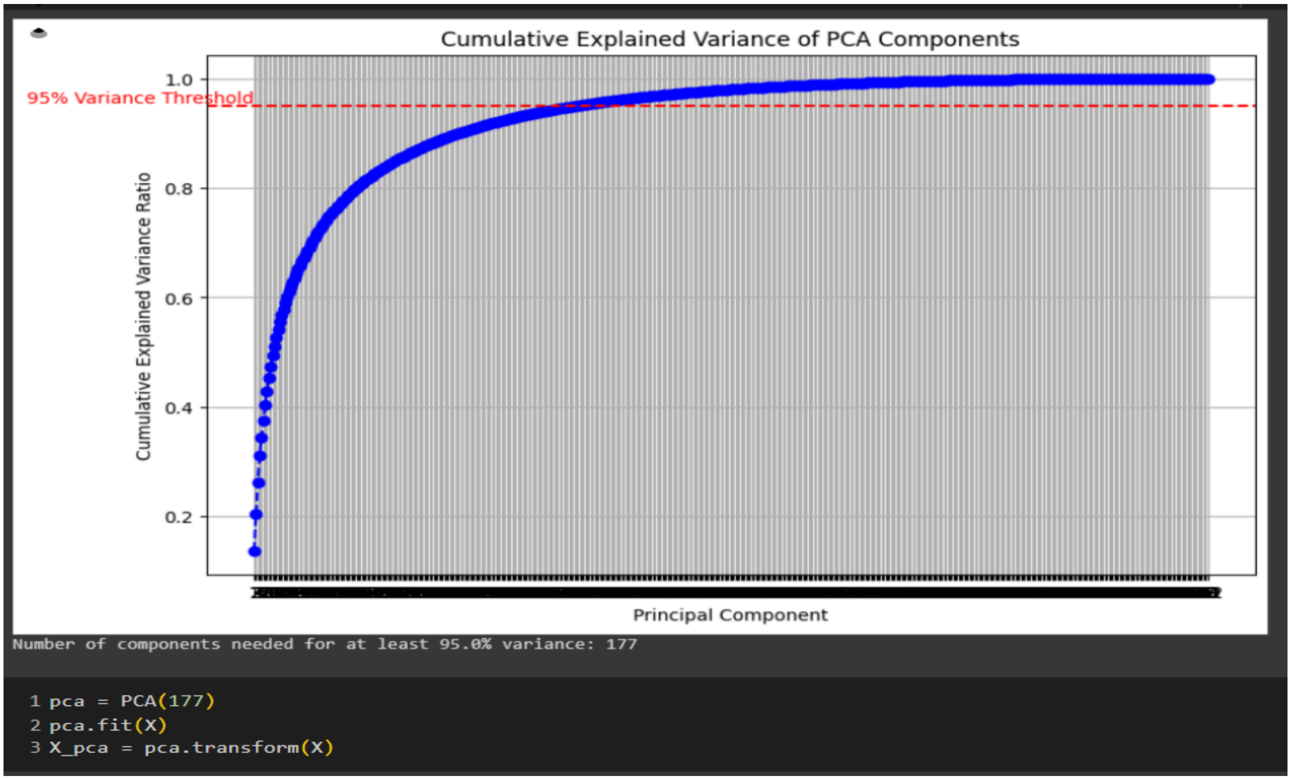
Colour Histogram: Extracted 521 features from an image. The below graph illustrates the distribution of how RGB values have been evaluated. This proved to be a good feature extraction method for my model.



LBP: Extracted 10 features from an image which are shown below. This method extracted features from the image with parameters, P: Number of circularly symmetric neighbor set points and R: Radius of a circle. This also proved to be a good feature extraction method for my model.



Tried PCA :



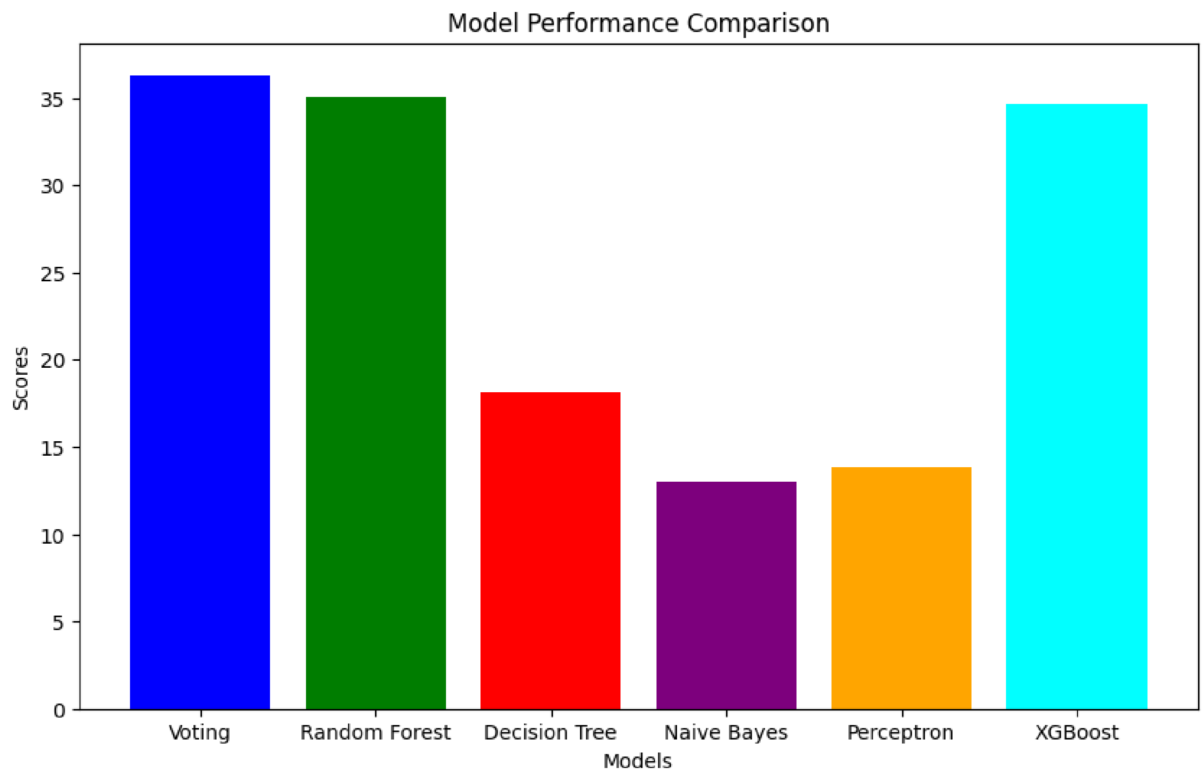
But it reduced the accuracy of my model for the extracted features

Part C) Model implementation and evaluation across 6 different models

The model implementation included feature extraction from Color Histogram + LBP and tried conducting PCA (Principle Components Analysis), but it reduced the accuracy of my model, followed by train-testing split and encoding of the labels, Tried different scaling techniques like MinMax, Standard but No scaled features proved to be better than that.

1) Ensemble Model : Voting Classifier
2) Random Forest
3) XGB Boosting
4) Decision Tree
5) Naive Bayes
6) Perceptron

Observations and Evaluations :



Voting Classifier turned out to be the best model for my features extracted with 36.31% accuracy, which included random forest, decision tree and Naive Bayes in the ratio of 3:1:1 Followed by good accuracy with random forest and XGB Boosting methods. However, Naive Bayes, Perceptron and Decision Tree alone did not prove out to be good models for evaluation,