SECTION-A

Part-(a)

(a) Given

Prior Probabiliy $P(D) = 0.8$

Standard deviation, $\sigma = 36\%$.

μ, mean profit increase

if issued dividends = 10%

if not issued dividends = 0%

we need to find what is the probability of divident given that company with 4% of increase.

i.e $P(D \mid X = 4\%)$

According to Bayes theorem

$$P(D \mid X = 4\%) = \frac{P(X = 4\% \mid D) \cdot P(D)}{P(X = 4\%)}$$

$P(X = 4\%)$ (Total probability) =

$$= P(X = 4\% \mid D)\, P(D) + P(X = 4\% \mid \neg D)\, P(\neg D)$$

Find $P(X = 4\% \mid D)$ and $P(X = 4\% \mid \neg D)$

To calculate these we use PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

we know, $\sigma = 0.36$

$x = 0.04$

$\mu = 0.10$ (in case of D)

$= 0$ (in case of $\neg D$)

$P(X=4\%|D)$

$$= \frac{1}{0.36\sqrt{2\times3.14}}\, e^{-\frac{(0.04-0.1)^2}{2(0.36)^2}}$$

$$= \frac{1}{2.26}\, e^{-\frac{0.003}{0.26}}$$

$$= 0.44\, e^{-0.01}$$

$$= 0.44(0.99) = 0.435$$

$P(X=4\%|7D)$

$$= \frac{1}{0.36\sqrt{2\times3.14}}\, e^{-\frac{(0.04-0)^2}{2(0.36)}}$$

$$= 0.44\, e^{-0.006} = 0.437$$

$$\therefore\ P(X=4\%) = 0.435\times0.8 + 0.437\times0.2$$

$$= 0.35 + 0.08$$

$$= 0.43$$

$$= \frac{0.435 \times 0.8}{0.43}$$

$$= 1.01\times0.8 = 0.80 = 80\%$$

Likelihood = 80%

Part-(b)

Assignment - 2

(b) we know formula of information gain

$$G(X) = H(Y) - H(Y|X)$$

Entropy of entire data $H(Y) = -\sum_{i=0}^{k} P(Y=y_i) \log_2 P(Y=y_i)$

$$H(Y|X) = -\sum_{j=1}^{v} P(X=x_j) \underbrace{\sum_{i=1}^{k} P(Y=y_i|X=x_j) \log_2 P(Y=y_i|X=x_j)}_{\text{Entropy of each classes of the particular attribute.}}$$

DATE
PAGE

Entropy of entire dataset $H(Y)$

$$= -\frac{7}{12}\log\frac{7}{12} - \frac{5}{12}\log\frac{5}{12}$$

$$= -0.58(-0.78) - 0.41(-1.3)$$

$$= 0.45 + 0.53 = 0.98$$

Entropy of all classes in class time feature

$H(Y|\text{Morning})$

$$= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}$$

$$= -0.75(-0.41) - 0.25(-2)$$

$$= 0.31 + 0.5 = 0.81$$

$$= \frac{4}{12}(0.81) = 0.27$$

$$H(Y|\text{NOON}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$$

$$= -0.5(-1) - 0.5(-1)$$

$$= 1 \Rightarrow \frac{4}{12}\times 1 = 0.3$$

$$H(Y|\text{Afternoon}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$$

$$= 1 \Rightarrow \frac{4}{12}\times 1 = 0.3$$

$$G(X) = 0.98 - (0.27 + 0.3 + 0.3)$$

Class time $G(X) = \cancel{0.87}\ 0.11$

DATE
PAGE

Entropy → Had proper sleep.

$$P(Y|\text{Yes}) = -\frac{6}{6}\log\frac{6}{6} - \frac{0}{6}\log\frac{0}{6}$$

$$= 0$$

$$= 0 \times \cancel{0.98} = 0$$

$$P(Y|\text{No}) = -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6}$$

$$= -0.17(-2.55) - 0.83(-0.27)$$

$$= 0.43 + 0.22 = 0.65$$

$$= \frac{6}{12}\times 0.65 = 0.32$$

$$G(X) = 0.98 - 0.32 = 0.66$$

Entropy → Weather

$$P(Y|\text{cool}) = -\frac{4}{5}\log\frac{4}{5} - \frac{1}{5}\log\frac{1}{5}$$

$$= -0.8(-0.32) - 0.2(-2.3)$$

$$= 0.25 + 0.46 = 0.71$$

$$= \frac{5}{12}\times 0.71 = 0.3$$

$P(Y|Rainy) = -\frac{2}{2}\log\frac{2}{2} - \frac{0}{2}\log\frac{0}{2}$

$= 0$

$P(Y|HOT) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$

$= -0.6(-0.7) - 0.4(-1.3)$

$= 0.42 + 0.52 = 0.94$

$= \frac{5}{12} \times 0.94 = 0.43$

$G(x) = 0.98 - (0.3 + 0.43)$

$= 0.98 - 0.73$

$= 0.25$

G(class Time) = 0.11

G(Had proper sleep) = 0.66

G(weather) = 0.25

As, 'Had proper sleep' have high entropy so we will split from here only.

classtime	HPS	weather	Attended ML
Morning	NO	rainy	no
Morning	NO	cool	yes
noon	NO	Hot	no
noon	NO	cool	no
Afternoon	NO	Rainy	NO
Afternoon	NO	HOT	NO

Entire dat.

Entropy of Had Proper sleep(NO) - HPS = H(Y)

$= -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}$

$= -0.17(-2.55) - 0.83(-0.27)$

$= 0.43 + 0.22$

$= 0.65$

Entropy of all attributes

~~H(x)~~ IG for classtime

$H(Y|morning) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= -0.5(-1) - 0.5(-1)$

$= 1$

$= \frac{2}{6}(1) = \cancel{0.3}\ 0.33$

$H(Y|noon) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}$

$= 0$

similarly, $H(Y|afternoon) = 0$

$G(X) = 0.65 - 0.33 = 0.32$

IG for weather

$H(Y|rainy) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}$

$= 0$

$H(Y|cool) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= 1 \times \frac{2}{6} = 0.33$

$H(Y|Hot) = 0$

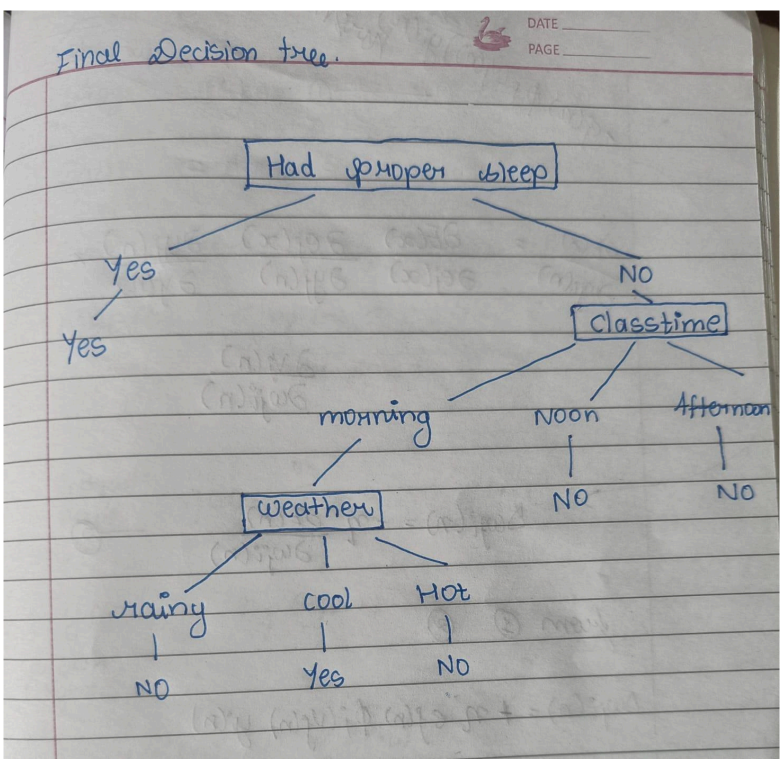
$G(X) = 0.65 - 0.33 = 0.32$

$\therefore$ G(classtime) = 0.32

G(weather) = 0.32

Since the entropy for both attributes are same we can split any of the attribute.

Final decision tree



Part-(c)

DATE

PAGE

(c) for perceptron

$w_t \to$ wt vector at time t

$x_i \to$ sample point, true label $\to y_i$

$y_i(w_i x_i) < r/2$ margin mistake

when incorrectly classified (mistake)

$$w_{t+1} = w_t + y_i x_i$$

let $w_0 = 0$, After 'T' updates:

$$w_T = \sum_{t=1}^{T} y_t x_t$$

$$\Rightarrow \|w_T\|^2 = \|\sum_{t=1}^{T} y_t x_t\|^2 \le T$$

(since $\|x_t\| = 1$)

$\Rightarrow$ let w^*: optimal wt vector that separates the data with margin r

$$w_T \cdot w^* = \sum_{t=1}^{T} y_t (x_t w^*) \ge Tr \ge \frac{Tr}{2}$$

By cauchy-Schwarz

$$w_T \cdot w^* \le \|w_T\| \|w^*\|$$

$\|w_T\| \le T$, let $\|w^*\| = 1$

$$\frac{Tr}{2} \le \sqrt{T}$$

$$\Rightarrow T \le \frac{4}{r^2}$$

$$\therefore T \le \frac{8}{r^2}$$

Part-(d)

(d)

(a) probability estimates for each feature given spam and not spam

$P(\text{Buy} \mid \text{spam}) = \frac{2}{2} = 1$

$P(\text{Buy} \mid \neg\text{spam}) = \frac{1}{2} = 0.5$

$P(\text{cheap} \mid \text{spam}) = \frac{1}{2} = 0.5$

$P(\text{cheap} \mid \neg\text{spam}) = \frac{1}{2} = 0.5$

(b) $P(\text{spam} \mid \text{cheap}, \neg\text{buy}) = \cancel{P(\text{cheap} \mid \text{spam}) \times}$

$$= \frac{P(\text{cheap} \mid \text{spam}) \times P(\neg\text{buy} \mid \text{spam}) \times P(\text{spam})}{P(\text{cheap}) \times P(\neg\text{buy})}$$

$$= \frac{0.5 \times 0 \times 0.5}{0.5 \times 0.25} = 0$$

$P(\neg\text{spam} \mid \text{cheap}, \neg\text{buy})$

$$= \frac{P(\text{cheap} \mid \neg\text{spam}) \times P(\neg\text{buy} \mid \neg\text{spam}) \times P(\neg\text{spam})}{P(\text{cheap}) \times P(\neg\text{buy})}$$

$$= \frac{0.5 \times 0.5 \times 0.5}{0.5 \times 0.25} = \frac{0.125}{0.125} = 1$$

$$= 1$$

so this will be classified as non-spam

(c) For zero probability we can use laplace smoothing.

without laplace

$P(buy=0 \mid spam) = 0$

with laplace

$$P(buy=0 \mid spam) = \frac{0+1}{2+2} = \frac{1}{4} = 0.24.$$

it would result in non-zero probabilities improving robustness of Naive bayes classifier

SECTION-C

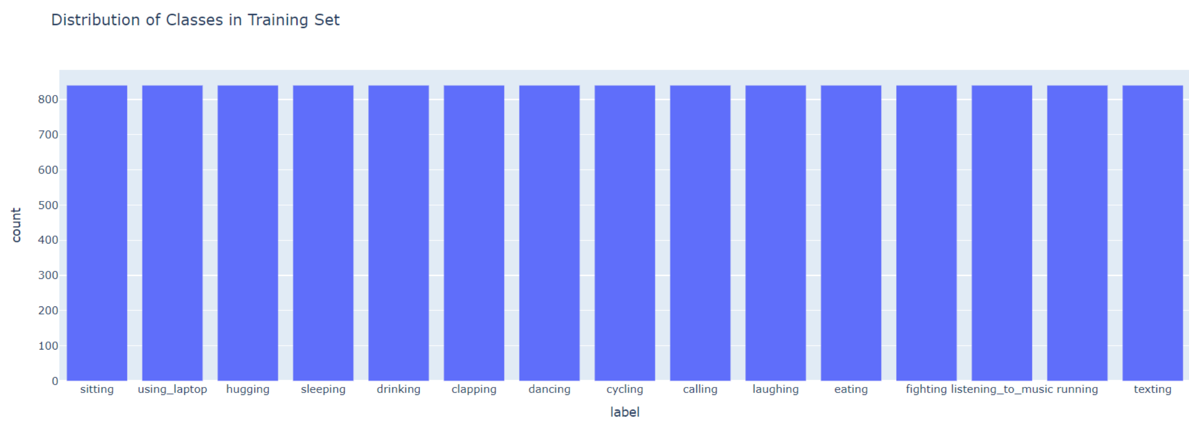
part-(a)

The given data for section c contains images categorized into 15 distinct classes, each with 840 images, making for a well-balanced distribution across categories. These categories represent various human activities such as sitting, using a laptop, hugging, and more.

```
filename       0
label          0
image_path     0
resized_img    0
dtype: int64
```

The dataset's consistency across classes indicates that there is no missing data in terms of filenames, labels, or image paths. Moreover, all images have been resized, showing a uniform preprocessing step applied to the dataset, which might simplify further analysis and model training.

I have visualized some of the images from the dataset and observed:



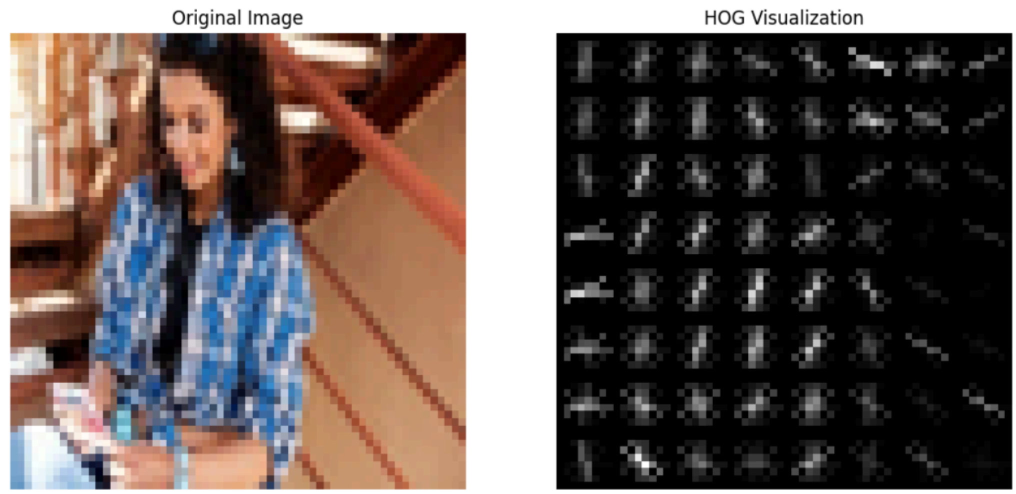
The figure above represents the **distribution of classes** in the dataset. As observed, each class has an identical number of images (840). This **uniform distribution** is a positive aspect since it eliminates the risk of bias during model training, which often occurs when one class is overrepresented compared to others.

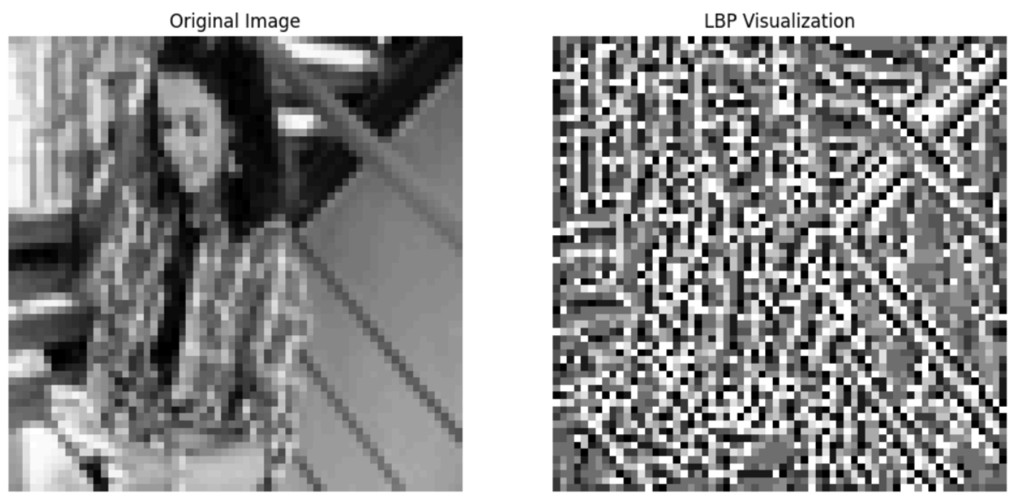
And I have visualized sample images from each class by printing some of them from each class. displaying a few sample images from each class will make it easier to recognize the variety of human actions being classified.

Class Imbalance and Potential Solutions

From the class distribution data, there is no noticeable class imbalance. Each class has exactly 840 images, making the dataset highly balanced. However, if class imbalance were an issue, **strategies such as data augmentation** (rotating, flipping, or cropping images) or **resampling techniques** (undersampling the majority class or oversampling the minority class) could be used to balance the dataset.

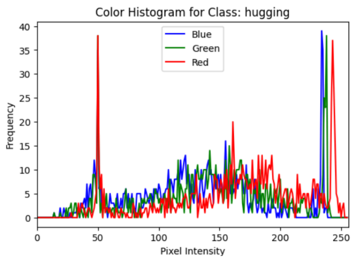
part-(b)

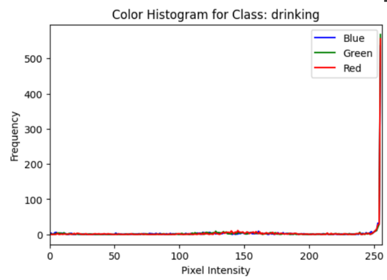


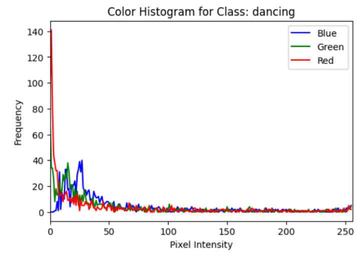


In feature extraction I have extracted several features like hog, lbp, hsv, color Histogram etc.
Histogram of Oriented Gradients (HOG):
HOG captures edge directions and gradients, making it effective for texture and shape recognition, especially in human detection tasks.
The HOG feature map displays an emphasis on structured regions, such as the edges of the stairs, clothing, and contours of the person.

Local Binary Patterns (LBP) Visualization:
LBP is a texture descriptor that encodes local patterns in an image, often used for facial recognition, texture classification, and object detection. he LBP feature map (bottom-right image) shows a dense and intricate texture representation, focusing on pixel-level differences. This feature may not be very useful for human activity recognition task.







Color Histogram
Color histograms summarize the distribution of colors (RGB or other color spaces) in an image. The above graph shows the intensities of the rgb values with their frequency. This feature will help us in this human activity recognition task. Like if we want to predict running then surroundings will be mostly green or white.

By many observations and execution it was observed that from all the features implemented in the code. The most important features are hog and color histogram.

Part-(c)

Now the dataset has been splitted into 80:20 ratio.
I used a random forest classifier in this part and then used a randomized search cv for the search of better accuracy. Among the models tested, **Random Forest** produced the best performance with an accuracy of **32.98%**. Despite the relatively low accuracy, the performance was consistent across most activity classes.

```
Accuracy after RandomizedSearchCV: 32.976190476190474
Classification Report:
                   precision    recall  f1-score   support

          calling       0.32      0.25      0.28       168
         clapping       0.39      0.26      0.31       168
          cycling       0.33      0.58      0.42       168
          dancing       0.34      0.40      0.37       168
         drinking       0.21      0.05      0.08       168
           eating       0.34      0.74      0.46       168
         fighting       0.33      0.57      0.41       168
          hugging       0.24      0.07      0.11       168
         laughing       0.31      0.39      0.35       168
listening_to_music       0.32      0.11      0.16       168
          running       0.38      0.27      0.31       168
          sitting       0.29      0.14      0.19       168
         sleeping       0.44      0.45      0.45       168
          texting       0.24      0.20      0.22       168
     using_laptop       0.31      0.47      0.37       168

         accuracy                           0.33      2520
        macro avg       0.32      0.33      0.30      2520
     weighted avg       0.32      0.33      0.30      2520
```

From the classification report we can say that The overall performance across different classes is moderate, with certain classes like "cycling" and "eating" achieving relatively higher recall and F1-scores. However, other classes like "drinking" and "hugging" performed poorly, indicating that the model struggled to differentiate between certain activities.

Other models like Naive Bayes and Decision Tree performed slightly worse, likely due to the complexity of the data and feature interactions.

So, **Random Forest** achieved the highest accuracy (32.98%) and was chosen as the best-performing model due to its ability to handle high-dimensional features and capture complex relationships within the data.

part-(d)
I have used pickling and stored the best model into the drive.