INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

Machine Intelligence and Expert Systems

TERM PROJECT

Biometric Authentication using mood data of Mouse Dynamics (k-NN Classifier)

PREPARED BY

Tirumala Durga Aravind Kumar Dangeti	(21EC65R03)
Apratim Sen	(18CH10009)
Astitva Sharma	(18EC10005)
Sarvottam Kumar Modi	(18CH10073)
Joel Antony Thomas	(17EC10023)

Problem statement

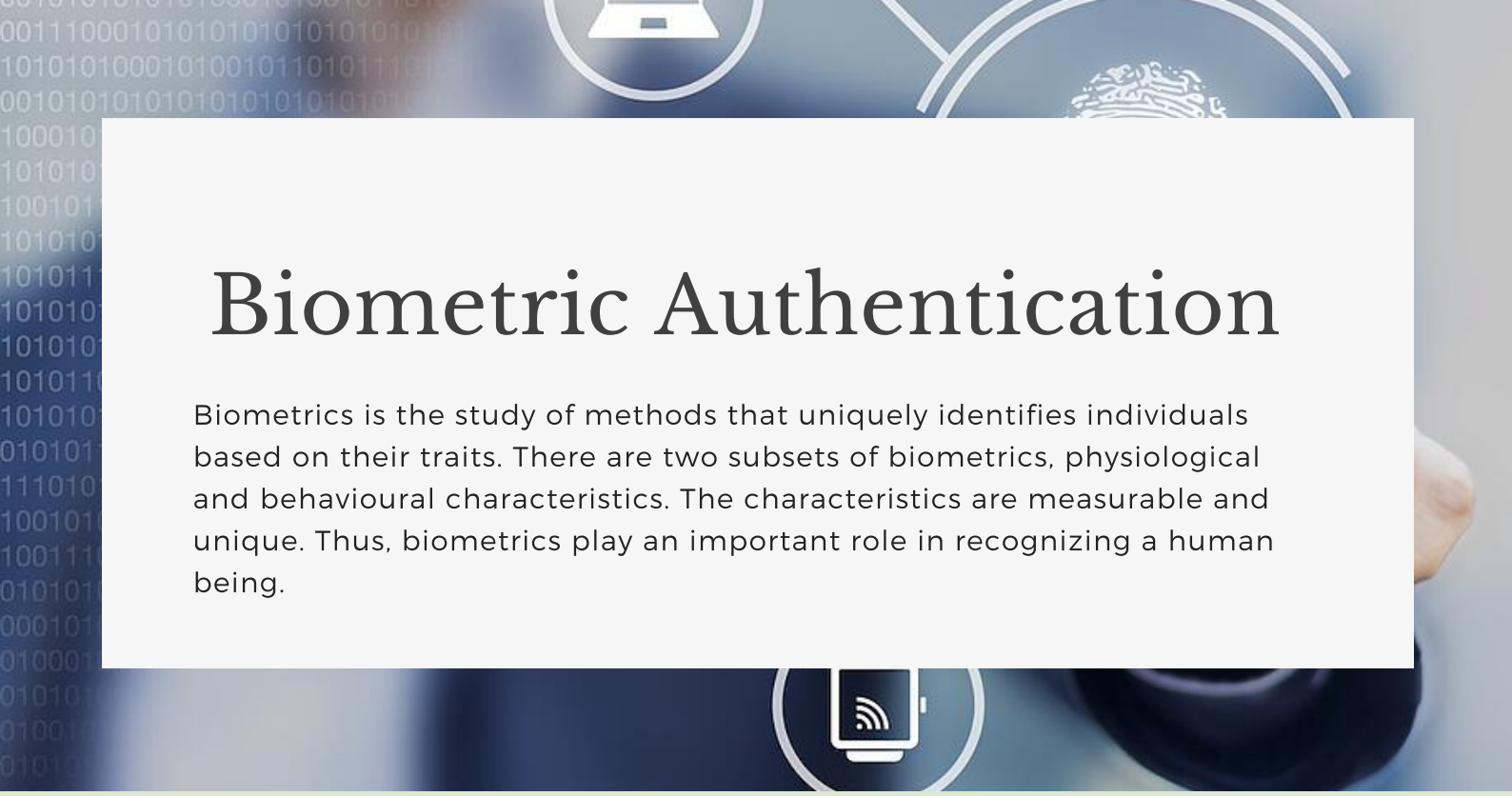
To create a continuous user authentication system for PCs/laptops to prevent threat against intruder, using biometrics involving mouse dynamics using the k-NN classifier by collecting data from your group and the group above yours

Why do we need this?

Today most computer systems identify users by means of passwords and other security phrases. However, once the user has logged in, there is no further means of security or tests which prevent unauthorized usage. Unattended computers with an active session pose a much bigger threat. Users often leave their system unlocked, which allows for unauthorized access.

This allows for three types of attacks.

1. A user of lower clearance can gain access to a terminal with higher clearance and access files or functions of the network to which he is not supposed to have access to.
2. User with the same or higher clearance can conceal his identity by performing malicious actions under the guise of a coworker.
3. A person who is not affiliated with the company in anyway can gain access to the internal network.



Biometric Authentication

Biometrics is the study of methods that uniquely identifies individuals based on their traits. There are two subsets of biometrics, physiological and behavioural characteristics. The characteristics are measurable and unique. Thus, biometrics play an important role in recognizing a human being.

Requirements of a Biometric

1. **Universality**: Every person should have the characteristics.
2. **Uniqueness**: No two person should have the same biometric characteristics.
3. **Permanence**: The biometric should not be variant with time. Behavioral biometrics can be highly variable with time.
4. **Collect-ability**: The characteristics must be measurable quantitatively and obtaining the characteristics should be easy.
5. **Performance**: Acceptable accuracy should be achieved after identification/ verification.
6. **Circumvention**: This property indicates to how difficult it is to fool the system by fraudulent techniques.

Types of Biometrics

The biometrics can be broadly classified into two types:

1. Physiological biometrics
2. Behavioral biometrics

Physiological Biometrics

Physiological biometrics involve physiological characteristics of a human being used as biometric. These biometrics are more reliable and accurate as they are not affected by any mental conditions such as stress or illness.

Example of Physiological Biometrics

1. **Fingerprint scanning** : Everyone has patterns of friction ridges on their fingers, and it is this pattern that is called the fingerprint. Fingerprints are uniquely detailed, durable over an individual's lifetime, and difficult to alter. Because there are countless combinations, fingerprints have become an ideal means of identification.

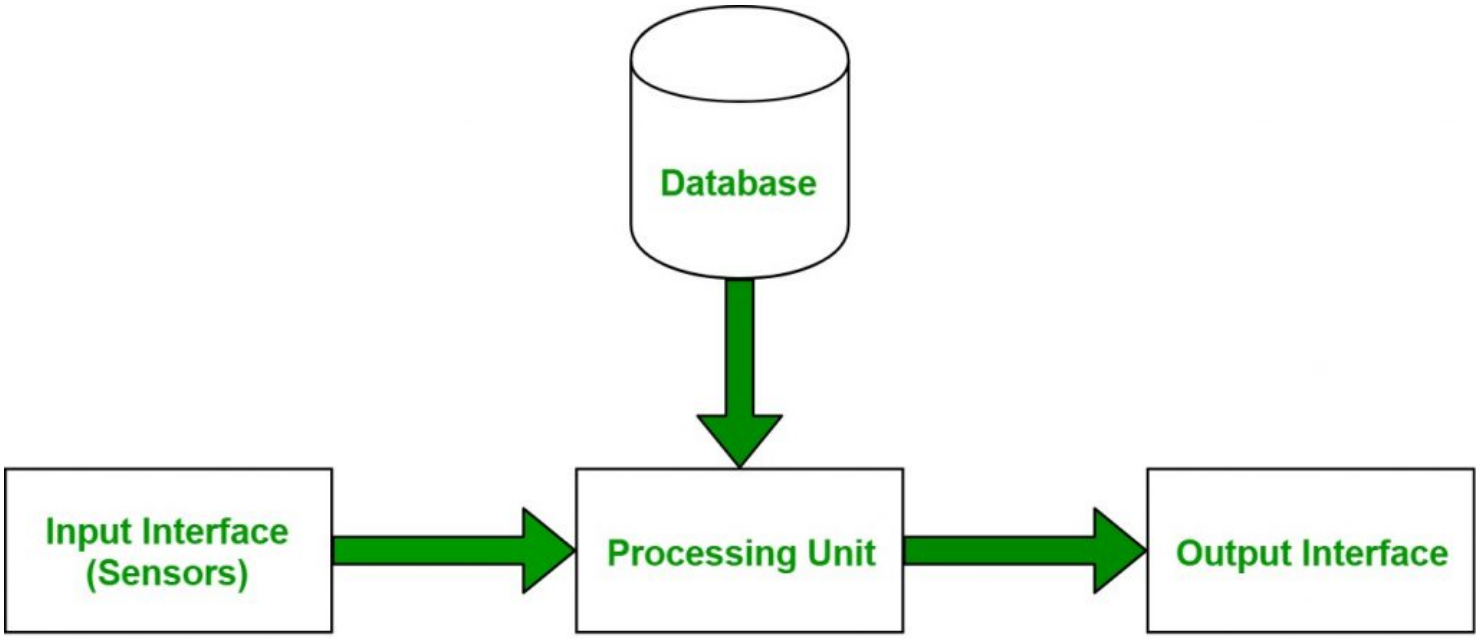
2. **Iris scans** : is an automated method of biometric identification that uses mathematical pattern-recognition techniques on video images of one or both of the irises of an individual's eyes, whose complex patterns are unique, stable, and can be seen from some distance. False matches are very low in this biometric.

3. **Hand geometry recognition**: is a biometric that identifies users from the shape of their hands. Hand geometry readers measure a user's hand along many dimensions including height, width, deviation, and angle and use it for biometric security.

4. **Facial recognition**: is a technology capable of matching a human face from a digital image or a video frame against a database of faces, typically employed to authenticate users through ID verification services, works by pinpointing and measuring facial features from a given image.

With the help of special devices (scanners, sensors, and other readers), a person's biometric data is stored in a database. The system saves this information, such as a fingerprint, and converts it into digital data. Then, when the finger is placed back on the scanner, the system compares the new data with what is stored in its database. Finally, the system will either confirm the person's identity and grant them access if there is a match or decline the request if not.

Modern smartphone cameras and video recorders can easily recognize faces with the help of built-in sensors powered by neural networks. In this sense, the image becomes a person's identifier. This technology can be used not only to unlock phones, but also for more complex tasks such as confirming purchases or accessing financial services.



Behavioral biometrics

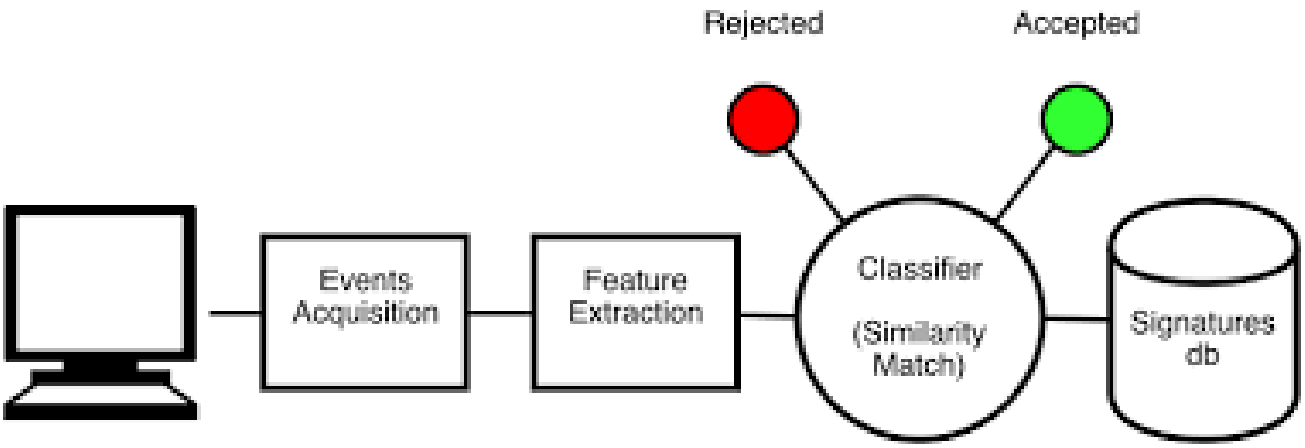
Behavioral biometrics is a recognition system that identifies a person based on dynamic or behavioral characteristics. This type is also known as passive biometrics, as it doesn't require a user's active participation to proceed with the authentication process. These dynamic authentication methods are based on the characteristics of a person's behavior. They evaluate a person's unique behavior and subconscious movements in the process of reproducing any action.

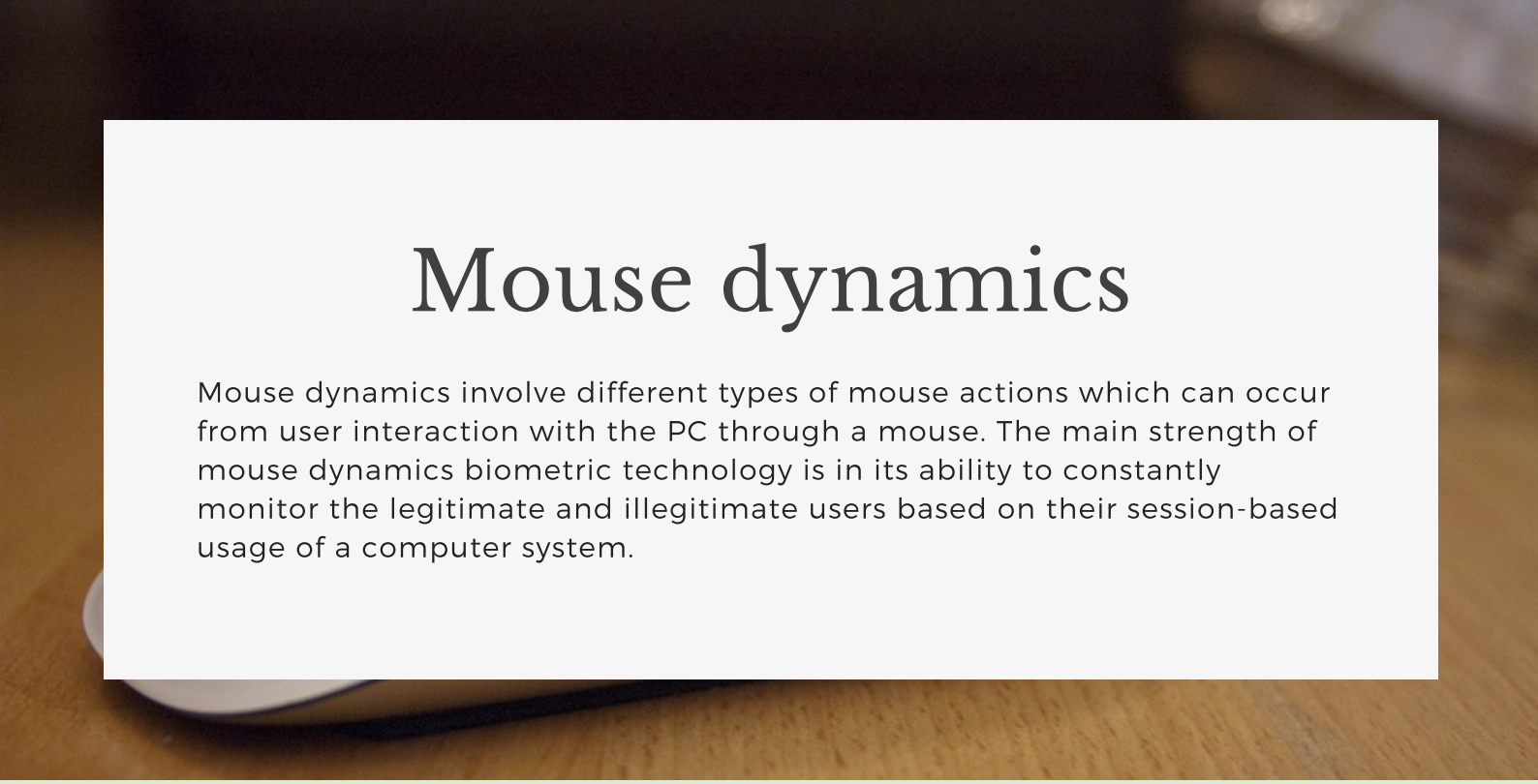
Example of Behavioral Biometrics

1. **Handwriting and Signature dynamics:** handwriting and signatures are partly unique to each individual, and can be used as a biometric
2. **Voice and speech rhythms**: The way each individual talks is unique, and can be used as a biometric.
3. **Gesture recognition:** identifying a user based on their gestures
4. **Walking dynamics:** the way how each individual walk.
5. **Mouse Dynamics:** This behavioral biometric is characterized by the way an individual moves the mouse or clicks on the screen of the desktop/laptop. Mouse actions like mouse movements, clicks, drag and drop etc. can be used as useful features.
6. **Keystroke dynamics:** refer to the detailed timing information which describes exactly when each key was pressed and when it was released as a person is typing at a computer keyboard.

Components of Behavioral Biometric Identification system

1. Feature extraction which captures the data generated by standard input devices such as a mouse or a keyboard.
2. Feature extraction and classifier module that constructs the users signature based on his/her behavioral biometrics.
3. A signature database consisting of behavioral signatures of registered users.





Mouse dynamics

Mouse dynamics involve different types of mouse actions which can occur from user interaction with the PC through a mouse. The main strength of mouse dynamics biometric technology is in its ability to constantly monitor the legitimate and illegitimate users based on their session-based usage of a computer system.

Possible mouse actions

1. **Mouse Move**: Mouse move is a simple movement involving no clicks. Mouse move can be between two click events or non-click events.
2. **Drag and Drop:** It is the action which starts by a mouse button held down followed by a movement and finally the button released. Generally, it is used to move/copy a file to a particular location.
3. **Point and Click**: It is a movement of the mouse ending in a click.
4. **Silence**: This action suggests no mouse movement.

Feature Extraction

The collected behavior patterns cannot be used directly by a detector or classifier. Instead, dynamic characteristics are extracted from these patterns.

Features which can be extracted

1. **Click Time**: It is the time required for the user to click a button.
2. **Pause Time**: It is the amount of time spent pausing between pointing to an object and actually clicking on it.
3. **Horizontal Velocity:** Horizontal Velocity is change in X coordinate value for the given change in time
4. **Vertical Velocity**: Vertical Velocity is change in Y coordinate value for the given change in time.
5. **Straightness**: The straightness feature characterizes the nature of movement of the mouse.

SPATIAL INFORMATION

1. Horizontal coordinates
2. Vertical coordinates
3. Path distance from the origin
4. Angle of the path with respect to X axis
5. Curvature of the path
6. Derivative of the curvature of the path

TEMPORAL INFORMATION

1. Input x values
2. Input y values
3. Input t values
4. Horizontal velocity
5. Vertical velocity
6. Tangential velocity
7. Tangential acceleration
8. Tangential jerk
9. Angular velocity

k-NN

k-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. k-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears, then it can be easily classified into a well suite category by using K- NN algorithm. It stores the dataset and at the time of classification, it performs an action on collected dataset and gives the required output.

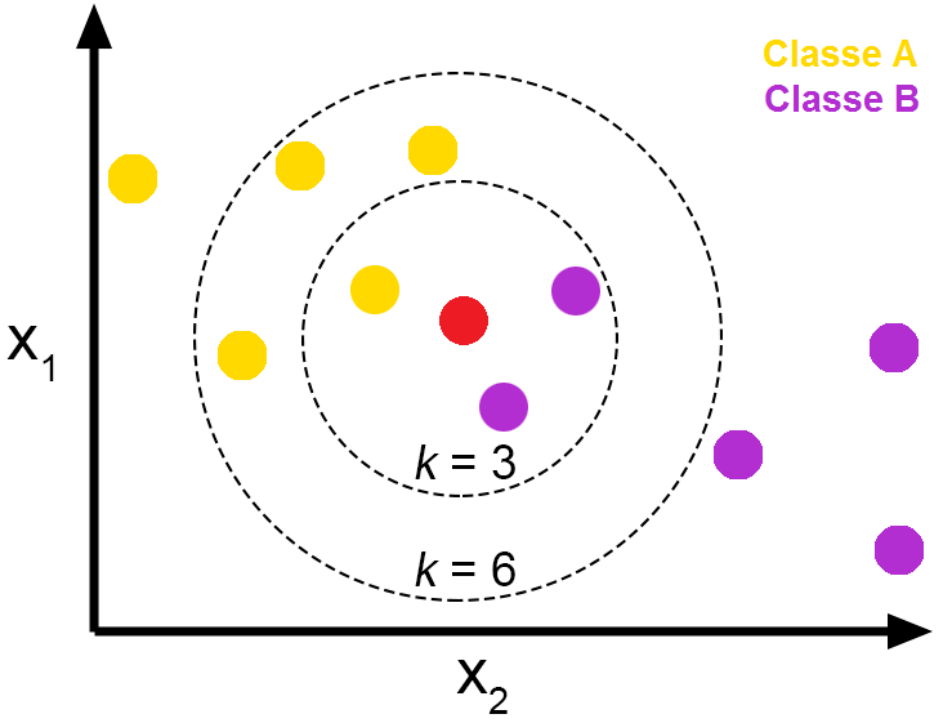
k-NN Algorithm

1. Select the number K of the neighbors
2. Calculate the Euclidean distance of K number of neighbors
3. Take the K nearest neighbors as per the calculated Euclidean distance.
4. Among these k neighbors, count the number of the data points in each category.
5. Assign the new data points to that category for which the number of the neighbor is maximum.

Note

Many people get confused between K-mean and K-nearest neighbor. Some of the differences are as follows-

1. K-mean is an unsupervised learning technique (no dependent variable) whereas KNN is a supervised learning algorithm (dependent variable exists)
2. K-mean is a clustering technique which tries to split data points into K-clusters such that the points in each cluster tend to be near each other whereas K-nearest neighbor tries to determine the classification of a point, combines the classification of the K nearest points



Pros of k-NN

1. Easy to understand
2. No assumptions about data
3. Can be applied to both classification and regression
4. Works easily on multi-class problems

Cons of k-NN

1. Memory Intensive / Computationally expensive
2. Sensitive to scale of data
3. Not work well on rare event (skewed) target variable
4. Struggle when high number of independent variables

Data & Results

```
**********************************************************************************
LOGGING TIME: 20211019_153420
CLIENT IP: 10.145.38.146
USERNAME: Joel
OS: Windows 10
**********************************Emotional
database**********************************************************************************
MM, 812, 675, 1636382539942
MM, 802, 667, 15
MM, 800, 665, 1
MM, 798, 664, 3
MM, 795, 662, 3
MM, 10, 31, 6
MM, 2, 24, 5
MM, 776, 645, 2
MM, 773, 642, 2
MM, 770, 639, 1
MM, 767, 636, 2
MM, 764, 632, 3
MM, 761, 629, 3
MM, 758, 626, 1
MM, 754, 621, 1
MM, 749, 618, 2
MM, 742, 610, 2
MM, 737, 606, 2
MM, 731, 601, 2
MM, 32, 24, 3
MM, 18, 15, 4
MM, 11, 10, 2
```

Collected data looks like this

Data & Results

Notation	Meaning
MC, n, t:	*Mouse Clicked, Click count, Relative time*
MP, n, t:	*Mouse Pressed, Button ID, Relative time*
MR, n, t:	*Mouse Released, Button ID, Relative time*
MM, x, y, t:	*Mouse Moved, x-coordinate, y-coordinate, Relative time*
MD, x, y, t:	*Mouse Dragged, x-coordinate, y-coordinate, Relative time*
MWM, x, y, w, a, s, t:	*Mouse Wheel Moved, x-coordinate, y-coordinate, Wheel rotation sense, Amount of scrolling, Scroll type, Relative time*

This is what the notation means

Data & Results

```
Pre-processing Done.

Count of different classes in Train set:
3    5276
1    2925
0    2177
4    1976
2    1568
5    1157
6     934
Name: Class, dtype: int64

Count of different classes in Test set:
3    1320
1     732
0     545
4     494
2     393
5     290
6     234
Name: Class, dtype: int64

Implementing K-Nearest Neighbors Model.

Number of mislabeled points out of a total 4008 points : 2248, Accuracy: 43.91218%

5 Fold Cross Validation Accuracy on Training Set: 23.861093962217556
```

When we test the k-NN classifier on the collected data (both from our own group, and from Group 16), we get testing accuracies of around 45 – 60%, and a k-fold(k=5) accuracy of around 30%.

As we change the value of k in the classifier, we are presented with different values for the accuracies, both final and k-fold.

Why the low accuracy?

Factors Affecting Performance

The data collection and the data processing may not be ideal and there are various factors, both human and computer, which could add to the error and the inconsistency of the results.

1. Environmental Conditions

a. Height of chair
b. Distance between mouse and body
c. Touchpad vs conventional mouse

2. User conditions

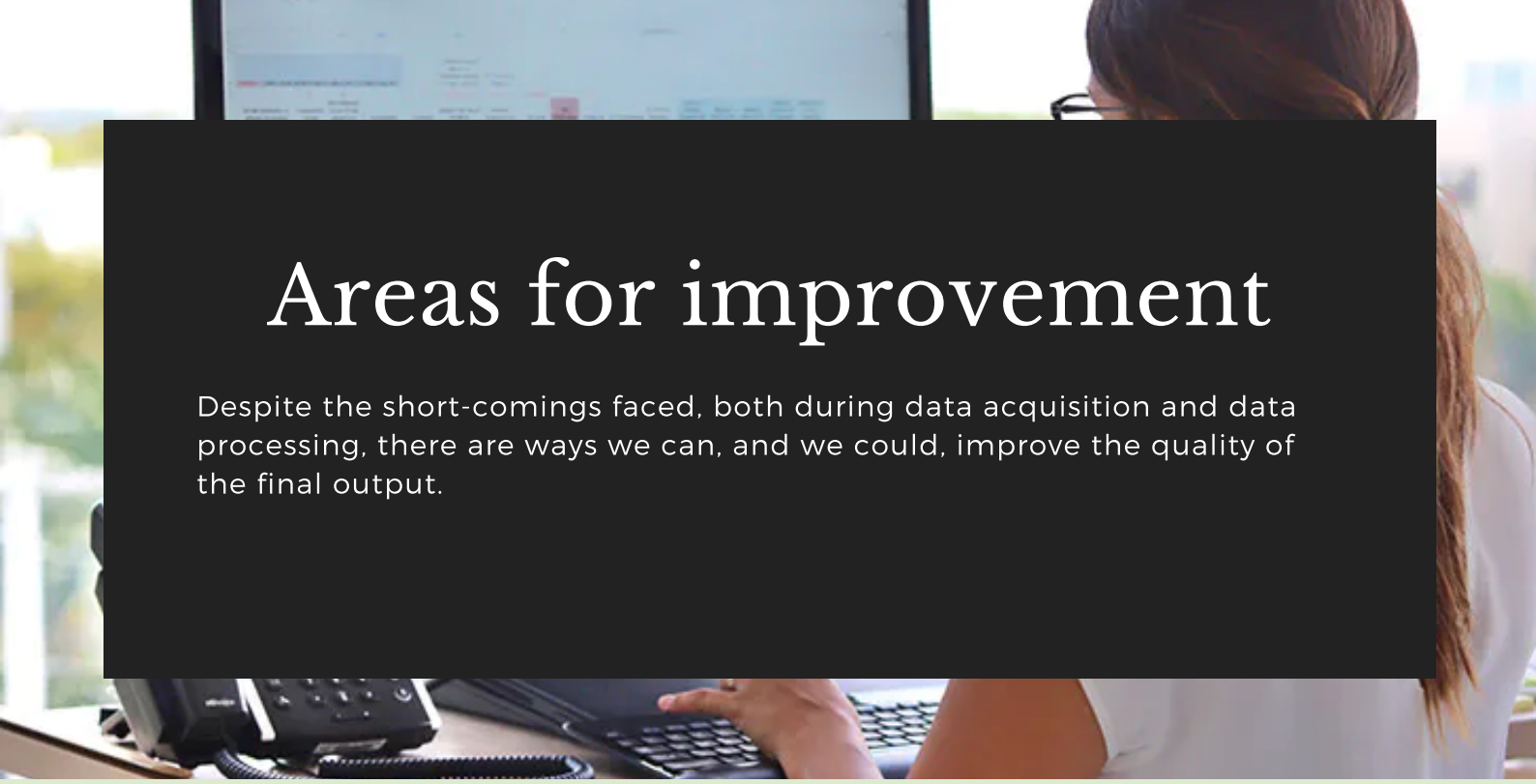
a. Mood
b. Knowledge & practice of application
c. Typing errors

3. GUI/mouse setting

a. Screen resolution
b. Pointer speed

4. Noise

a. Hardware error
b. Software error



Areas for improvement

Despite the short-comings faced, both during data acquisition and data processing, there are ways we can, and we could, improve the quality of the final output.

1. Mood analysis

a. A separate label for mood could constitute a useful feature

2. Data collection

a. More amount of data

b. Standard environment

c. Standard computer settings

3. Data Preprocessing

a. Noise removal, smoothening and error nullification

4. Training

a. State-of-the-art classification algorithms