**Assessing Impact of Digital Lens Usage on Eye Dryness using Schirmer's Effect**

CAPSTONE PROJECT REPORT

By

Team 16

Naveen Rampa [21R11A0589]

J Sinduja [21R11A0574]

E Saikiran [22R15A0514]

Lakshanya S [21R11A05H3]

A Bhavya Sri [21R11A05A5]



**Department of Computer Science and Engineering**

**Accredited by NBA**

**Geethanjali College of Engineering and Technology**

**(UGC Autonomous)**

(Affiliated to J.N.T.U.H, Approved by AICTE, New Delhi)

Cheeryal (V), Keesara (M), Medchal.Dist.-501 301

March 2024

**ABSTRACT**

Digital eye strain (DES) is a growing concern in today's digital age, caused by excessive use of electronic devices like computers and smartphones. This condition, marked by various eye and vision problems, has become more common due to remote work and online learning, especially during the COVID-19 pandemic. This project focuses on understanding the risk factors of DES among college students, who often spend long hours using digital screens for studying and leisure activities.

The goal is to analyze survey data and develop a model to predict DES risk factors. Using tools like Anaconda Python Jupyter, we explore the dataset, clean the data, and build models to find the best predictors of DES. The project also involves creating a user-friendly web application with Flask API, making it easy for users to access and utilize the model's predictions.

The project aims to provide insights into DES prevalence and its impact on college students, along with suggestions for reducing its effects. The final report will detail our research methods, findings, and recommendations, contributing to a better understanding of DES and guiding future efforts to address this issue.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Chapter Name** | **Page No** |
|  | **ABSTRACT** | 2 |
| **1** | **INTRODUCTION** | 4 |
| **2** | **PROBLEM STATEMENT** | 5 |
| **3** | **OBJECTIVES** | 5 |
| **4** | **METHODOLOGY** | 6 |
|  | **4.1 Data Source** | 6 |
|  | **4.2 Exploratory Data Analysis** | 9 |
|  | 4.2.1 Checking for Data Consistency | 11 |
|  | 4.2.2 Revised data frame | 12 |
|  | 4.2.3 Data Visualization | 13 |
|  | **4.3 Data Pre-Processing** | 24 |
|  | 4.3.1 Replacing Null values | 24 |
|  | 4.3.2 Removing Rows with Incorrect Data | 25 |
|  | 4.3.3 Making the Target Variable Categorical | 25 |
|  | 4.3.4 Dividing Multiple Target Variables | 27 |
|  | 4.3.5 Creating Training and Testing Datasets | 27 |
| **5** | **ALGORITHMS** | 28 |
| **6** | **Results(Accuracy) obtained for the models** | 30 |
| **7** | **Final Implementation** | 31 |
|  | **Conclusion** | 33 |

**1. INTRODUCTION**

The Schirmer's eye test is a widely used diagnostic tool in ophthalmology to assess tear production and diagnose conditions related to dry eye syndrome. Named after the German ophthalmologist Otto Schirmer, who developed the test in the early 20th century, it remains a cornerstone in evaluating ocular surface health. Dry eye syndrome is a common condition characterized by insufficient tear production or poor tear quality, leading to symptoms such as eye discomfort, redness, itching, and blurred vision. The Schirmer's test provides valuable information about tear film dynamics by measuring the amount of tears produced over a specified time period.



Fig 1.1: Schirmer’s eye test

Not detecting Schirmer's eye abnormalities and addressing dry eye syndrome in its early stages can result in worsening symptoms, ocular surface damage, visual impairment, and increased susceptibility to infections and complications. Early diagnosis and appropriate management are essential for preserving ocular health, maintaining visual function, and improving patients' quality of life.

**2. PROBLEM STATEMENT**

Understanding the impact of digital screen usage on eye health and associated symptoms involves analyzing various factors like age, duration of screen time, online platforms, nature of activities, screen illumination, years of exposure, daily screen hours, types of devices used, distance from the screen, nighttime usage, blinking frequency, difficulty in focusing, frequency and severity of complaints, observed ocular symptoms, and specific eye examination results. By examining these variables we aim to uncover patterns and correlations to develop strategies for maintaining optimal eye health in the digital age.

**3. OBJECTIVES**

1. To develop a predictive model for detecting abnormalities in Schirmer's eye test results.

2. To classify Schirmer's test outcomes as normal or abnormal using machine learning classification techniques.

3. To compare and evaluate different machine learning classification algorithms to determine the most effective model for detecting Schirmer's test abnormalities.

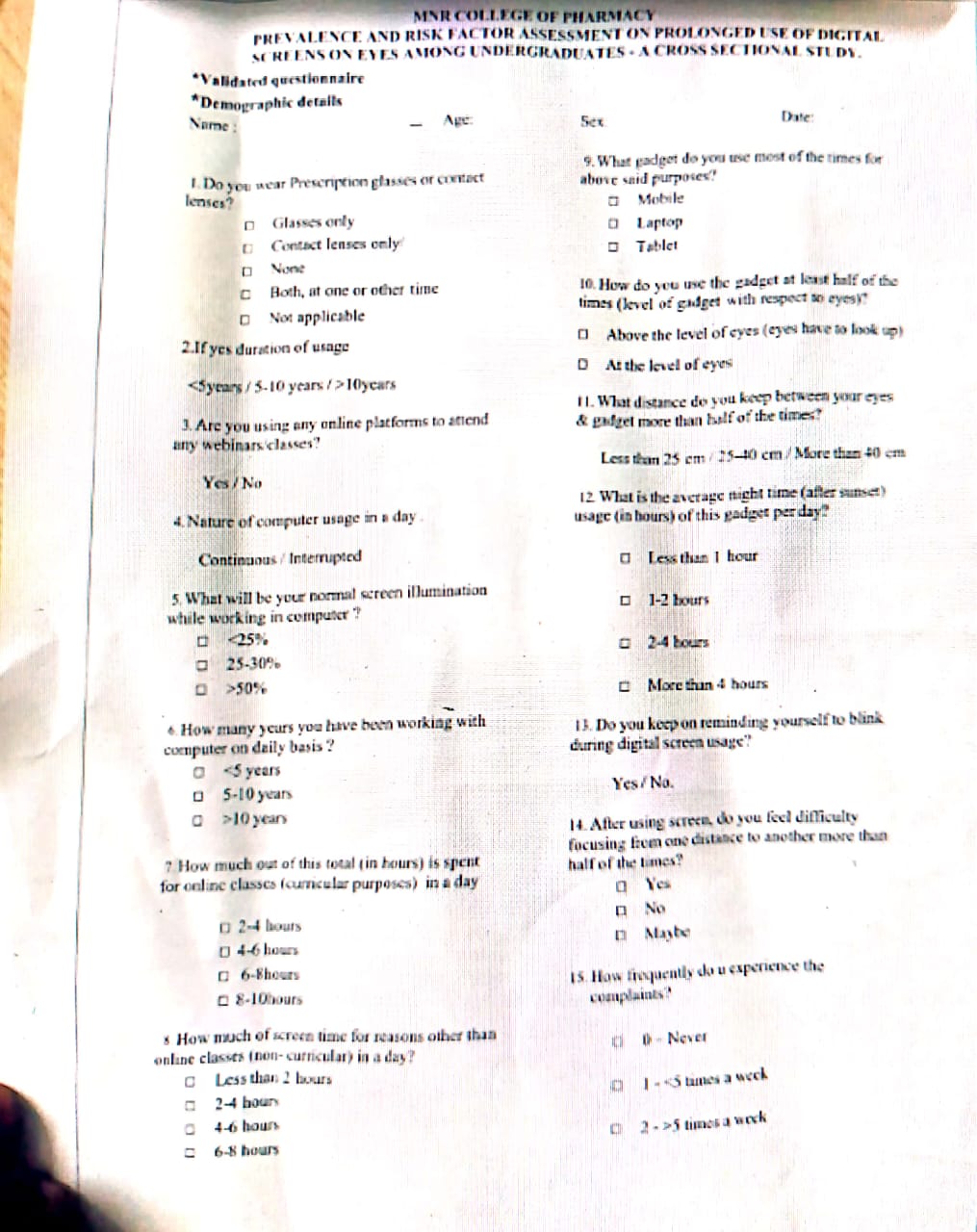
4. To present key performance metrics, such as accuracy, for evaluating the predictive performance of the developed models.

5. Create a user-friendly web app to visualize and use the developed model.

**4. METHODOLOGY**

**4.1 Data Source**

The dataset is procured from a survey and a form which represents the values of the columns is also given.



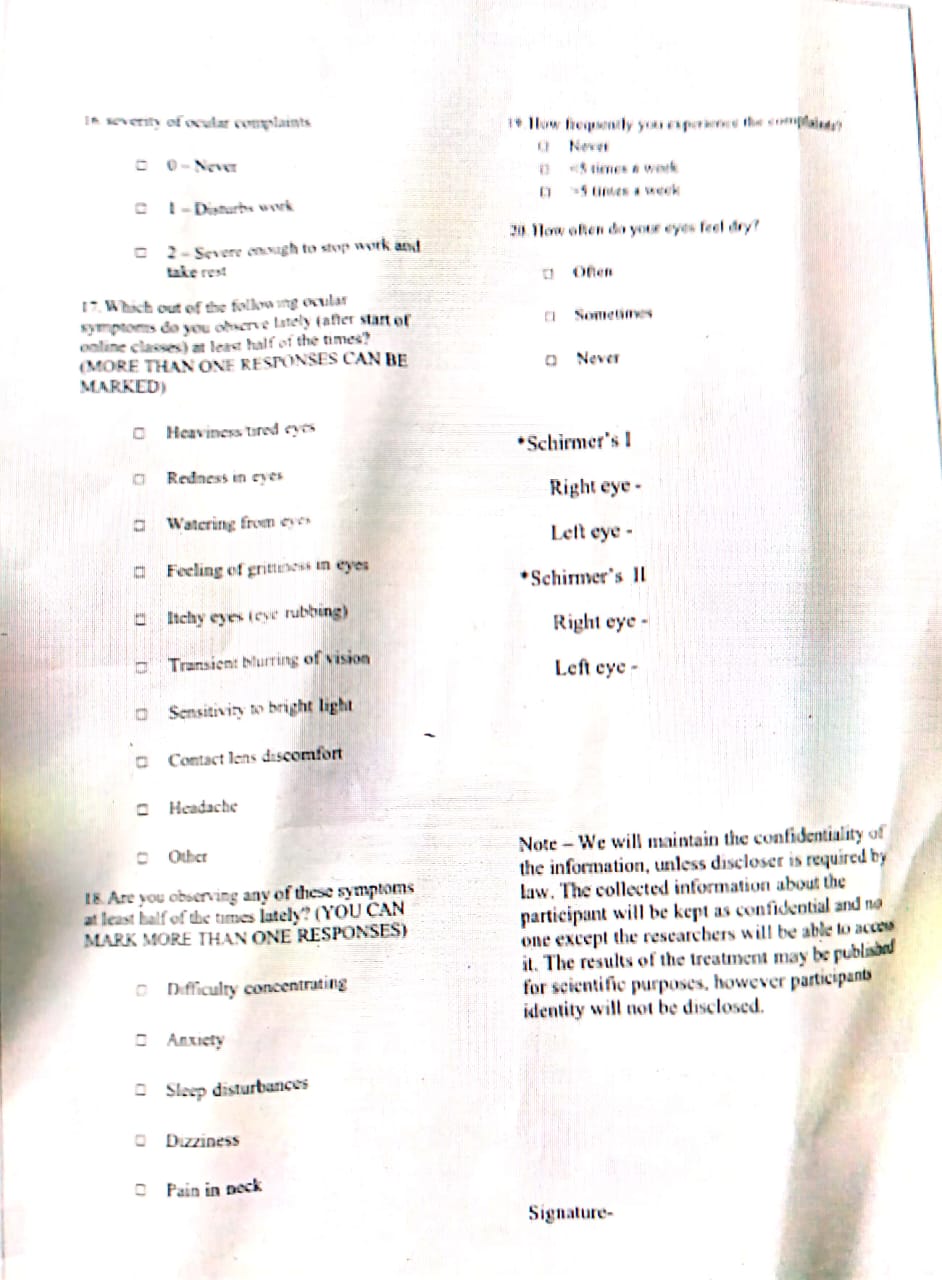


Fig : Survey form related to the given dataset.

The data set provided to us comprised of 28 features and 300 observations

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Feature Name** | **Description** |
| 1 | Name | Identifier for the individual participating in the study |
| 2 | Age | Age of the participant |
| 3 | Sex | Gender of the participant |
| 4 | wearables | Use of wearable devices |
| 5 | Duration | Usage of wearables over a period of time |
| 6 | onlineplatforms | Whether the participants use online platforms or not |
| 7 | Nature | Nature of computer usage in a day |
| 8 | screenillumination | Intensity of screen illumination |
| 9 | workingyears | Number of years working with digital screens on daily basis |
| 10 | hoursspentdailycurricular | Time spent daily on curricular activities |
| 11 | hoursspentdailynoncurricular | Time spent daily on non – curricular activities |
| 12 | Gadgetsused | Type of electronic gadget used |
| 13 | levelofgadjetwithrespecttoeyes | Height or level of electronic gadgets relative to the eyes |
| 14 | Distancekeptbetweeneyesandgadjet | Proximity of electronic gadgets to the eyes |
| 15 | Avgnighttimeusageperday | Average night time usage of the gadget |
| 16 | Blinkingduringscreenusage | Reminding oneself to blink during digital screen usage |
| 17 | Difficultyinfocusingafterusingscreens | Difficulty in focusing vision after screen usage |
| 18 | freqquencyofcomplaints | Frequency of reported ocular complaints |
| 19 | Severityofcomplaints | Severity of reported ocular complaints |
| 20 | RVIS | Visual discomfort or discomfort related to vision |
| 21 | Ocularsymptomsobservedlately  Symptomsobservingatleasthalfofthetimes | Recent observation of ocular symptoms |
| 22 | Ocularsymptomsobservedlately  Symptomsobservingatleasthalfofthetimes | Frequency of experiencing symptoms |
| 23 | Complaintsfrequency | Frequency of reported ocular complaints |
| 24 | frequencyofdryeyes | Frequency of experiencing dry eyes |
| 25 | Schimers1Lefteye | Measurement of tear production in the left eye using Schirmer's test. |
| 26 | Schimers1righteye | Measurement of tear production in the right eye using Schirmer's test |
| 27 | Schimers2Lefteye | Measurement of tear production after anesthetic instillation in the left eye using Schirmer's test |
| 28 | Schimers2righteye | Measurement of tear production after anesthetic instillation in the right eye using Schirmer's test |

**4.2 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an essential step in the machine learning process, where the main goal is to analyze and understand the characteristics of the data before applying any modeling techniques. EDA is the process of summarizing the main characteristics of the data, such as the distribution, the relationship between variables, and identifying any patterns or anomalies that may exist.

EDA is an important step because it allows us to gain a deeper understanding of the data and the underlying relationships between variables, which can help inform decisions about feature engineering, data preprocessing, and model selection. By performing EDA, we can identify any missing or erroneous data, outliers, and inconsistencies in the data, which can be addressed before training any machine learning models.

* **Information about the Features & their data types**

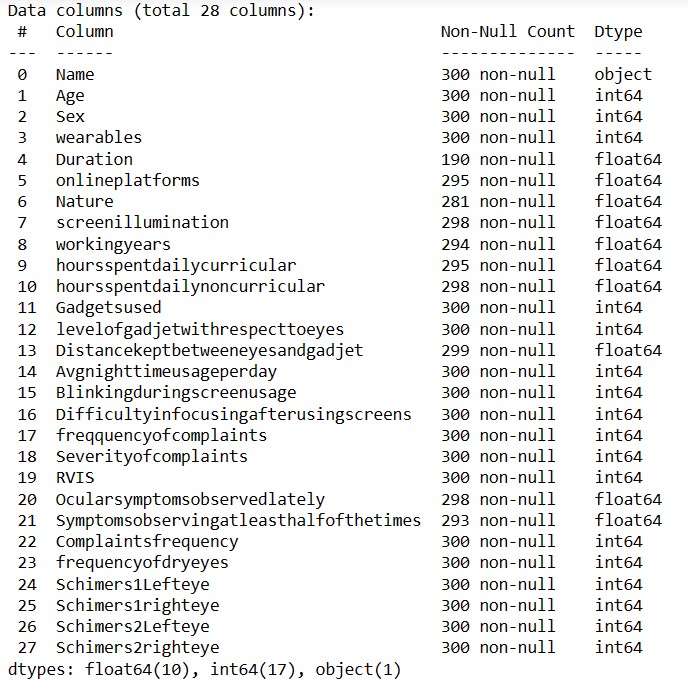


Fig 4.2.1: Dataset Information

* **Observations:**
* Figure 4.2.1 outlines the data types and non-null entry counts of the dataset columns. Upon initial review, it's evident that the 'Name' column is categorical, while others appear numerical but are also categorical. Considering the survey form, we infer that these numerical values represent coded options. Further analysis of unique values will validate this observation.
* In the given dataset, there are missing values. Addressing these missing values requires data pre-processing, where we either remove or replace them.
* Within the dataset, multiple target variables include “Schimers1Lefteye, Schimers1righteye”, and “Schimers2Lefteye”, “Schimers2righteye”. These variables likely represent different measurements or aspects related to eye examinations.

**4.2.1 Checking for Data Consistency**

* No duplicates found.



Fig 4.2.1.1: Number of Duplicate rows in a dataset

* Unique Values.

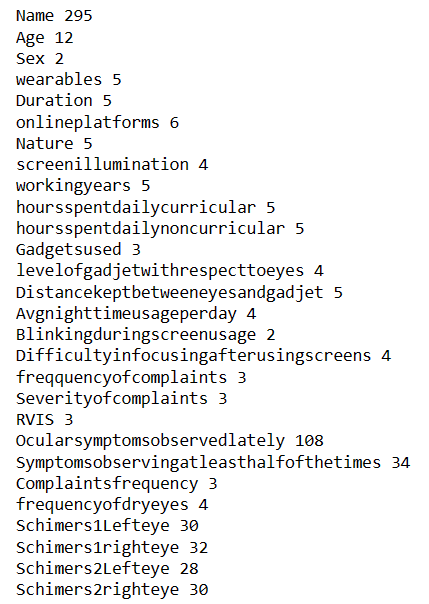


Fig 4.2.1.2: Unique value count for each column

**Observations:** From the above unique value count (Fig 4.2.1.2) we can observe that Name has almost all the values unique and since it acts as an identifier it can be dropped from the dataset.

Other than the target columns, 'Name' column, ‘Ocularsymptomsobservedlately’, ‘Symptomsobservingatleasthalfofthetimes’ the remaining columns contain a very limited number of unique values. This proves that the numerical values are label encoded and each value correspondes to the number of options given in the survey form.

Target variables(“Schimers1Lefteye, Schimers1righteye”, and “Schimers2Lefteye”, “Schimers2righteye”) are numeric in nature.

* More about ‘Ocularsymptomsobservedlately’ and ‘Symptomsobservingatleasthalfofthetimes’

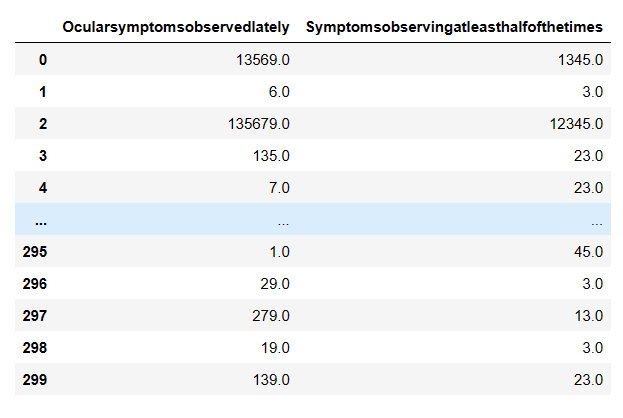


Fig 4.2.1.3: Values of the columns ‘Ocularsymptomsobservedlately’ and ‘Symptomsobservingatleasthalfofthetimes’

Upon initial observation, the columns appear to contain typical numerical values. However, upon cross-referencing with the survey form provided to us, it became apparent that these columns actually represent multiple options concatenated together. This realization occurred because these two columns allow respondents to select multiple options from a list of choices.

**4.2.2 Revised data frame**

After dropping column ‘Name’, the data frame looks like below.

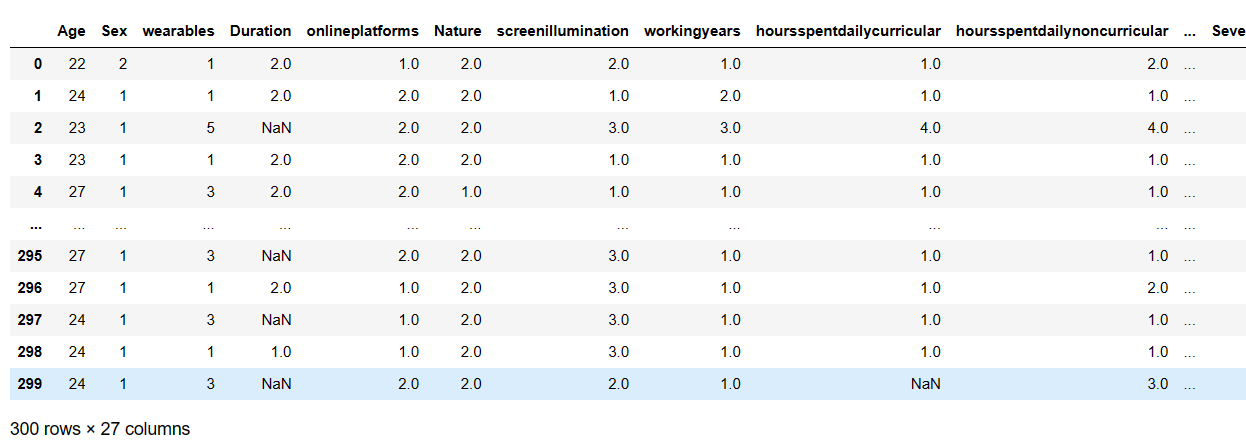


Fig 4.2.2.1: Revised DataFrame.

**4.2.3 Data Visualization**

* **Analysing the target variables**

1. **“Schimers1Lefteye”**

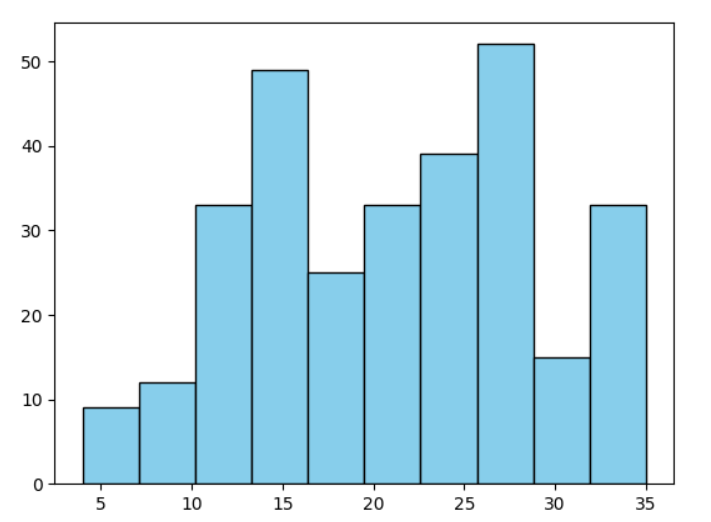
****

Fig 4.2.3.1: Histogram showcasing Distribution of values in Schimers1Lefteye.

**Observations:**

* The x-axis represents the values (ranging from 0 to 35).
* Prominent peaks are visible around the values of 15 and 30 on the x-axis.
* Valleys or lower points are noticeable around 10, 20, and 35 on the x-axis.

1. **“Schimers1righteye”**

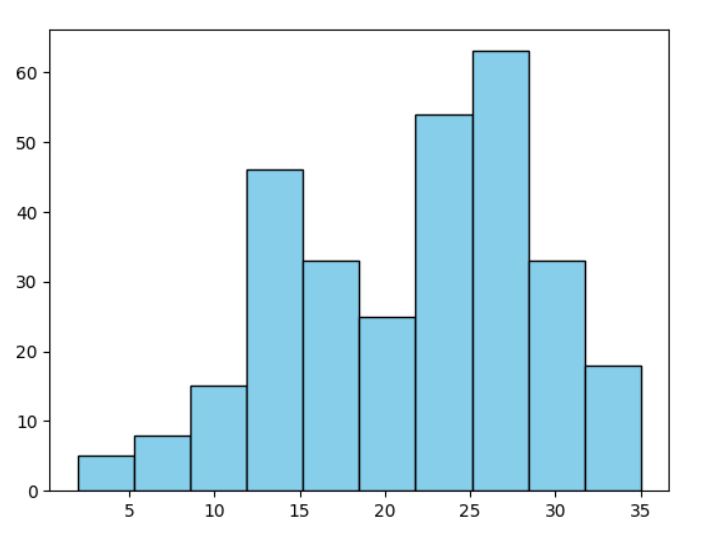


Fig 4.2.3.2: Histogram showcasing distribution of values in Schimers1righteye.

**Observations:**

* Prominent peaks are visible around the values of 25 and 30 on the x-axis.
* Valleys or lower points are noticeable around 10, 20, and 35 on the x-axis.

1. **“Schimers2Lefteye”**

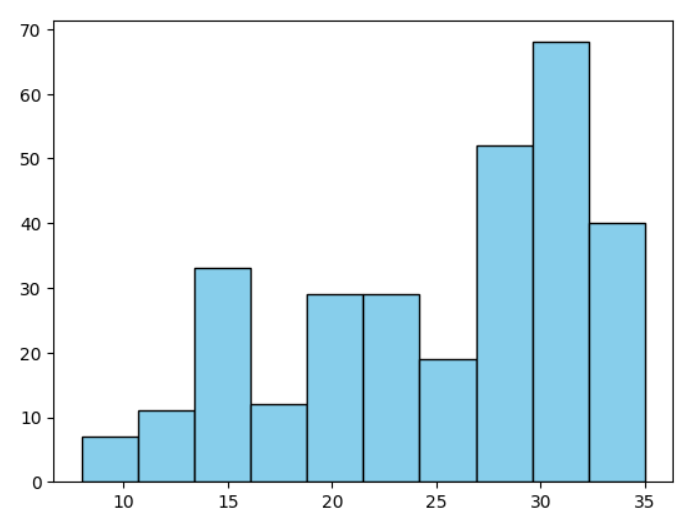
****

Fig : Histogram showcasing distribution of values in Schimers2Lefteye.

**Observations:**

* The bars initially increase in height, peak around the value of 30, and then decrease.
* The distribution appears **skewed to the right**, as the tail extends further in that direction.

1. **“Schimers2righteye”**

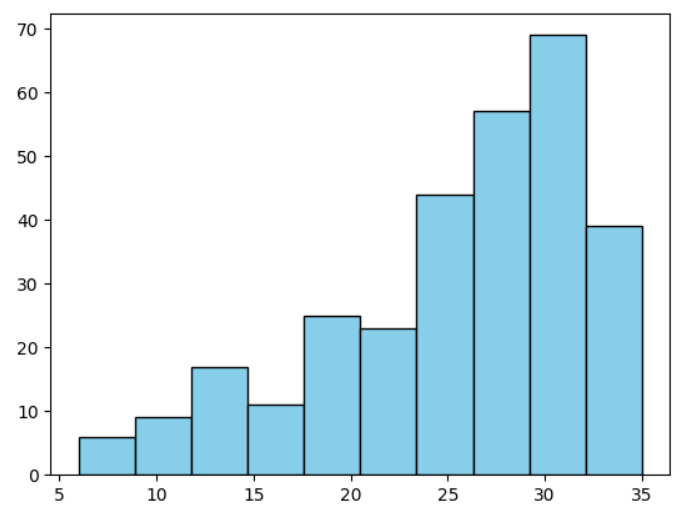
****

Fig : 4.2.3.1

**Observations:**

* The bars initially increase in height, peak around the value of 30, and then decrease.
* The distribution appears **skewed to the right**, as the tail extends further in that direction.
* **Frequency distribution of Age in the dataset**

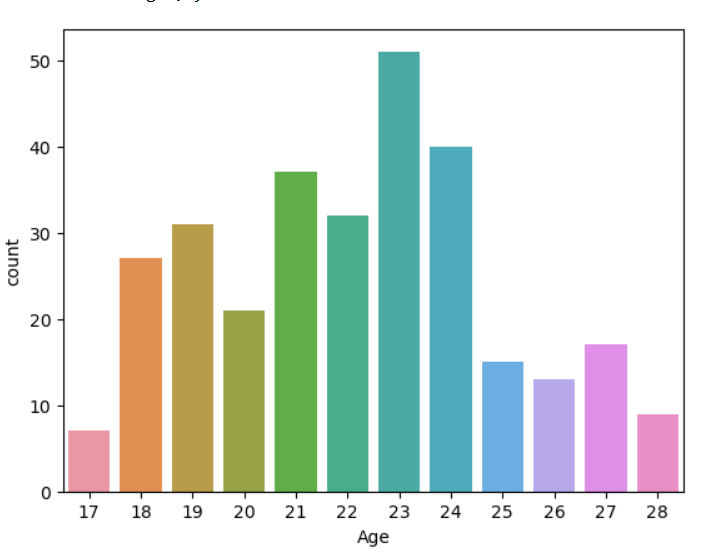


Fig : 4.2.3.2 Graph showcasing frequency distribution of Age

**Observation:**

* The graph displays a count of individuals across different age groups, ranging from 17 to 28 years old.
* The age group of 23 has the highest count, indicating that this age group is the most represented in the dataset.
* Age groups 21 and 24 follow closely behind, with slightly lower counts than the 23-year-olds.
* Age groups of 17, 26, and 28 have significantly lower counts compared to the other ages, suggesting fewer individuals fall into these age brackets within the dataset.
* **Analysis on wearables**

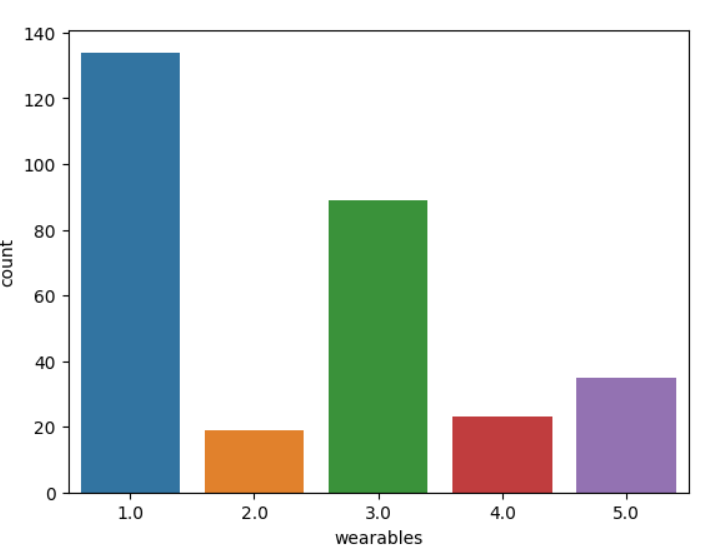
****

Fig : 4.3.2.3

**Observations:**

* **Type 1.0 Wearables(Glasses only)** : This category has the highest count, nearly 140, indicating that it’s the most common type of wearable in the dataset. It suggests that Type 1.0 might be a standard or basic model that is widely used.
* **Type 3.0 (None): With a count close to 80, this type is the second most prevalent in the dataset. This implies that there are close to 80 participants who do not wear any glasses or contacts.**
* **Other Types**: Types 2.0(Contacts only), 4.0(Both,at one or other time), and 5.0(Not applicable) have significantly lower counts, ranging from around 20 to 40. These types might represent more niche or less commonly selected options.
* **Overall Distribution**: The distribution suggests that the dataset is dominated by Type 1.0 wearables, with Type 3.0 next.
* **Onlineplatforms**

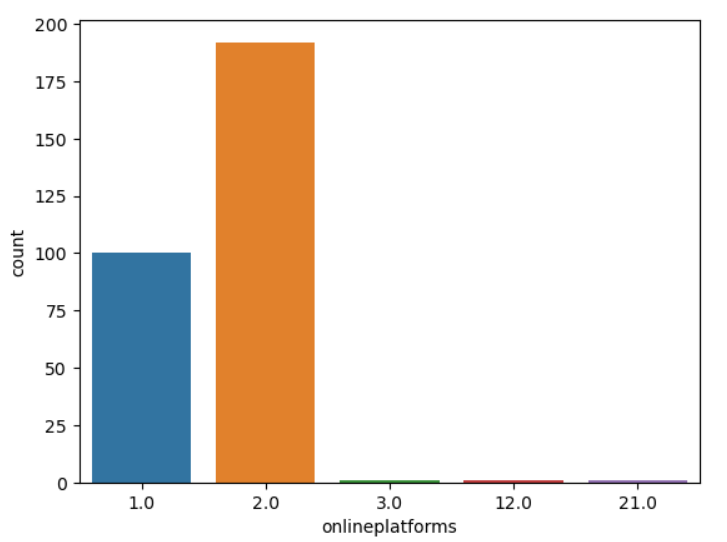
****

Fig : 4.3.2.4

This column indicates whether the participants use online platforms, represented by values 1 or 2. However, the options 3, 12, and 21 are outliers and need to be removed from the dataset for consistency. Although the count of these values is significantly lower than that of 1 and 2, removing them ensures data consistency.

* **Nature**

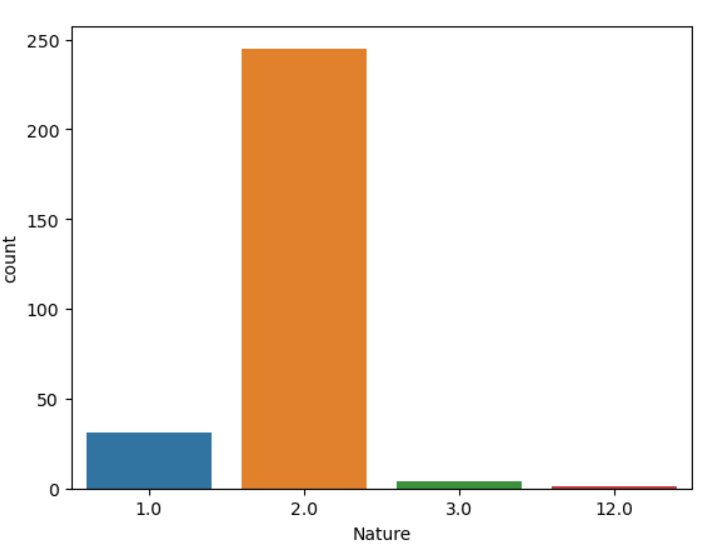
****

Fig : 4.3.2.5

The frequency distribution reveals that the majority of respondents indicate interrupted usage of their computers, as indicated by a higher count in the "2.0 - Interrupted usage" category compared to "1.0 - Continuous usage." This trend is noteworthy among the surveyed users. Additionally, the presence of values other than 1 and 2 in the graph suggests potential errors, considering that the survey form only offers two options. Given the low frequency of these outliers, it's advisable to remove them from the dataset.

* **ScreenIllumination**

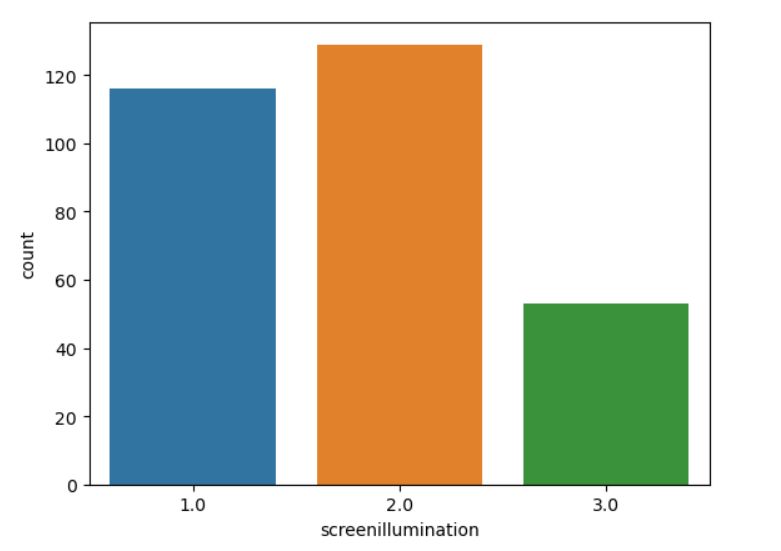
****

Fig :4.3.2.6 Graph representing the value count of screenillumination column.

* Screen Illumination 1.0 (Less than 25%): The count for this level is just under 120. This suggests that at this level of screen illumination, the count of the variable is relatively high.
* Screen Illumination 2.0 (25 to 30%): The count exceeds 120 at this level, indicating that this level of screen illumination results in the highest count of the variable among the three levels.
* Screen Illumination 3.0 (greater than 50%): The count is around 40 at this level, which is significantly lower than the other two levels. This implies that this level of screen illumination results in the lowest count of the variable.
* **WorkingYears**

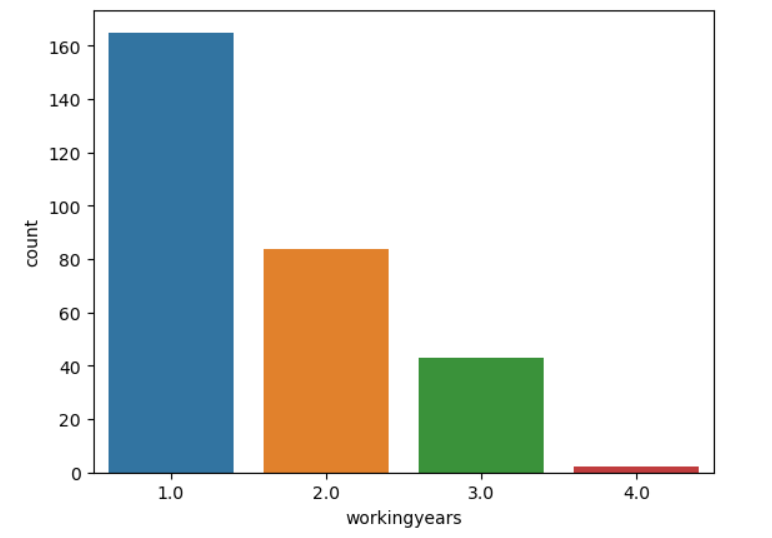
****

Fig :4.3.2.7 Frequency of workingyears

1. – Less than 5 years, 2.0 – 5 to 10 years, 3.0 – Greater than 10 years and 4.0 – Outlier

**Observations**

* Many participants have been recorded to use the gadgets for less than 5 years which may suggest they might be students in college or freshly graduated students.
* The usage of gadget is gradually decreased as the number of working years are increased. The category 2.0 and 3.0 support this observation.
* **Hoursspentdailycurricular**

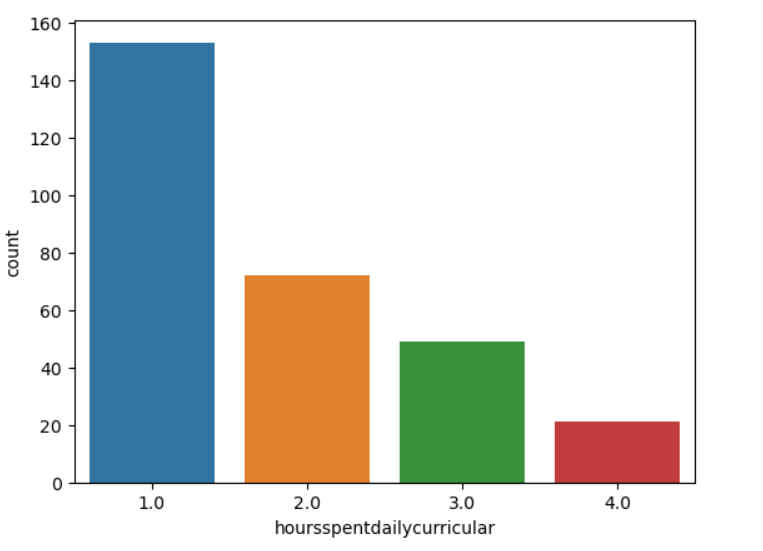


Fig.4.3.2.8

1. - 2 to 4 hours, 2.0 - 4 to 6 hours, 3.0 - 6 to 8 hours and 4.0 – 8 to 10 hours

**Observations**

Hours Spent Daily on Curricular Activities (1.0): The frequency for this level is about 150. This suggests that a significant number of people spend around 1 hour daily on curricular activities.

Hours Spent Daily on Curricular Activities (2.0): The frequency for this level is significantly lower than for 1.0, indicating that fewer people spend around 2 hours daily on curricular activities.

Hours Spent Daily on Curricular Activities (3.0 and 4.0): The frequencies for these levels are much lower, indicating that very few people spend 3 or 4 hours daily on curricular activities.

* **HoursSpentDaily on Non-Curricular Activities**

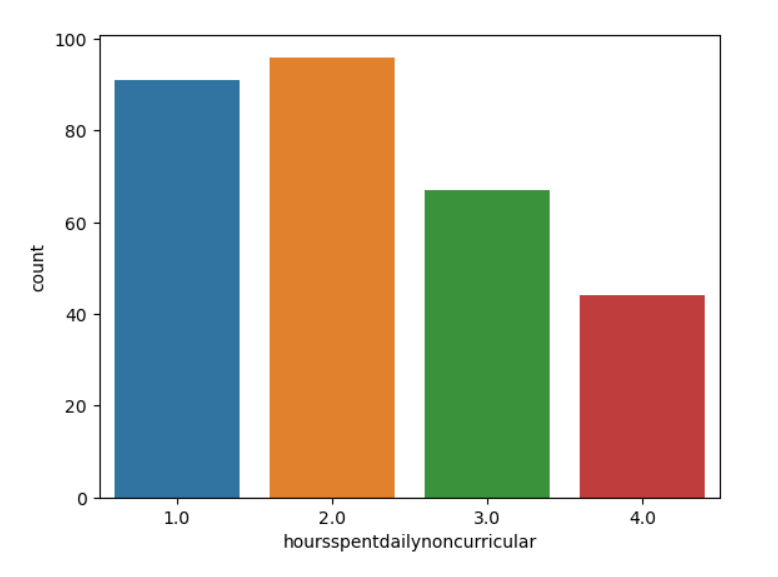


Fig. 4.3.2.9

1.0 – Less than 2hours, 2.0 – 2 to 4 hours, 3.0 – 4 to 6 hours and 4.0 – 6 to 8hours

**Observations**

* The most frequent range of hours spent daily on non-curricular activities is 2 to 4 hours, with a frequency of around 80.
* Less frequent, but still common, are the categories of less than 2 hours and 4 to 6 hours, with frequencies of around 60 and 40 respectively.
* The least frequent category is 6 to 8 hours, with a frequency of around 20.

Overall, the data suggests that most people spend less than 6 hours per day on non-curricular activities. There seems to be a gradual decrease in frequency as the number of hours spent per day increases.

* **Most Commonly used gadget**

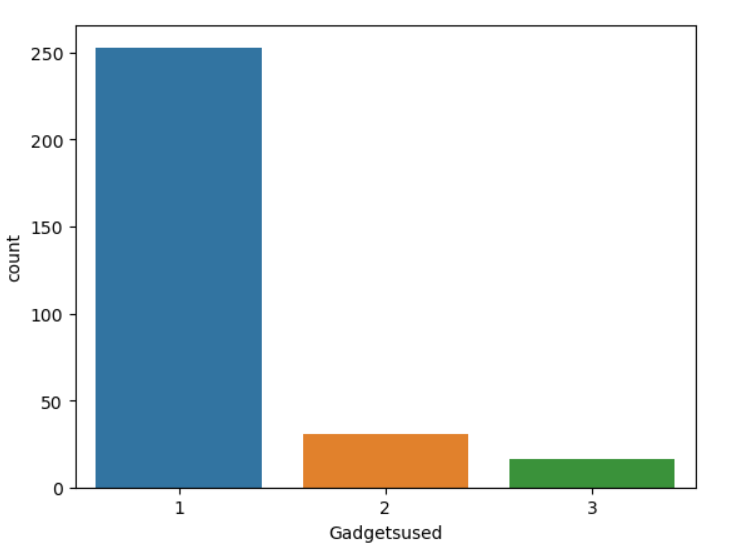


Fig : 4.3.2.10 Graph representing the frequency of gadgetsused

**Observations**

* Category 1 represents mobile gadgets, which are the most commonly used gadgets among the survey participants compared to category 2(Laptops) and category 3(tablets).
* **Frequency of experiencing dry eyes**

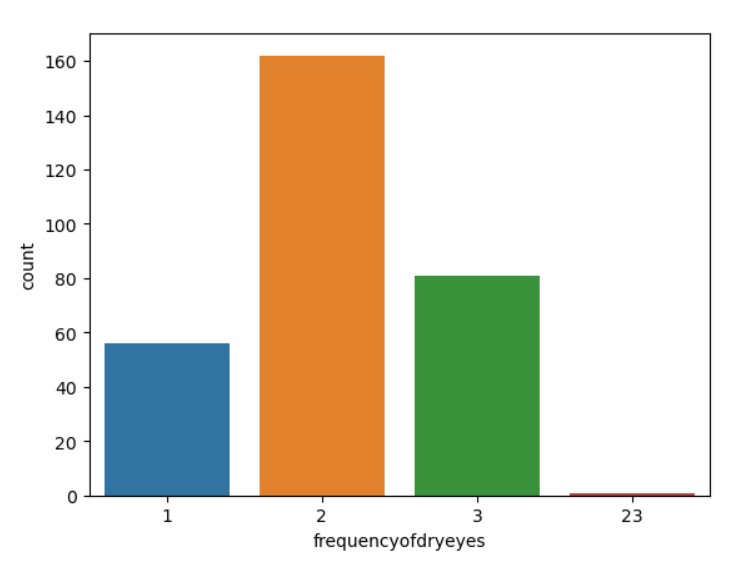


Fig :4.3.2.11 Frequency of dry eyes.

1 – Often , 2 – Sometimes, 3 – Never and 23- is an outlier in the column which can be removed since the frequency is near to 0.

**Observations**

* Often (1.0): The frequency for this level is about 40. This suggests that a smaller group of people often experience this condition or behaviour.
* Sometimes (2.0): The frequency for this level is the highest, about 150, indicating that the majority of people sometimes experience this condition or behaviour.
* Never (3.0): The frequency for this level is around 80, indicating that a significant number of people never experience this condition or behaviour.
* The category named 23 is an outlier and has to be removed while preprocessing data.

**4.3 Data Pre-Processing**

**4.3.1 Replacing Null values -**

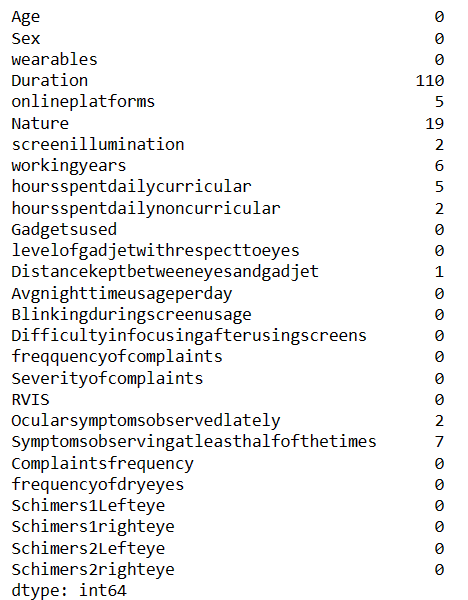
****

Fig :4.3.1.1 Sum of Null values.

The above Fig shows the total count of missing values in each column. Given that all columns are categorical, replacing missing values with the mode of each column appears logical. However, upon cross-referencing with the provided survey form, it's evident that the 'Duration' column is contingent on the preceding 'Wearables' column. Individuals who select 'None' or 'Not Applicable' for wearables may not need to specify a duration. Therefore, we segregated respondents based on their wearables selection. For those who wear wearables, missing values in 'Duration' were replaced with the mode, while missing values for non-wearables were replaced with zero.

Missing values in other columns were replaced with their respective mode values. This approach ensures data integrity while considering the survey context.

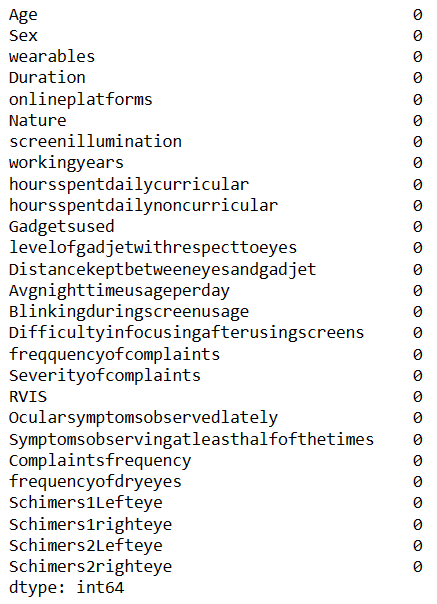


Fig :4.3.1.2 After Removing Missing values from the dataset.

* + 1. **Removing Rows with Incorrect Data**

During exploratory data analysis (EDA), we identified several columns with incorrect values. Subsequently, we removed the corresponding rows from the dataset due to their low frequency. The provided information indicates the total number of rows in the dataset after the removal of these erroneous entries.



Fig :4.3.2.1 Shape of the Dataset.

The figure illustrates that the dataset contains 281 rows and 27 columns.

**4.3.3 Making the Target Variable Categorical**

In order to enhance the comprehensibility of the output for a wider audience, the results from the schirmer tests have been segmented into three distinct categories. These categories are delineated based on the measured values in millimeters ('mm'). Specifically, the categories are defined as follows:

1. Dry: This category encompasses results falling within the range of 1 to 4 mm on the schirmer test. Individuals falling into this category may exhibit symptoms indicative of dryness in their eyes.

2. Slight Dryness: Results falling between 5 to 14 mm on the schirmer test fall into this category. It indicates a mild degree of dryness, where individuals may experience some discomfort or irritation but to a lesser extent compared to the 'Dry' category.

3. Normal: This category encompasses results exceeding 15 mm on the schirmer test. Individuals within this range typically demonstrate normal tear production and are less likely to experience symptoms of dryness or discomfort.

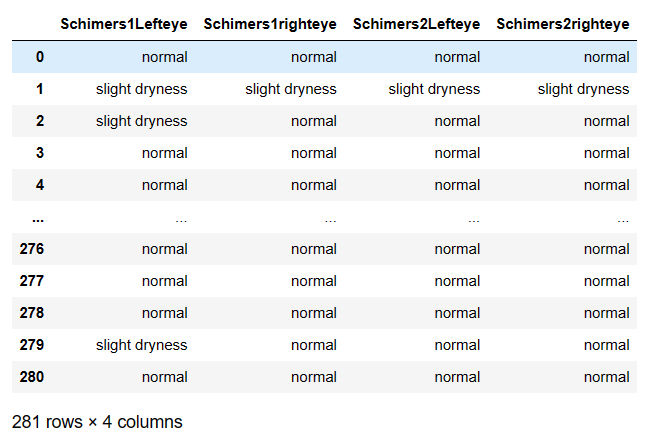


Fig :4.3.3.1 After categorising the target columns.

Furthermore, the data is encoded numerically as follows: 1 represents 'Dry,' 2 represents 'Slight Dryness,' and 3 represents 'Normal.'

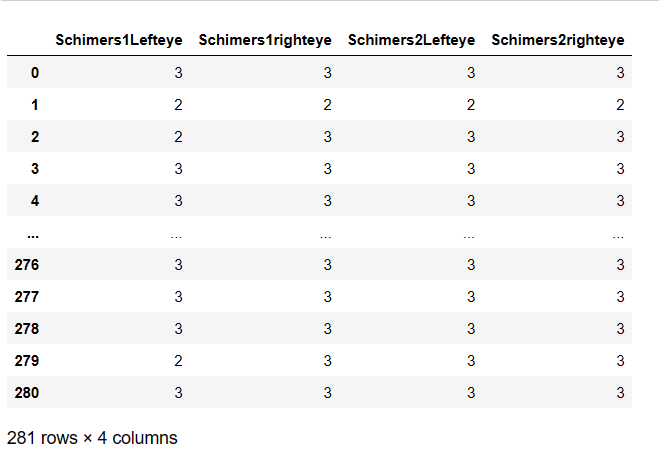


Fig :4.3.3.2 After Encoding the categories.

**4.3.4 Dividing Multiple Target Variables**

Recognizing the complexities involved in predicting multiple target variables simultaneously, our approach involves developing four distinct models, each tailored to predict one of the four target variables. These models are designed to utilize the same set of input variables, ensuring consistency and comparability across predictions. This strategy allows for a focused and specialized approach to address the unique characteristics and dependencies present in each target variable, ultimately enhancing the predictive accuracy and interoperability of our analyses.

**4.3.5 Creating Training and Testing Datasets**

All columns, except for "Name" and the target variables, are utilized as input for building the model and predicting the outcome.

The dataset has been partitioned into training and testing datasets with a split ratio of 70% for training data and 30% for testing data.

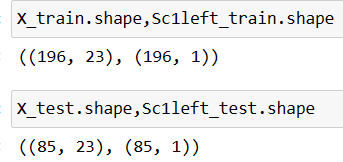
****

Fig :4.3.5.1 Train and split data for Schimers1lefteye

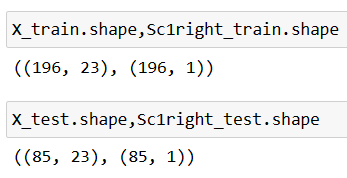


Fig :4.3.5.2 Train and split data for Schimers1righteye

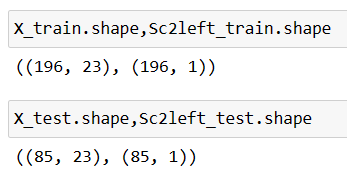


Fig :4.3.5.3 Train and split data for Schimers2lefteye

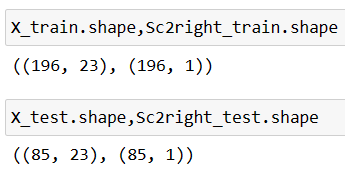


Fig :4.3.5.4 Train and split data for Schimers2righteye

The dataset has been separated into two parts, as explained earlier. Now we're all set to create our models using this data.

1. **ALGORITHMS**

The following algorithms are well suited for building a classification models.

**1. Logistic Regression:**

- Logistic Regression is a linear model used for binary and multiclass classification tasks.

- It estimates probabilities using the logistic function and assigns the class with the highest probability.

- Despite its name, it's used for classification rather than regression.

- It's computationally efficient and interpretable but may struggle with complex relationships in the data.

**2. Gradient Boosting Classifier:**

- Gradient Boosting Classifier is an ensemble learning technique that combines multiple weak learners (usually decision trees) sequentially to create a strong learner.

- It's highly effective for a wide range of classification problems, including multiclass classification.

- It builds trees iteratively, with each tree correcting the errors of the previous ones.

- It's known for its high predictive accuracy but may require more computational resources and tuning of hyperparameters.

**3. SVC (Support Vector Classifier):**

- SVC is a powerful supervised learning algorithm used for classification tasks, including multiclass classification.

- It works by finding the hyperplane that best separates the classes in the feature space.

- SVC can handle high-dimensional data effectively and is robust to overfitting.

- It may not perform well on very large datasets due to its computational complexity.

**4. Random Forest Classifier:**

- Random Forest Classifier is an ensemble learning method that builds multiple decision trees during training and outputs the class that is the mode of the classes (classification) of the individual trees.

- It's highly accurate and robust to overfitting, even for high-dimensional data.

- It's capable of handling large datasets with high dimensionality and is relatively easy to use.

**5. Decision Tree Classifier:**

- Decision Tree Classifier is a simple yet powerful supervised learning algorithm used for classification tasks.

- It splits the data into subsets based on the feature value that results in the best separation of the target classes.

- It's intuitive and easy to interpret, making it suitable for understanding the underlying structure of the data.

- However, decision trees can be prone to overfitting, especially with complex datasets.

1. **Results(Accuracy) obtained for the models**

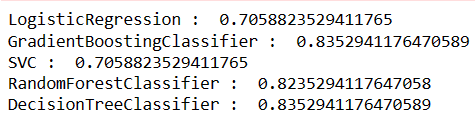
****

Fig : Accuracies of Schimers1lefteye models

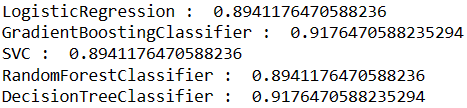


Fig : Accuracies for Schimers1righteye models

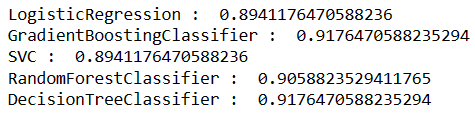


Fig : Accuracies for Schimers2lefteye models

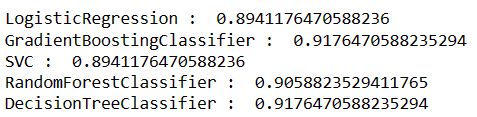


Fig : Accuracies for Schimers2righteye models

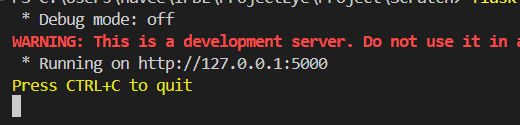
**Observations**

The above figure’s indicates that the accuracies of the Gradient Boosting Classifier and Decision Tree Classifier are identical. Even after splitting the data multiple times and building various models, the results remained nearly the same.

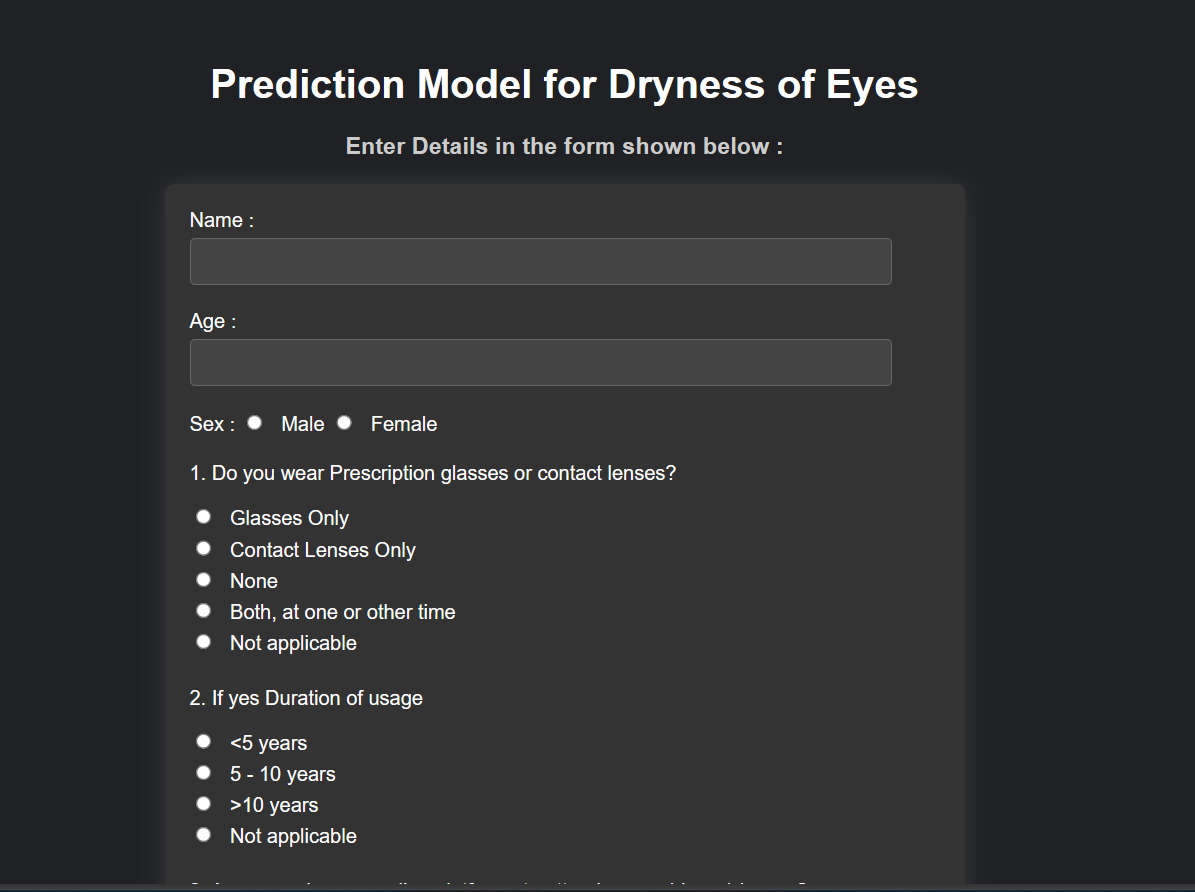
1. **Final Implementation**

For the final implementation, a web application is developed using Flask. The models, which were built using Jupyter, are transformed and saved into pickle files. These pickle files are then loaded onto the API. To use the web application and obtain results, the following steps need to be followed.

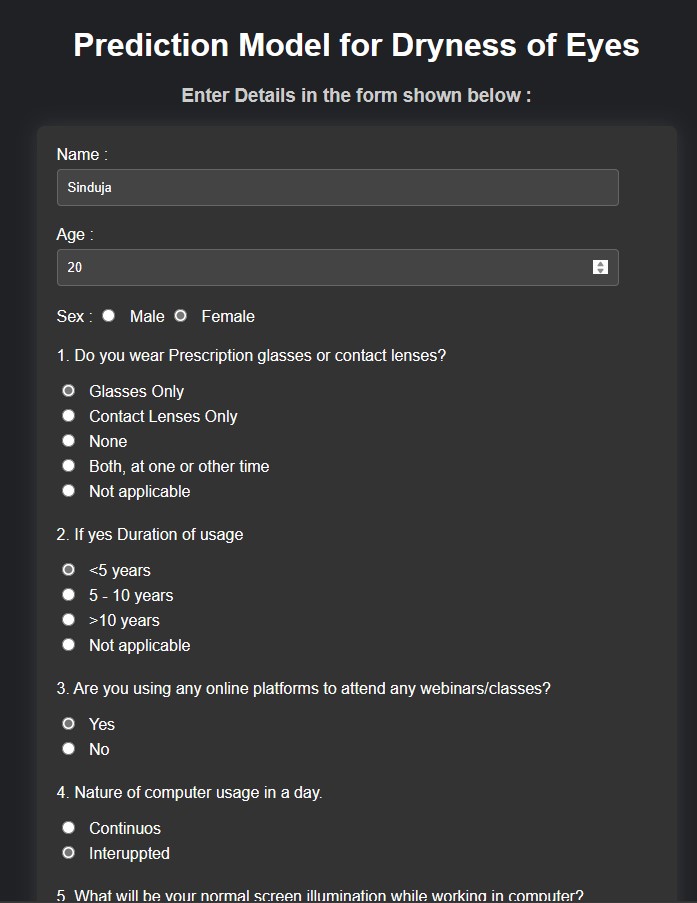
1. Launch the web app by executing the Python code, which will deploy the application on localhost.



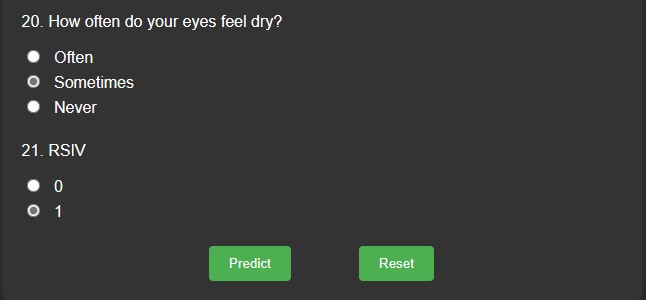
1. Open the deployed web app in a browser. If the deployment is successful the following webpage is rendered.

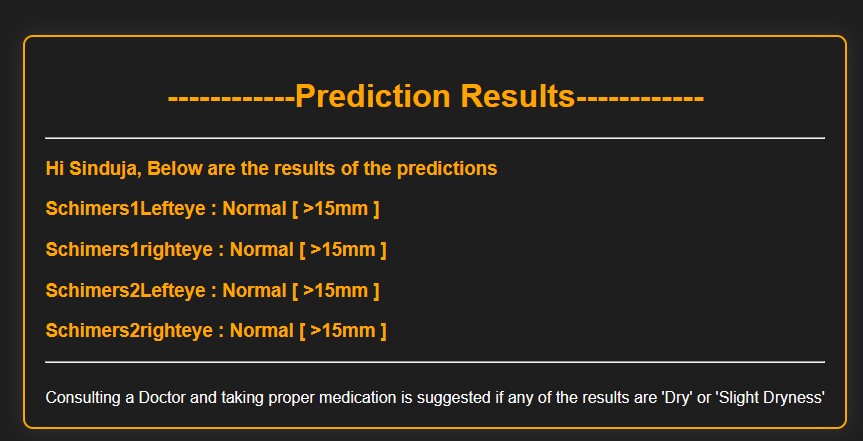


1. Fill the form which appeared on the screen. The form must be filled carefully as the options represent the input for the predictive model.



1. After filling all the fields the use have to click on predict button to acquire the results.





**Conclusion**

In our study, we employed several popular machine learning algorithms, including Logistic Regression, Gradient Boosting Classifier, SVC (Support Vector Classifier), Random Forest Classifier, and Decision Tree Classifier. Upon evaluation, we found that both the Gradient Boosting Classifier and Decision Tree Classifier exhibited comparable accuracy levels. However, we opted to utilize the Gradient Boosting Classifier for our final model implementation due to its reputation as an advanced algorithm.

Although the Decision Tree Classifier demonstrated similar performance, the Gradient Boosting Classifier offers several advantages, such as better handling of complex relationships within the data, robustness against overfitting, and the ability to sequentially improve model performance through the ensemble of weak learners.

Consequently, the final implementation of our model, which yielded the digital eyes results, was derived from the Gradient Boosting Classifier. This choice was driven by the algorithm's ability to provide reliable and accurate predictions, making it well-suited for our specific task.