Spectacles Prediction using Machine Learning Algorithms

A Project Report in partial fulfillment of the degree

Bachelor of Technology

in

Electronics & Communication Engineering/Computer Science & Engineering

By

19K41A04B6	S. Amulya
19K41A0570	K. Sudeeptha
19K41A0573	G. Sathvika

Under the Guidance of
Dr. V. Venkataramana

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
S.R.ENGINEERING COLLEGE(A), ANANTHASAGAR, WARANGAL
(Affiliated to JNTUH, Accredited by NBA) Dec-2021.

SR Engineering College
Innovation . Creativity . Entrepreneurship

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled "Spectacles prediction using Machine Learning Algorithms" is a record of bonafide work carried out by the student(s) S. Amulya, K. Sudeeptha, G. Sathvika bearing Roll No(s) 19K41A04B6, 19K41A0570, 19K41A0573 during the academic year 2021-2021 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Electronics & Communication/Computer Science Engineering** by the Jawaharlal Nehru Technological University, Hyderabad.

.

Supervisor **Head of the Department**

External Examiner

ABSTRACT

Eye defects in youth are becoming more prevalent these days, especially due to the pandemic situation. Digital eye strain due to excessive usage of gadgets can result in a prescription for spectacles, which is a basic test in any ophthalmology clinic, but due to the current situation of the pandemic, visiting a clinic might be risky. In this project, we propose an AI (ML) model for spectacles prediction based on a few input parameters through an app. AI machine learning models are used widely in the medical sector. Here, we applied 3 different machine learning models (Logistic regression model, Decision tree algorithm and Random forest algorithm) to extract maximum accuracy prediction using the dataset we collected from Youngsters. In our project, the highest accuracy of 96% is achieved using Logistic regression model, which we used for developing a web application.

Table of Contents

1. INTRODUCTION

Usage of digital devices especially smartphones significantly increased in the previous decade. Moreover, COVID pandemic has further shifted much of the work towards digital device assisted applications. In today's era, people across all ages are spending a lot of time in front of these devices. This also implies a surge in Digital Eye Strain cases, which is one of the emerging health issues. Researchers have linked this problem with symptoms such as dry eyes, altered blinking pattern, visual fatigue etc. [1]. Digital eye strain is a phenomenon that comes about from over use of digital screens, placing a strain on the eyes and leading to future vision problems. One study has found that children who use technology more often are more likely to need glasses for vision correction than their peers [4].

In the modern era, the use of digital screens is quite common for children and teenagers. Besides, the instigation of unlimited e-classes for such children has rested overt burden on already overburdened eyes. And this way unknowingly we are pushing a cohort of children into a higher risk of digital eye strain due to current trend of unregulated e-learning [2]. Teenagers spend more time than ever staring at digital screens—on computers, tablets, TVs, smartphones, and other devices. All that screen time can take a toll on their wellbeing, including how their eyes may feel. Research shows that children begin zooming in on digital media devices, such as their parents' tablets or smartphones, as young as 6 months old. By their teens, studies have found, kids spend nearly 9 hours a day using screened-based media, watching TV, playing video games, and using social media. Especially if they're having fun, they might keep playing and watching to the point of eye-rubbing exhaustion [5].

In a recent survey conducted in 2021, 72% of teenagers said that they were on social media "almost constantly." That's up from 27% in 2018 [6]. In general, social media use increases inversely with age, so younger generations are using social media more and more often. Today's social media platforms, including YouTube, Facebook, Snapchat, and Instagram, all require a computer, phone, or tablet. This means that young people who are on these platforms often are constantly being exposed to screens. In addition, features of these platforms, such as videos and endless scrolling, make it difficult for users to look away and get the breaks in digital activity that are critical to avoiding digital eye strain.

We contacted an ophthalmologist to gather information regarding the symptoms of digital eye strain. Our findings are as follows:

- *Eye fatigue:* Muscles around the eye, like any others, can get tired from continued use. Concentrating on a screen for extended periods can cause concentration difficulties and headaches.
- *Blurry vision:* Gazing at the same distance for an extended time can cause the eye's focusing system to spasm or temporarily "lock up." This condition, called an *accommodation spasm*, causes a child's vision to blur when he or she looks away from the screen.
- *Dry eyes:* Studies show that people blink significantly less often when concentrating on a digital screen, which can leave eyes dry and irritated.
- Frequent headaches

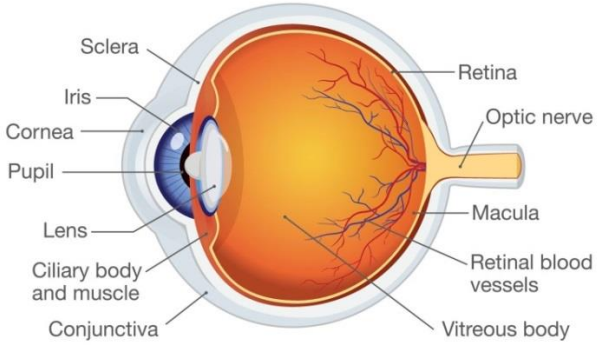


Figure 1: Human eye anatomy [7]

According to the ophthalmologist, treatment for eyestrain consists of making changes in our daily habits or environment. Some people may need treatment for an underlying eye condition. For some people, wearing glasses that are prescribed for specific activities, such as for computer use or for reading, helps reduce eyestrain. Doctor also suggests taking regular eye breaks to help the eyes focus at different distances

2. LITERATURE REVIEW

The paper published by R. Kaur and A. Guleria, titled “Digital Eye Strain Detection System based on SVM” in 2021, tries to effectively detect when a user is under strain so that he or she can take timely precautions. In this paper, a few more relevant features have been added, like facial gestures and decrease in glabella length along with the features from previous theoretical studies. A supervised method of statistical features linked to suggested symptoms is

proposed for classifying videos recorded in real time when user is under strain using SVM (Support Vector Machine, is a supervised machine learning algorithm that can be used for both classification and regression challenges). The main finding of this paper is an explicit feature set comprising of two newly proposed features along with four other appropriate features derived from previous theoretical studies. The proposed system shows considerable increase in accuracy when tested on YawDD (Yawning Detection Dataset), the best possible dataset available for the use case. [1]

3. DESIGN:

3.1 Requirement Specifications (S/W & H/W)

Hardware Requirements

- ✓ **System** : Processor Intel(R) Core (TM) i5-8265U CPU @ 1.60GHz, 1800 MHz, 4 Cores, 8 Logical Processors
- ✓ **RAM** : 8 GB
- ✓ **Hard Disk** : 557 GB
- ✓ **Input** : Keyboard and Mouse
- ✓ **Output** : PC

Software Requirements

- ✓ **OS** : Windows 10
- ✓ **Platform** : Google Colaboratory / Jupyter Notebook
- ✓ **Deployment software** : Stream lit
- ✓ **Program Language** : Python

3.2 Flow chart

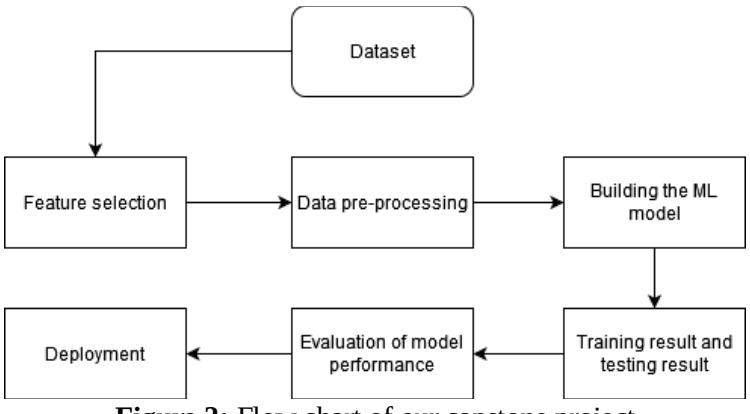


Figure 2: Flow chart of our capstone project

In the above flow chart we described our work flow in this project for developing ML models.

4. DATASET:

We collected data 250 samples from youngsters of age 17 to 22 years through a survey using Google forms. To finalize the input features, we contacted an ophthalmologist to know more about the digital eye strain symptoms. Upon consideration, we finalized 5 input features to predict whether or not a person has spectacles.

Input features:

- Number of hours spent on Mobile phone per day
  - Value : 0 - 24 (Numeric input)
- Number of hours spent on Laptop/Computer per day
  - Value : 0 - 24 (Numeric input)
- Number of hours spent on TV per day
  - Value : 0 - 24 (Numeric input)
- Frequent headaches
  - Value : Yes/No (Binary input)
- Blurry vision
  - Value : Yes/No (Binary input)

Output feature:

- Spectacles
  - Value : Yes/No (Binary output)

In total, we have 6 features, out of which 3 have numerical data and 3 have binary data, which can be visualized as below. For binary data visualization (Scale: On x-axis, 1=Binary 0, 2=Binary 1)

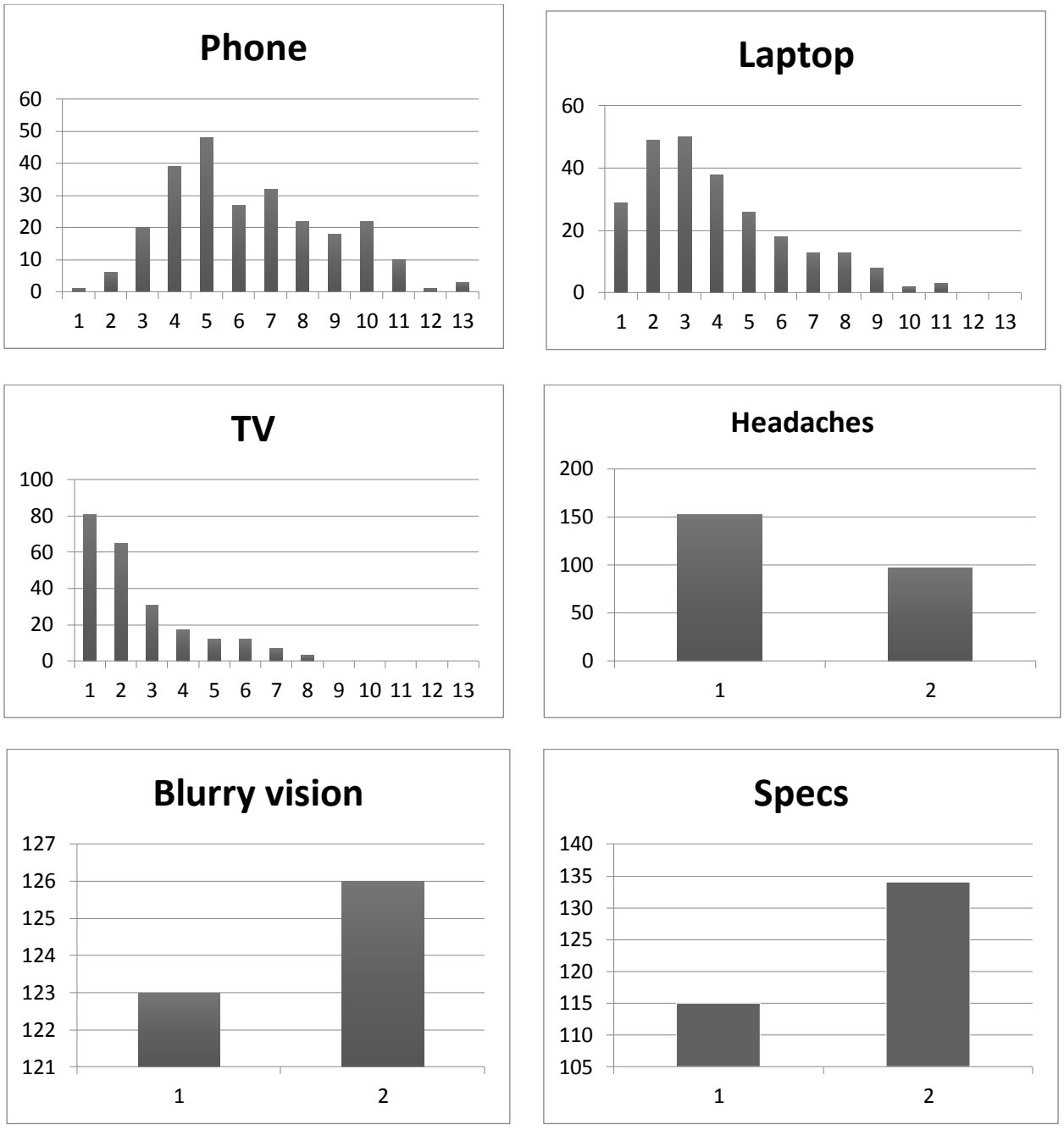


Figure 3: Visualizing attributes of the dataset

5. DATA PREPROCESSING:

Real-world data collection has its own set of problems. It is often very messy which includes missing data, presence of outliers, unstructured manner, etc. Before looking for any insights from the data, we have to first perform preprocessing tasks which then only allow us to use that data for further observation and train our machine learning model. We

use missing values treatment, outliers detection, normalization and data split to process our data before feeding it to the machine learning model.

Data info:

```
RangeIndex: 249 entries, 0 to 248
Data columns (total 6 columns):
#   Column         Non-Null Count  Dtype
---  ------         --------------  -----
0   phone          249 non-null    int64
1   Laptop         249 non-null    int64
2   TV             249 non-null    int64
3   Headaches      249 non-null    int64
4   Blurry vision  249 non-null    int64
5   Specs          249 non-null    int64
dtypes: int64(6)
```

Missing values treatment:

The real world's dataset often has many missing values which can be treated by using certain methods. But in our data, there are no missing values, because we collected the data manually through google forms survey and made sure not to miss out any data. To treat the missing values, we generally use the following strategies:

- Remove the entire row (If missing values are less in number)
- Replace the missing value with either mean or median
- Replace the missing value with most frequent value in the column (This is generally used only for large dataset)

```
print(data.isnull().sum())
```

```
phone            0
Laptop           0
TV               0
Headaches        0
Blurry vision    0
Specs            0
dtype: int64
```

Fig 4: Printing missing values from our dataset

Outliers detection and treatment:

Outliers are data points that don't fit the pattern of rest of the numbers. They are the

extremely high or extremely low values in the data set. A simple way to find an outlier is to examine the numbers in the data set. We can also detect outliers by Z-score method, its formula is given by**:**

- **Z=(x- μ)/ σ**
    - Where, x is each value in dataset,
    - μ is mean
    - σ is standard deviation

Normalization:

Normalization is a technique for organizing data in a database. Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0. It is important that a database is normalized to ensure only related data is stored in each table and to avoid biasing towards huge values. When we normalize the data while feeding it to the model, we also have to de-normalize it. This process can be done using the formulas below:

- $x_{nor} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$
- $y_i = y_{nor}(y_{max} - y_{min}) + y_{min}$

Data split:

To train any machine learning model irrespective what type of dataset is being used, we have to split the dataset into training and testing data. The reason to split the data is to give the machine learning model an effective mapping of input to outputs and to evaluate the model performance. We pass the training data to train our machine learning model and then test the model on testing data. In our model, we used 70:30 data split. That is the data is split 70% for training and 30% for testing. We can do the data split using train_test_split module in python.

Correlation matrix:

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

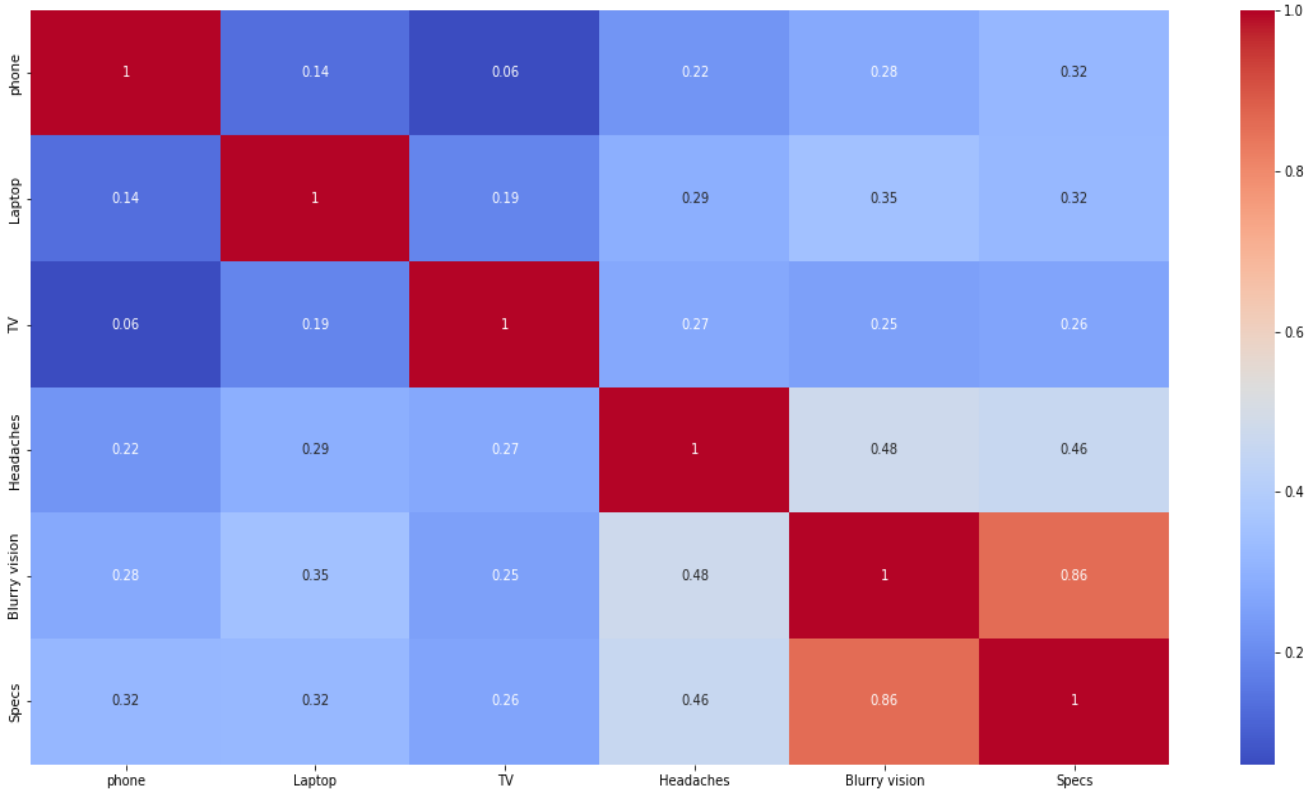


Fig 5: Correlation matrix of our dataset

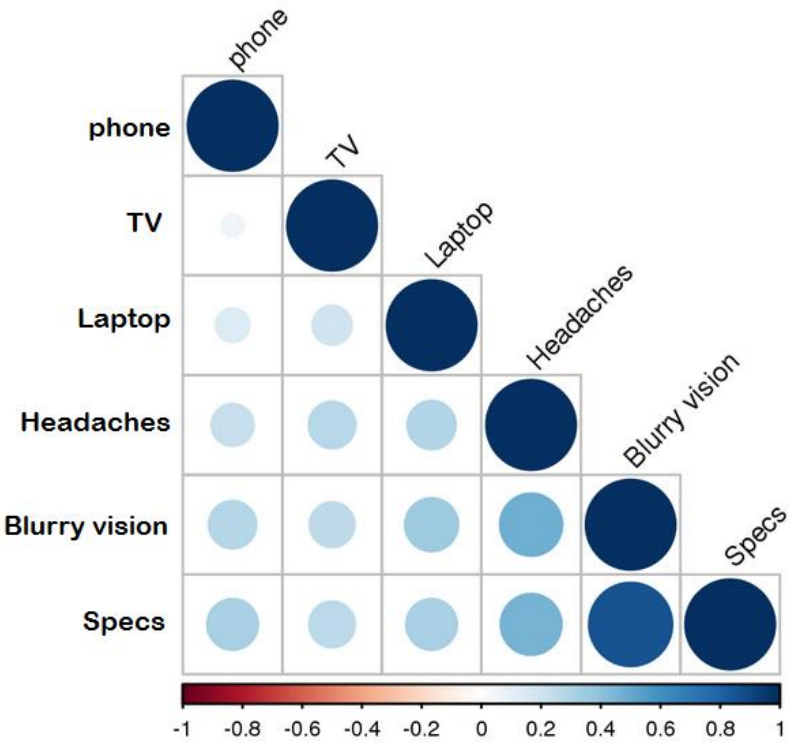


Fig 6: Correlogram of our dataset

6. METHODOLOGY:

This section talks about the algorithms used for the project. We used three different algorithms like Logistic Regression, Decision tree and Random forest.

Logistic Regression

Logistic regression is a supervised algorithm for study classification. The likelihood of a destination variable was predicted. The nature of the target or dependent variable is dichotomous, meaning that only two possible classes are available [7]. Here, we use binary logistic regression which uses the sigmoid function to convert real numbers into binary and to calculate accuracies we use binary cross entropy loss function.

Steps for Logistic Regression:

Step 1: Read the data

Step 2: Data pre-processing (Normalization, missing values treatment, outliers)

Step 3: Initialize the model parameters, number of iterations and eta

Step 4: Gradient calculation (For all iterations and data samples)

Step 5: Read the model parameters

Step 6: Print confusion matrix and accuracy for training and testing data

Step 7: Model performance evaluation and deployment

Decision tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label**.** Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. This process is then repeated for the sub tree rooted at the new node [8]. We can perform decision tree using entropy and gini index.

Steps for decision tree using entropy:

Step 1: Determine the Root of the Tree.

Step 2: Calculate Entropy for The Classes.

Step 3: Calculate Entropy After Split for Each Attribute.

Step 4: Calculate Information Gain for each split.

Step 5: Perform the Split.

Step 6: Perform Further Splits until leaf nodes

Step 7: Complete the Decision Tree.

The decision trees of our model using entropy and gini index are shown below:

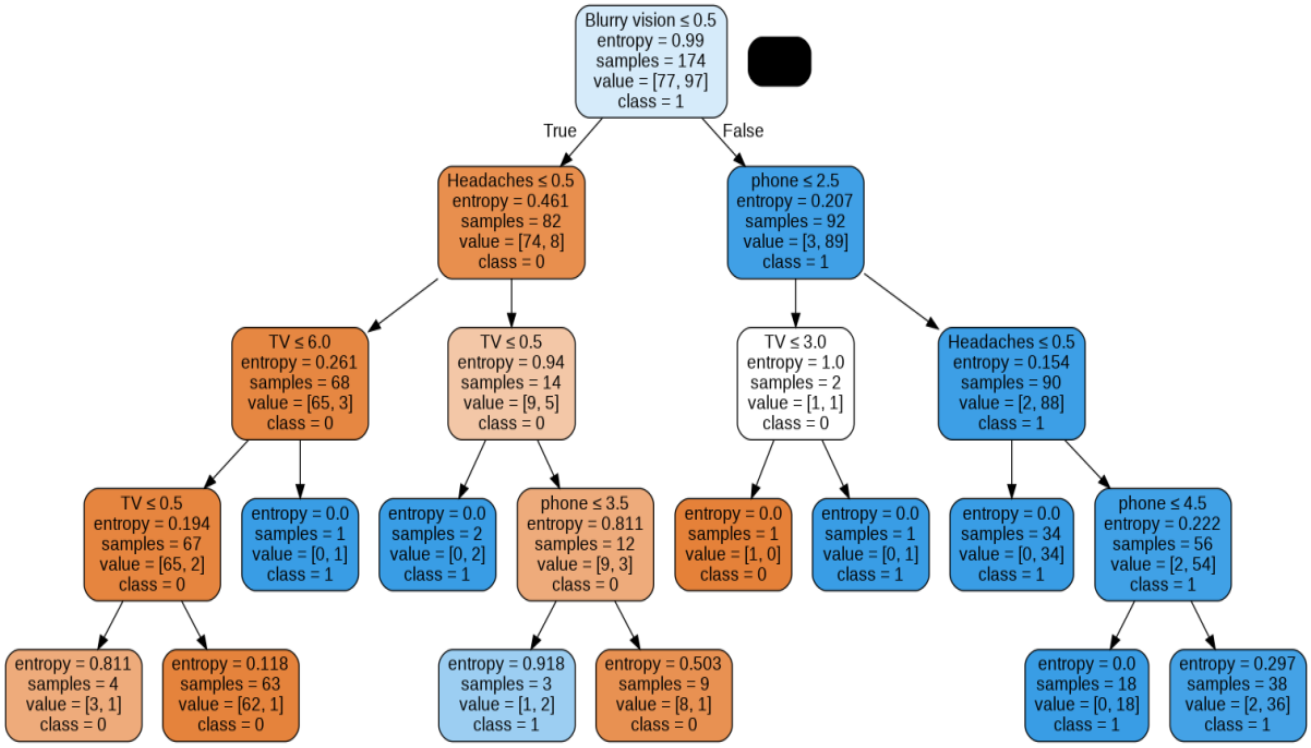


Fig 7: Decision tree using entropy

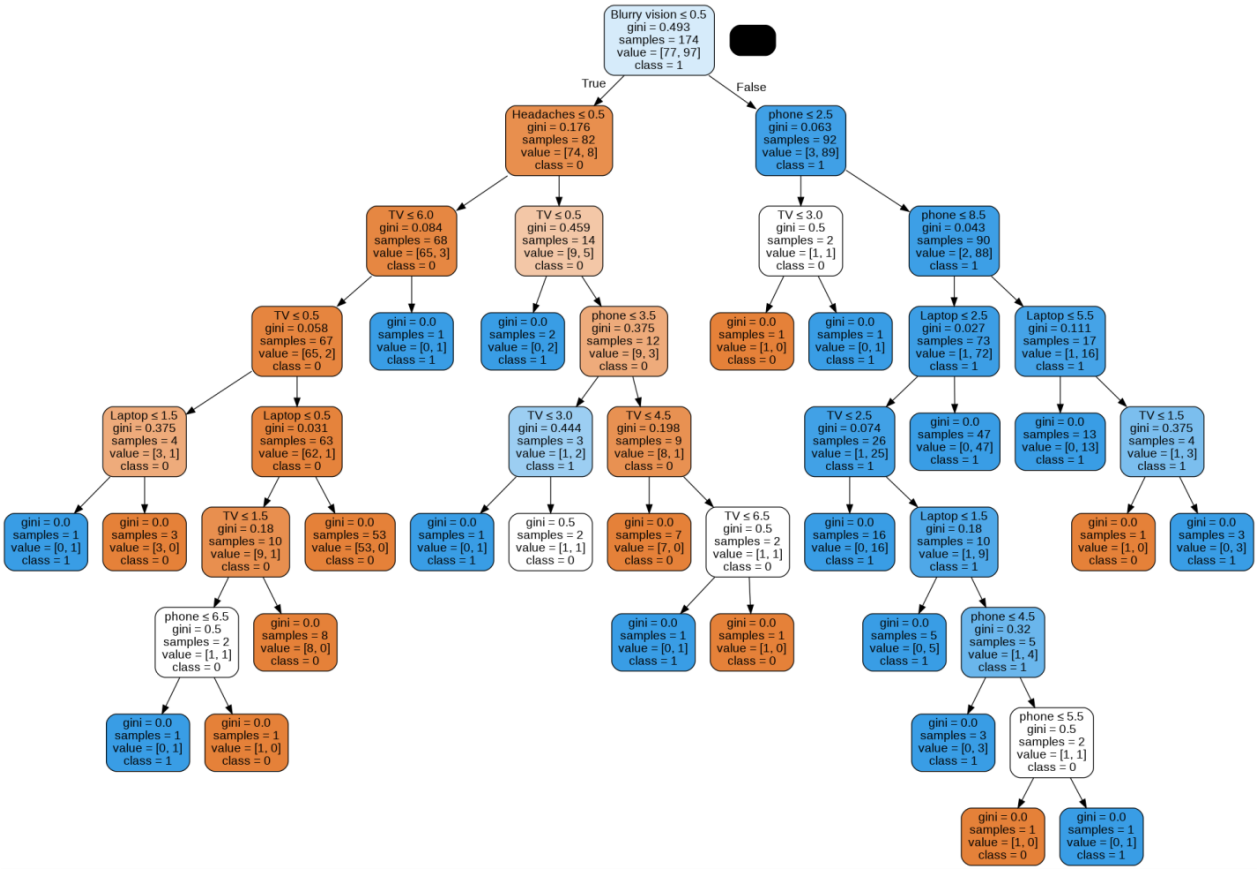


Fig 8: Decision tree using gini index

Random forest:

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. Random forest establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the over fitting of datasets and increases precision [9].

Steps of random forest:

Step 1: Selection of random samples from the dataset.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output

Step 4: Final output is considered on majority voting or averaging

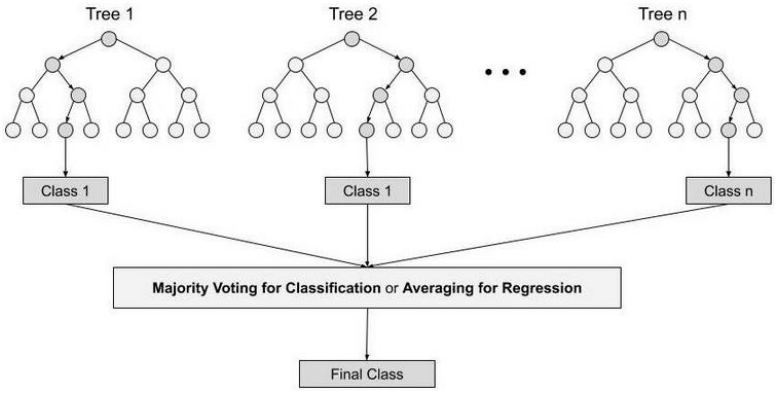


Fig 9: Random forest

7. RESULTS:

We used 3 different machine learning algorithms in this project, and all the algorithms have high level accuracies for our dataset. We consider testing accuracy for final deployment of the project. Here, the testing accuracy of Logistic regression is highest with 96%, the next highest is decision tree algorithm which gave 90% accuracy and lastly, random forest algorithm with accuracy 82%. As the logistic regression has highest accuracy of all the 3 models, we use that model to develop the web application for the project. Accuracies with training and testing data of all the three models are represented in graphical and tabular form below.

Training accuracy:

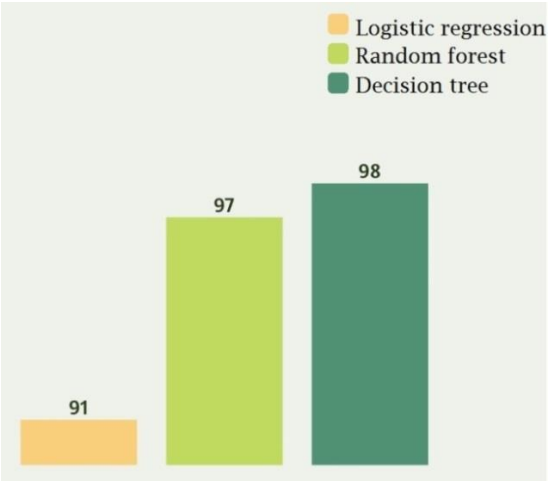


Fig 10: Training accuracy of all the models

Testing accuracies:

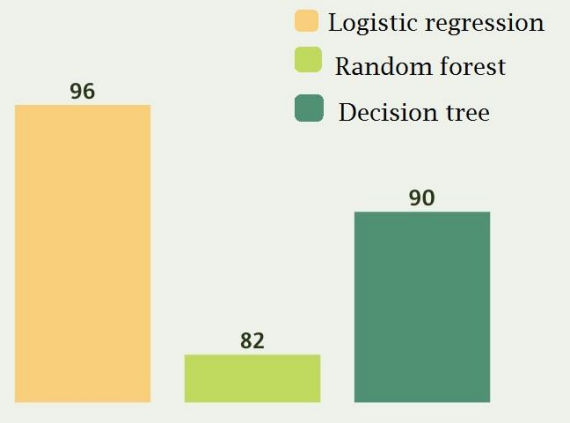


Figure 11: Testing accuracies of all the models

Algorithm	Training accuracy	Testing accuracy
Logistic Regression	91	96
Random forest	97	82
Decision tree	98	90

Table 1:. Accuracy

Cost	1	2	3	4	none
Entropy	93%	93%	95%	96%	99%
Gini	93.6%	93.6%	95%	96%	99%

Table 2: Decision tree training accuracies

Cost	1	2	3	4	None
Entropy	90%	89%	86%	86%	77%
Gini	90.6%	89%	88%	86%	77%

Table 3: Decision tree testing accuracies

Cost	1	2	3	4	None
Entropy	89%	93%	94%	95%	97%
Gini	92%	93%	94%	95%	95%

Table 4: Random forest training accuracies

Cost	1	2	3	4	None
Entropy	90%	90%	99%	87%	86%
Gini	90%	87%	86%	84%	82%

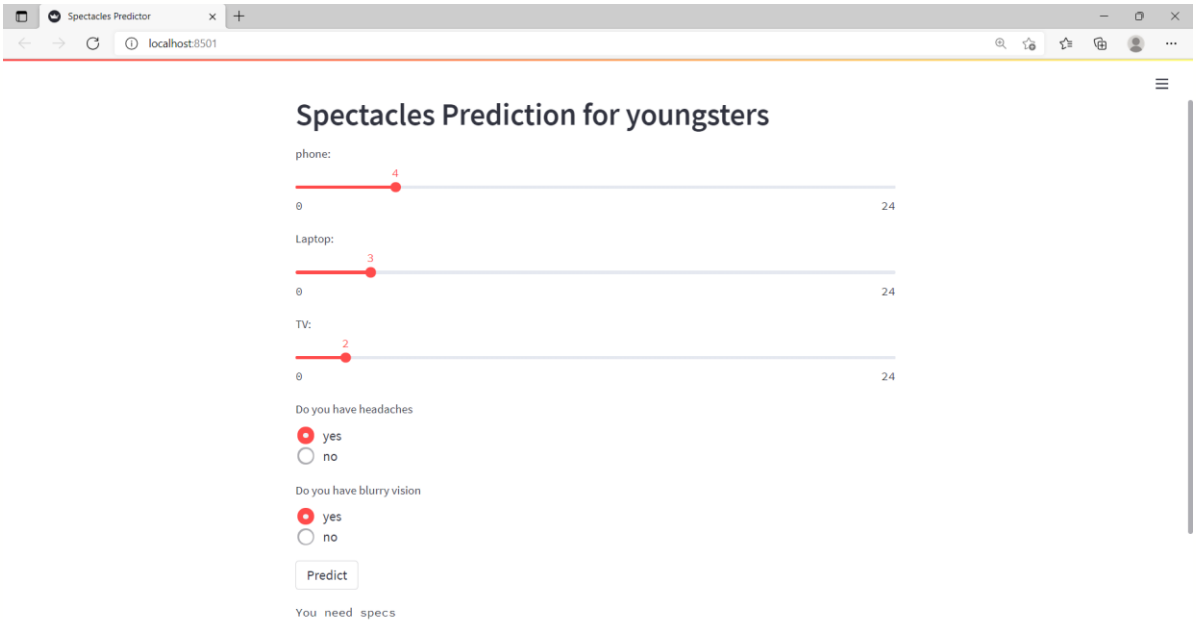
Table 5: Random forest testing accuracies

Confusion matrix

- Logistic regression: `[[21,  0],`
  `[ 2, 27]]`
- Decision tree: `[[30,  8],`
  `[ 7, 30]]`
- Random forest: `[[30,  8],`
  `[ 5, 32]]`

8. APPLICATION:

We have developed a web application through streamlit software by implementing the logistic regression model (as it has the highest testing accuracy)



9. CONCLUSION:

In conclusion, the spectacles prediction model is developed as a web application with an accuracy of 96% using logistic regression model. We can further improve the scope of this project by moving the web application from local host to cloud based environment.

10. REFERENCES:

[1] R. Kaur and A. Guleria, "Digital Eye Strain Detection System Based on SVM," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021, pp. 1114-1121, doi: 10.1109/ICOEI51242.2021.9453085

[2] Bhattacharya, Sudip & Saleem, Sheikh & Singh, Amarjeet. (2020). Digital eye strain in the era of COVID-19 pandemic: An emerging public health threat. Indian Journal of Ophthalmology. 68. 1709. 10.4103/ijo.IJO_1782_20.

[3] Rosenfield, Mark. (2016). Computer vision syndrome (a.k.a. digital eye strain). Optometry in practice. 17. 1-10.

[4] Eye strain risk for children and teens [internet] available at: https://blog.eyeglasses.com/visionmagazine/eye-strain/

[5] Give your children's eyes a screen-time break: here's why [internet] available at: https://www.healthychildren.org/English/health-issues/conditions/eyes/Pages/What-Too-Much-Screen-Time-Does-to-Your-Childs-Eyes.aspx

[6] How much time do people spend on social media in 2021[internet] available at: https://techjury.net/blog/time-spent-on-social-media/#gref

[7] Introduction to logistic regression [internet] available at: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[8] Decision tree [internet] available at: https://www.geeksforgeeks.org/decision-tree/

[9] Random forest [internet] available at: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/