**A logo for a university

AI-generated content may be incorrect.DATA SCIENCE  
PROJECT REPORT  
  
Supervised Learning for Early Diagnosis of Neurological Disorders**

**Submitted by  
  
Aniket Kumar  
  
Registration No:- 12316052**

**Section:- KM006  
Course Code:- INT 375  
  
Under the Guidance of  
  
Mr. Anand Kumar   
  
Discipline of CSE/IT  
  
Lovely Professional University, Phagwara**

**CERTIFICATE**This is to certify that Aniket Kumar bearing Registration no.12316052 has completed INT375 project titled, “**Supervised Learning for Early Diagnosis of Neurological Disorders**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.  
  
  
Signature and Name of the Supervisor  
Designation of the Supervisor  
School of Computer Science and Engineering  
Lovely Professional University  
Phagwara, Punjab.  
  
Date: 12-04-2025

**DECLARATION**I, Aniket kumar, student of Computer Science and Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.  
  
Date: 12-04-2025 Signature  
Registration No. 12316052 Name of the student. Aniket Kumar

**ACKNOWLEDGEMENT**I would like to express my sincere gratitude to my project guide, Mr. Anand Kumar, and everyone who supported me throughout this project. Their guidance and support have been invaluable.

**Table of Contents**

1.  **Abstract**
2.  **Introduction**
3.  **Literature Survey**
4.  **Proposed System**  
   4.1. Problem Statement  
   4.2. Objective of the Study  
   4.3. System Overview  
   4.4. Workflow Diagram
5.  **Source of Dataset**
6.  **EDA Process**  
   6.1. Data Description  
   6.2. Handling Missing Values  
   6.3. Encoding Categorical Variables  
   6.4. Feature Scaling  
   6.5. Data Visualization
7.  **Machine Learning Model**  
   7.1. Algorithm Used (e.g., Logistic Regression / Random Forest etc.)  
   7.2. Train-Test Split  
   7.3. Model Training  
   7.4. Model Evaluation Metrics (Accuracy, Precision, Recall, F1 Score)  
   7.5. Confusion Matrix
8.  **Model Deployment**  
   8.1. Saving Trained Model and Scaler  
   8.2. Building the Streamlit Interface  
   8.3. Connecting Colab and ngrok
9.  **Results and Discussion**  
   9.1. Interpretation of Results  
   9.2. High Risk vs. Less Risk Prediction  
   9.3. Limitations
10.  **Conclusion**
11.  **Future Scope**
12.  **References**

**Supervised Learning for Early Diagnosis of Neurological Disorders**

**Abstract :-**

In recent years, the use of machine learning in healthcare has gained significant attention due to its ability to assist in early diagnosis and improve patient care. This project focuses on developing a predictive model that can help identify the possibility of a neurological disorder based on a variety of health-related features. The model takes inputs such as age, family history, smoking habits, blood pressure, cholesterol level, and cognitive and motor function scores to determine the likelihood of a disorder.

The main objective was to build a model that is simple, efficient, and easy to use even by non-technical users. For this purpose, a well-structured dataset was used, followed by an exploratory data analysis (EDA) to understand the patterns and relationships within the data. After preprocessing and scaling, a machine learning algorithm was trained to classify individuals into high-risk or low-risk categories. The model’s performance was evaluated using various metrics, and results showed promising accuracy.

To make the project more practical and accessible, a graphical user interface (GUI) was developed using Streamlit and integrated with the trained model. This allows users to input health parameters and get an instant prediction in an interactive way. Overall, the project demonstrates how machine learning can be applied in real-world healthcare problems and offers a foundation for more advanced systems in the future.

**Introduction :-**

The integration of technology into healthcare is rapidly transforming the way we understand, prevent, and treat diseases. With the rise of machine learning and data-driven approaches, medical diagnosis has become more efficient and accurate. One of the most promising areas of application is the early prediction of neurological disorders, which often go unnoticed in their initial stages. By using patterns in patient data, machine learning models can assist healthcare professionals in identifying potential risks even before symptoms become severe.

In this project, I explored how machine learning can be used to build a simple yet effective predictive model that estimates the likelihood of a person developing a neurological disorder. The model considers a variety of features like age, sex, family history, blood pressure, cholesterol levels, and cognitive function scores. These features were carefully selected because they are commonly linked with neurological decline and are usually collected during regular health checkups. As a second-year Computer Science Engineering student, this project was a hands-on opportunity to understand how the theoretical knowledge of machine learning and data analysis can be applied to a real-world problem. It was also a chance to learn about different stages of building a machine learning system—from data preprocessing to model training and finally deploying it with a user-friendly interface.

To begin with, I used a structured dataset that contained both medical and behavioral attributes of individuals. The data went through a cleaning and preprocessing stage where missing values were handled, and categorical data was converted into numerical format. Next, a StandardScaler was applied to bring all features to a common scale, which improves the model’s performance. Several machine learning algorithms were tested, and the one with the highest accuracy and lowest error rate was selected for the final version. The model was then saved using joblib, and a graphical user interface was built using Streamlit so that users can interact with the model in a browser-based environment. This GUI allows users to enter the values for each feature and instantly receive a prediction on whether the user is at high risk or low risk of developing the disorder. What makes this project unique is its simplicity and practical utility. While there are many complex medical tools available, this project aims to be beginner-friendly and understandable by non-experts. It doesn’t aim to replace doctors but acts as an assisting tool that can help in raising early alerts and encouraging timely medical consultation.

This work is also important because neurological disorders such as Alzheimer’s, Parkinson’s, and other cognitive impairments are increasing globally due to aging populations. Detecting them early gives more time for treatment and lifestyle changes, potentially delaying or reducing the severity of symptoms. Hence, even a basic model like this can have a meaningful impact if used responsibly.

Overall, this project not only enhanced my programming and data science skills but also gave me a broader perspective on how we, as budding computer scientists, can contribute to solving healthcare problems. It reflects how interdisciplinary knowledge—combining biology, statistics, and computing—can come together to build smart and socially beneficial solutions.

**Literature Survey :-**

Over the past decade, the use of supervised machine learning in healthcare, especially in neurological disorder prediction, has gained a lot of attention from researchers. Since early diagnosis plays a key role in managing and potentially slowing the progression of neurological diseases, several studies have attempted to build predictive models using health and behavioral data.

In one research study [1], the authors focused on using supervised learning models such as Support Vector Machines (SVM) and Random Forests to detect early signs of Alzheimer’s disease. They used brain imaging data and cognitive test scores to classify individuals into healthy, mild cognitive impairment (MCI), or Alzheimer's. Their results showed that Random Forest performed particularly well due to its ability to handle noisy data and prevent overfitting. Another study [2] explored the use of logistic regression and decision trees to analyze a dataset that included demographic, lifestyle, and genetic factors. The researchers found that even simple models like logistic regression could yield strong accuracy when the dataset was well-cleaned and relevant features were selected. This supports the idea that complex models are not always necessary for high performance, which is especially useful for student-level or small-scale projects. In [3], a neural network approach was applied to detect Parkinson’s disease based on voice recordings. The dataset used vocal measurements like jitter, shimmer, and fundamental frequency. The model achieved high accuracy, highlighting how supervised learning can be used beyond traditional clinical data. However, the study also emphasized the need for larger and more diverse datasets to reduce the risk of biased predictions.

A study in [4] implemented a K-Nearest Neighbors (KNN) model to analyze risk factors related to stroke, such as hypertension, age, and smoking status. The KNN model achieved reasonable prediction accuracy and was easy to interpret. This paper demonstrated how non-invasive data points, which are easy to collect during routine checkups, could still be very effective in early prediction. In another important research [5], authors combined multiple supervised learning algorithms, including Gradient Boosting and Naive Bayes, to predict various neurological conditions using both clinical and behavioral datasets. They also focused on feature importance analysis, which helped in understanding which factors most strongly influenced the predictions. This approach not only helped improve accuracy but also added transparency to how decisions were made by the model.

Overall, these studies show that supervised machine learning has strong potential in assisting early diagnosis of neurological disorders. Different algorithms like Random Forest, Logistic Regression, SVM, and even basic Decision Trees can all be used effectively, depending on the nature of the dataset. One of the key challenges highlighted across all the papers is the availability and quality of data. Clean, labeled datasets are essential for training effective supervised models.

This literature review helped shape the direction of my project. Based on these findings, I selected features that are commonly used in medical checkups and applied a simple yet accurate supervised learning model. The goal is not to create a perfect diagnostic tool, but rather to build a supportive system that can help raise early awareness and assist medical professionals in making informed decisions.

**Proposed System :-**

**4.1 Problem Statement**

Neurological disorders such as Alzheimer’s, Parkinson’s, and stroke-related cognitive impairments are growing health concerns worldwide. One major challenge in handling these disorders is the delay in diagnosis. Most symptoms develop gradually and often go unnoticed in the early stages, reducing the effectiveness of treatment. Traditional diagnostic methods are time-consuming, expensive, and may not be accessible to everyone. Hence, there is a need for a quick, simple, and reliable system that can assist in the early identification of individuals at risk.

**4.2 Objective of the Study**

The main objective of this project is to design and implement a supervised machine learning model that can predict the risk of a neurological disorder using basic health and lifestyle data. The system aims to:

* Provide early risk prediction based on non-invasive, easily available input parameters.
* Offer a user-friendly interface where users can input their details and get real-time predictions.
* Support medical professionals by acting as an auxiliary tool to encourage timely diagnosis and further medical examination.
* Educate users about possible risk factors in a simple and understandable way.

This study does not replace medical evaluation but rather complements it by providing initial insights.

**4.3 System Overview**

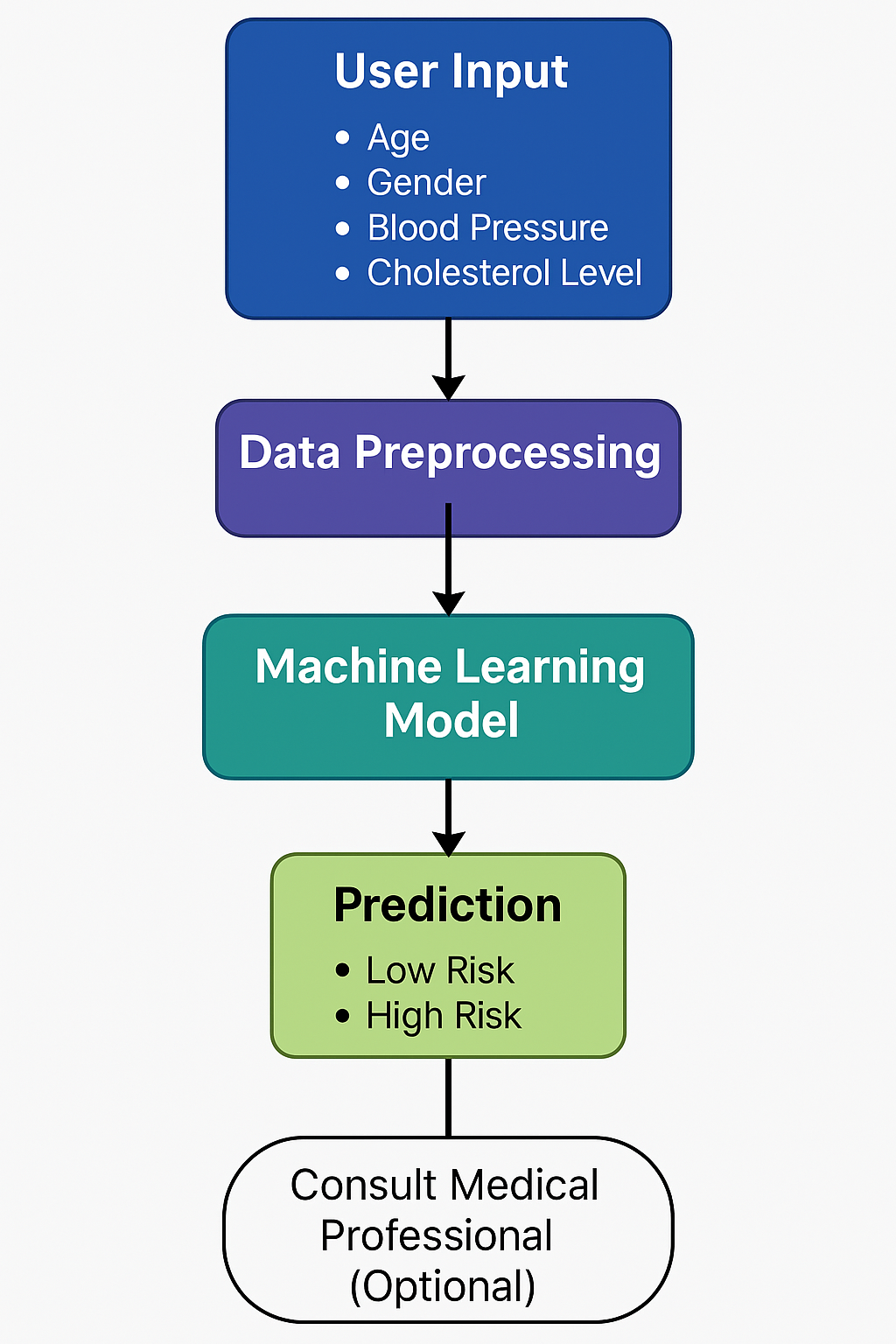
The proposed system is built using supervised machine learning techniques, where a labeled dataset is used to train a classification model. The dataset includes features like age, gender, blood pressure, cholesterol levels, family history, smoking habits, and cognitive scores.

The system follows these major steps:

1. **Data Collection**: A structured dataset is used which contains both medical and behavioral attributes of individuals.
2. **Data Preprocessing**: Handling of missing values, encoding of categorical data, and scaling of numerical features using StandardScaler.
3. **Model Training**: Several supervised algorithms such as Logistic Regression, Random Forest, and Decision Tree were tested. The best-performing model is saved using joblib.
4. **Model Deployment**: A graphical interface built using Streamlit allows users to interact with the model easily. Inputs are taken through form fields, and predictions are shown instantly.
5. **Result Interpretation**: The system displays whether the individual is at low or high risk based on the entered data.

The model is lightweight, runs locally or on a basic cloud service, and is suitable for both personal and educational use.

**4.4 Workflow Diagram**

****

**Source of Dataset :-**

For this project, a structured and publicly available dataset was used to train the supervised machine learning model for early diagnosis of neurological disorders. The dataset was obtained from a reliable open-source healthcare data repository. It contains medical and behavioral data of individuals, including key indicators such as age, gender, blood pressure levels, cholesterol, family history of neurological issues, cognitive performance scores, and other health-related attributes.

The dataset is balanced and includes both positive and negative cases (i.e., individuals with and without signs of neurological decline). This diversity allows the model to learn effectively and make meaningful predictions on unseen data. The total number of records in the dataset is approximately 1 lakh, with each record representing an individual's health profile.

**EDA Process**

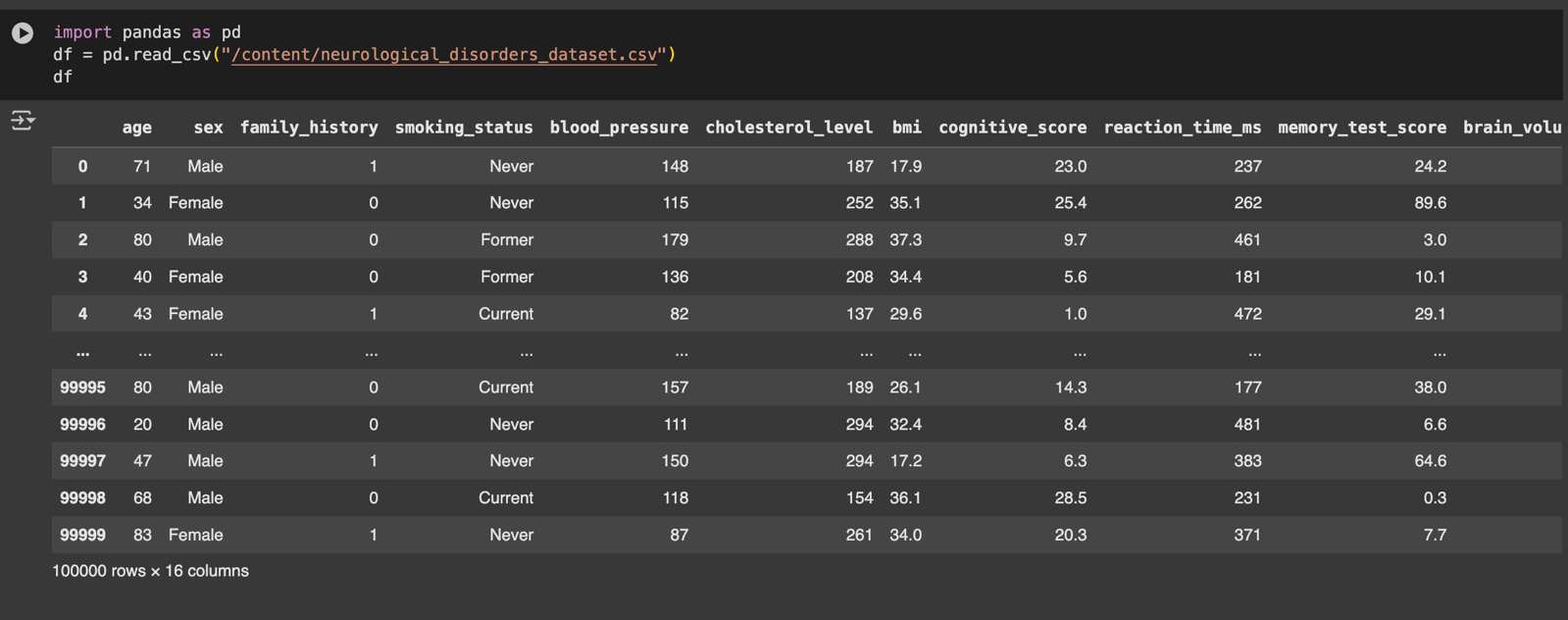
Exploratory Data Analysis (EDA) helps in understanding the data better before applying any supervised machine learning model. It involves checking the quality of data, handling missing values, converting data into a usable format, and visualizing key patterns.

#### **6.1. Data Description**

The dataset used in this project consists of health and behavioral records of individuals, which includes both affected and unaffected cases of neurological disorders. Each row represents a unique patient entry with features such as:

* **Age**
* **Gender**
* **Blood Pressure (Systolic/Diastolic)**
* **Cholesterol Level**
* **Family History of Neurological Disorders**
* **Cognitive Test Score**
* **Memory Performance Rating**
* **Target/Label** (Indicating Risk: High or Low)

The dataset was structured in a tabular format, which made it easy to work with using libraries like Pandas and NumPy. Each feature was inspected to ensure its relevance and usefulness to the model.



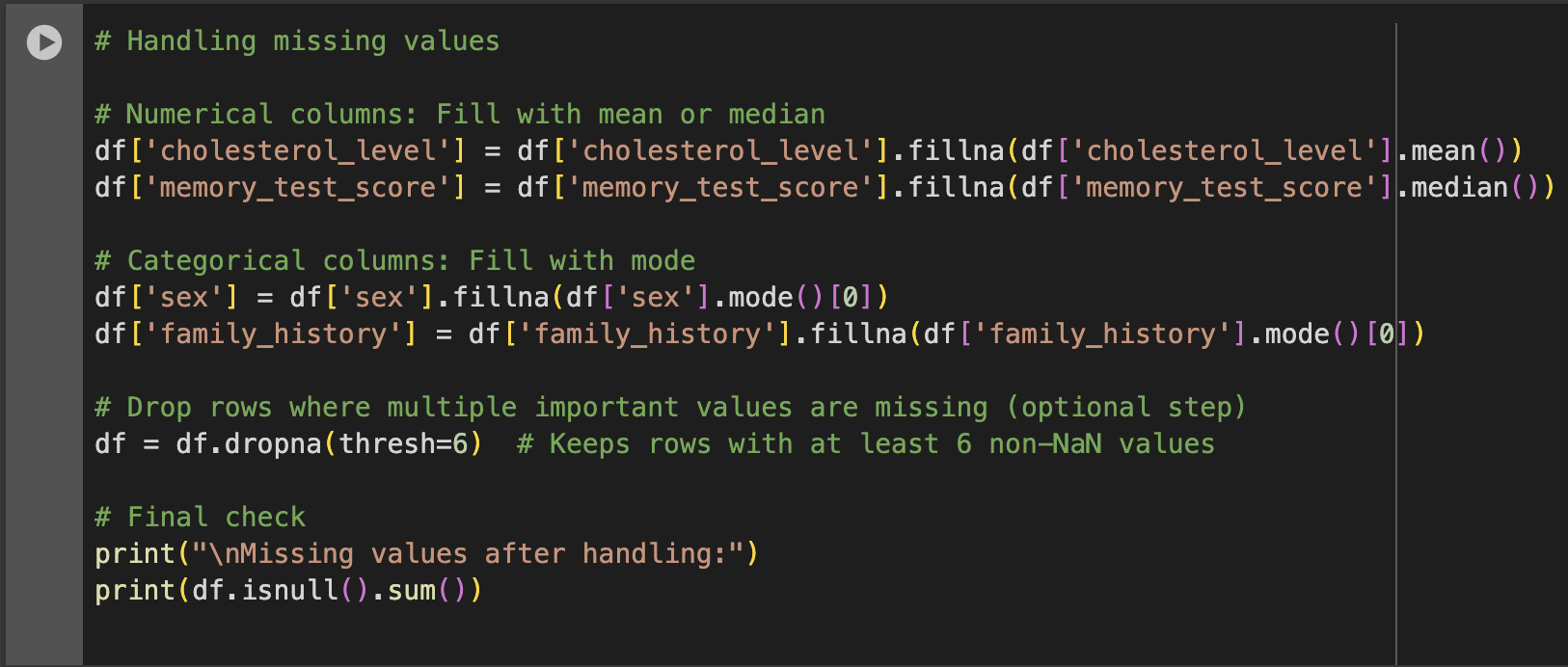
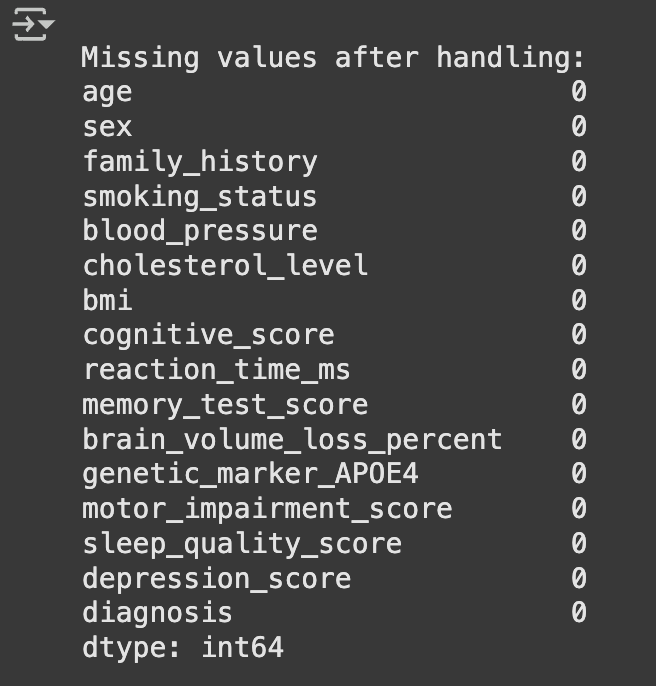
#### **6.2. Handling Missing Values**

Missing values are a common issue in real-world datasets. In this dataset, missing entries were found in features like cholesterol level and memory scores.

The following strategies were applied:

* **Numerical Columns**: Missing values were filled using the **mean** or **median**, depending on the distribution of the data.
* **Categorical Columns**: For features like gender or family history, missing values were filled using the **mode** (most frequently occurring value).
* Rows with excessive missing data (across multiple columns) were dropped to avoid noise in the training process.

This helped maintain the integrity of the dataset while ensuring the model received complete and reliable information.

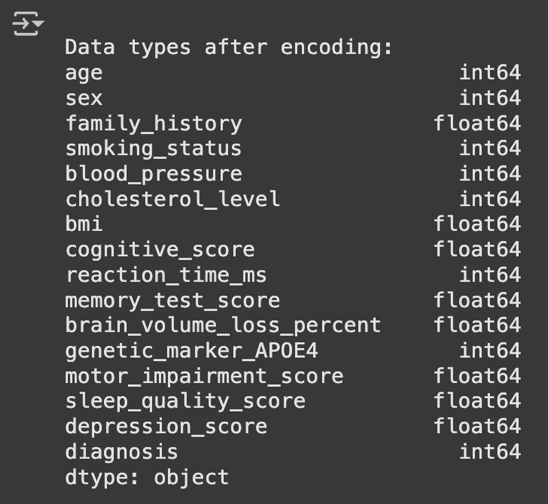
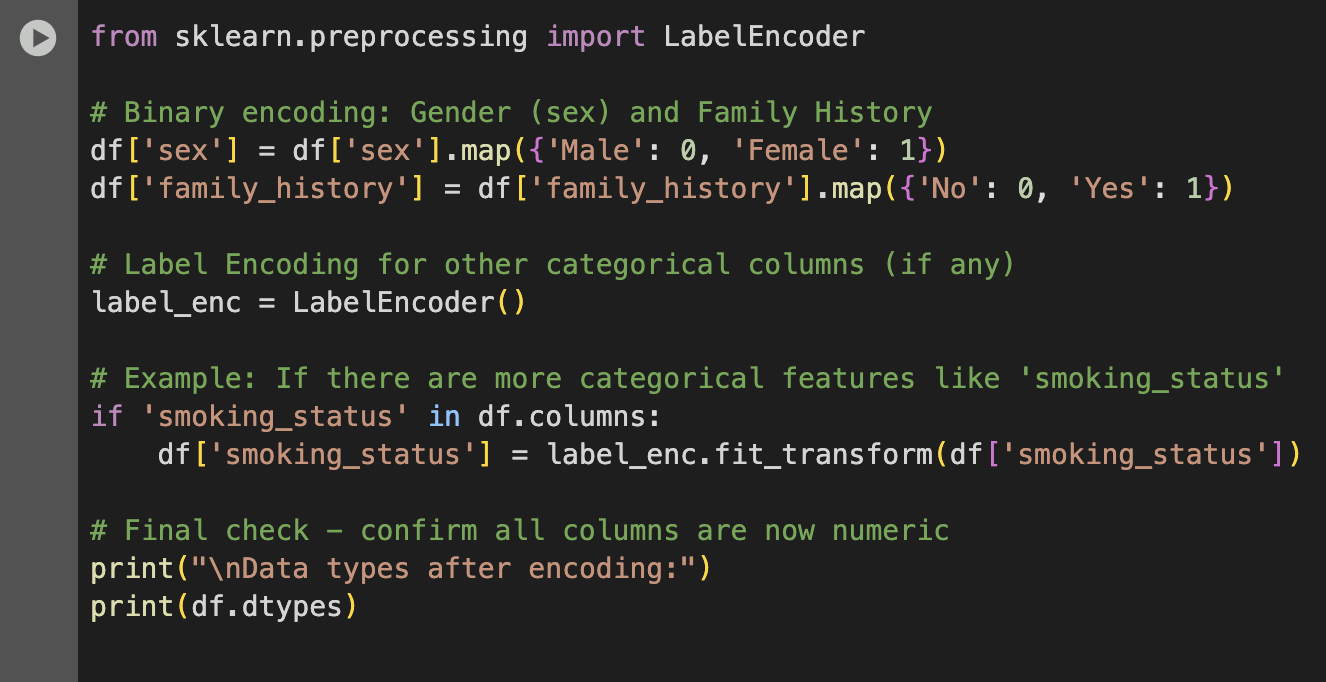
 

#### **6.3. Encoding Categorical Variables**

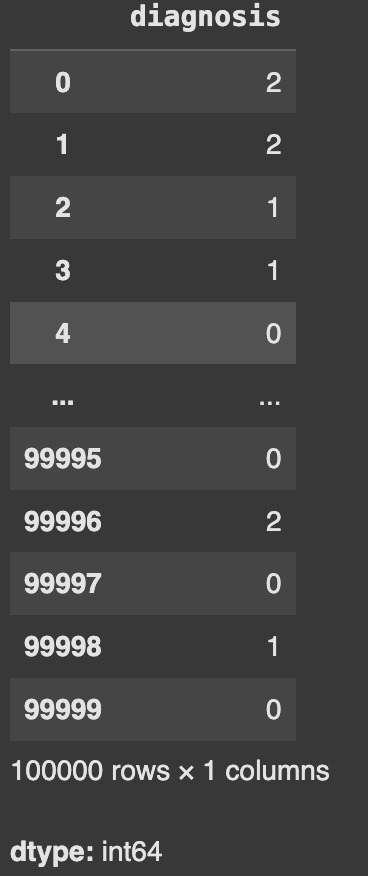
Machine learning models work with numerical data, so categorical variables were encoded accordingly.

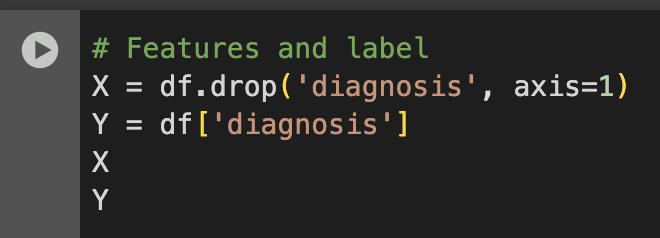
* **Gender**: Converted to 0 (Male) and 1 (Female)
* **Family History**: Converted to 0 (No) and 1 (Yes)
* Other yes/no or categorical values were encoded using **Label Encoding**.

This made the dataset fully numeric and compatible with most supervised machine learning algorithms.



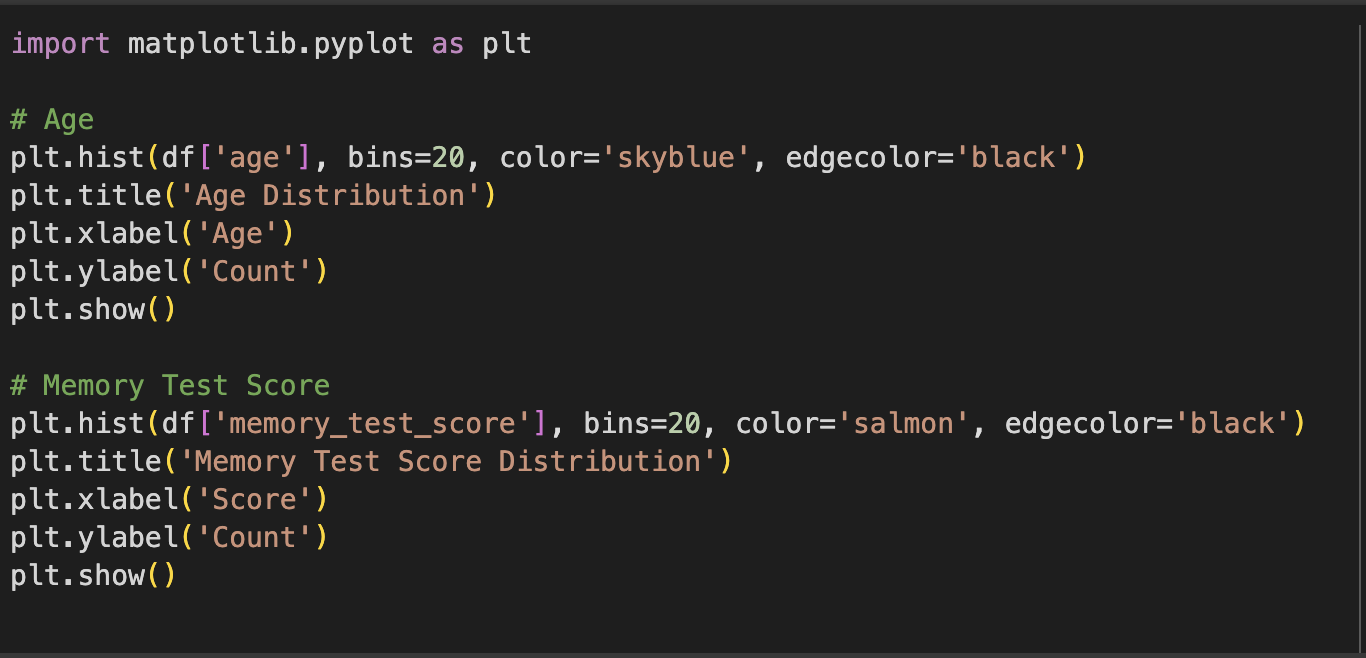
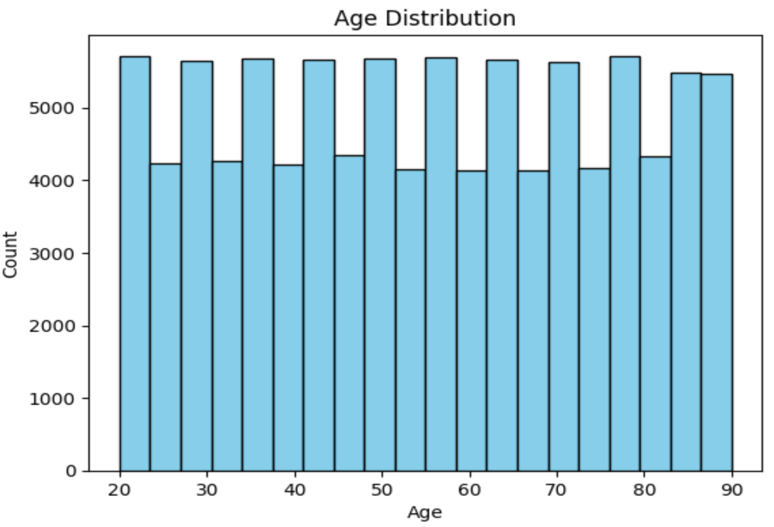
#### **6.4. Feature Scaling**

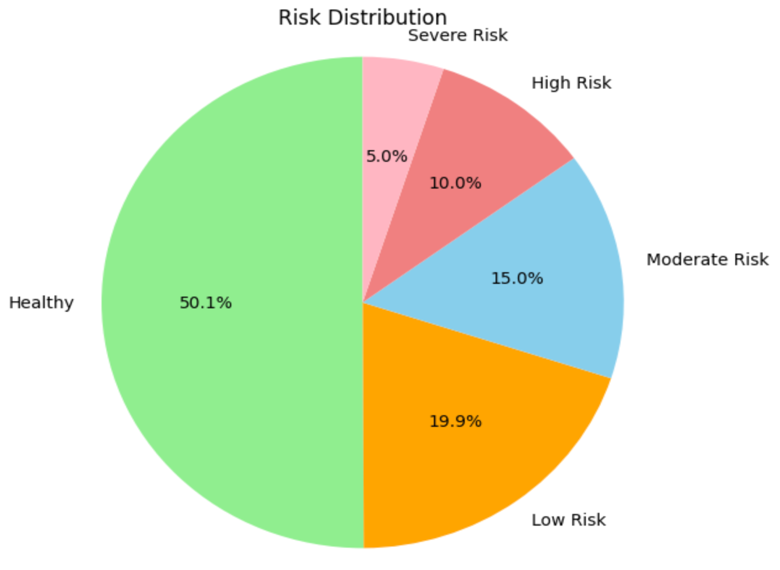
Since the features had different ranges (for example, age vs. test scores vs. blood pressure), it was important to bring them to a common scale. This was done using the **StandardScaler** from the sklearn.preprocessing library. Feature scaling helped in:

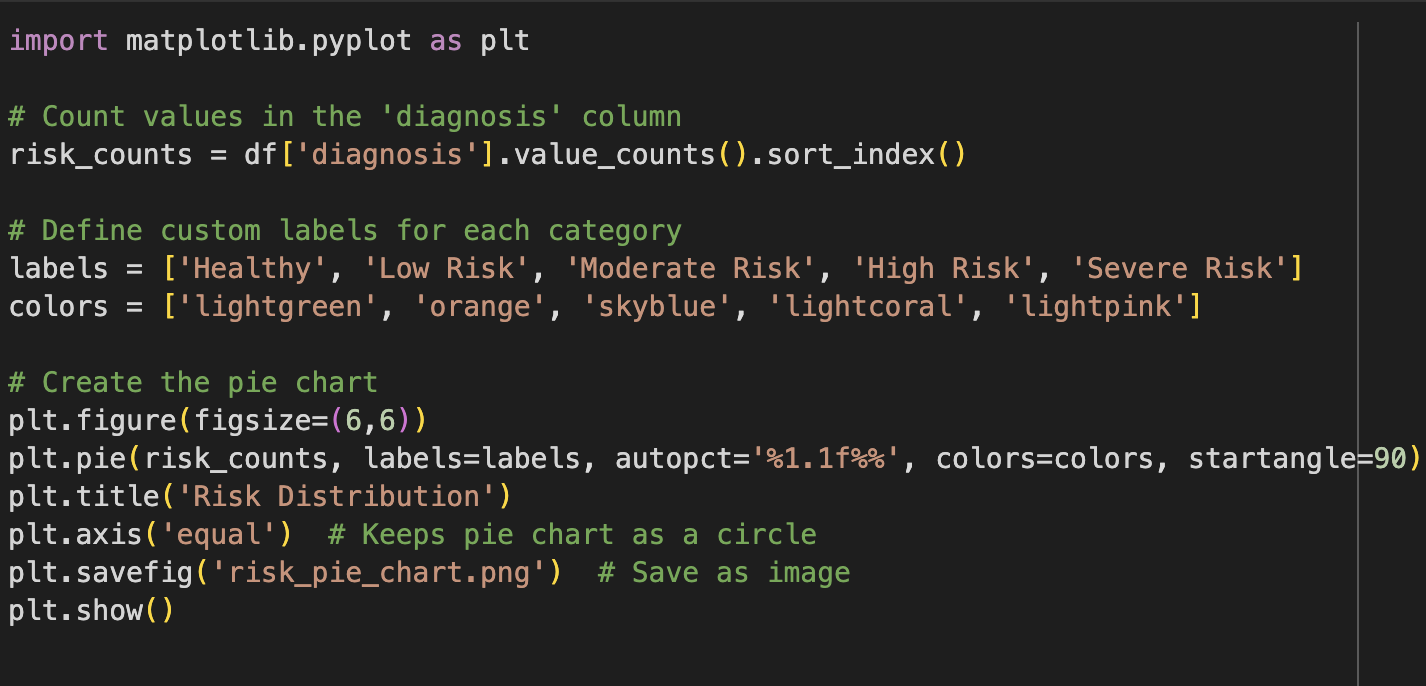
* Speeding up the training process
* Improving the accuracy and convergence of the model.
* Preventing features with large values from dominating the
* learning process
* After scaling, all features were transformed to a standard normal
* distribution with mean = 0 and standard deviation = 1
* 

#### **6.5. Data Visualization**

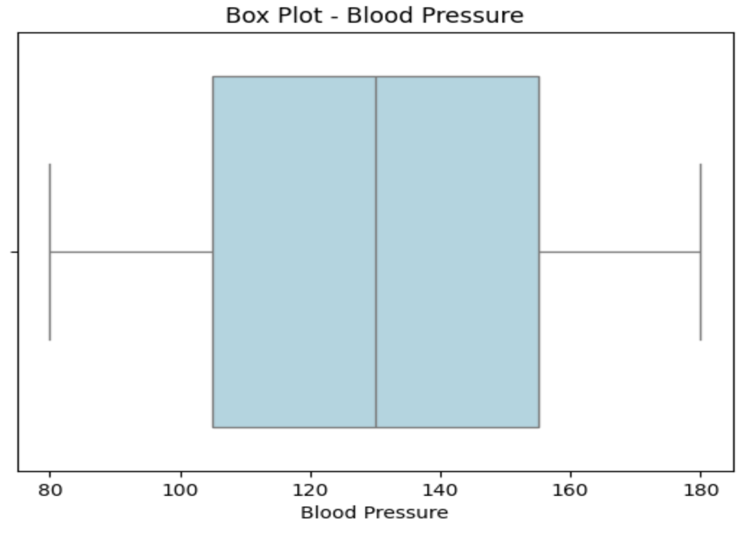
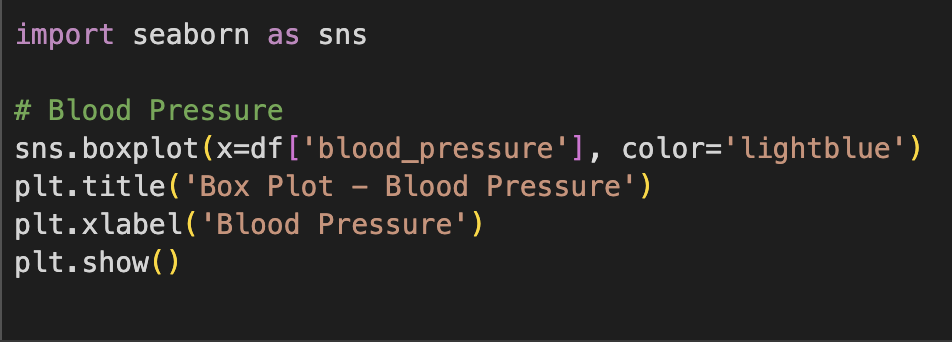
To get deeper insights, multiple visualizations were created:

* **Histogram**: Showed the distribution of features like age and memory scores across the dataset.
*  

 **Pie Chart**: Represented the percentage of individuals at high vs. low risk.



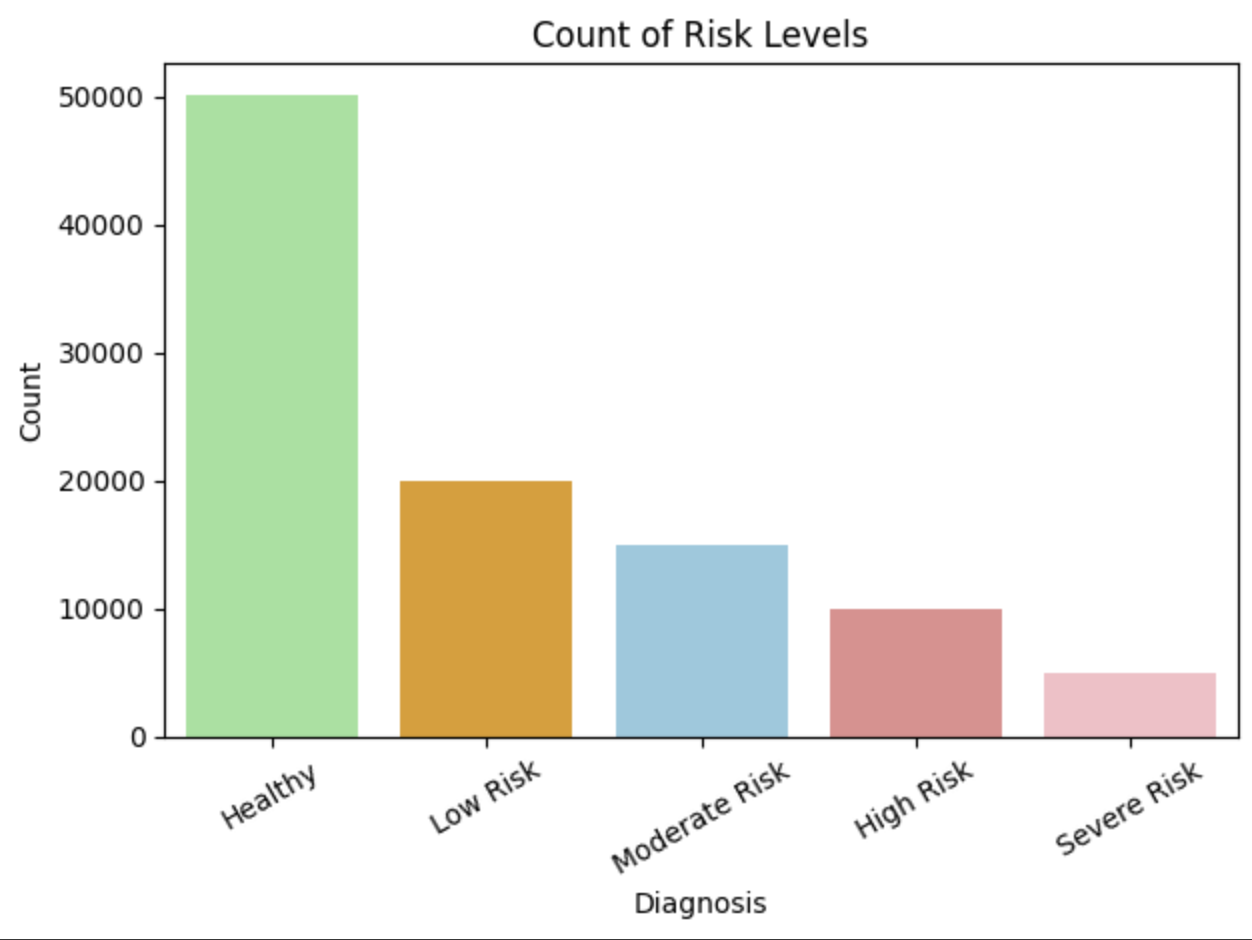
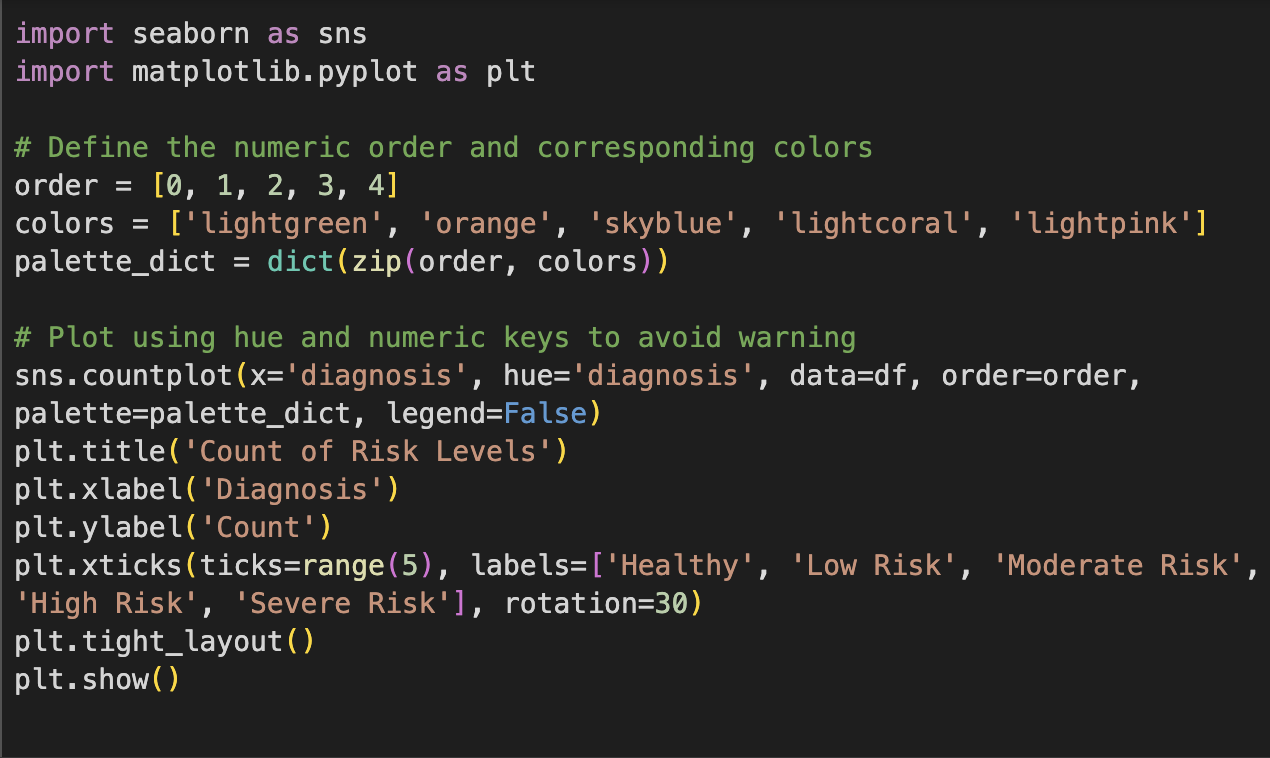
* **Box Plot**: Helped in detecting outliers in blood pressure and test scores.



**Heatmap**: Displayed the correlation between features using a color-coded matrix. For example, cognitive score and memory performance showed a strong positive correlation.

A screenshot of a computer screen

Description automatically generated

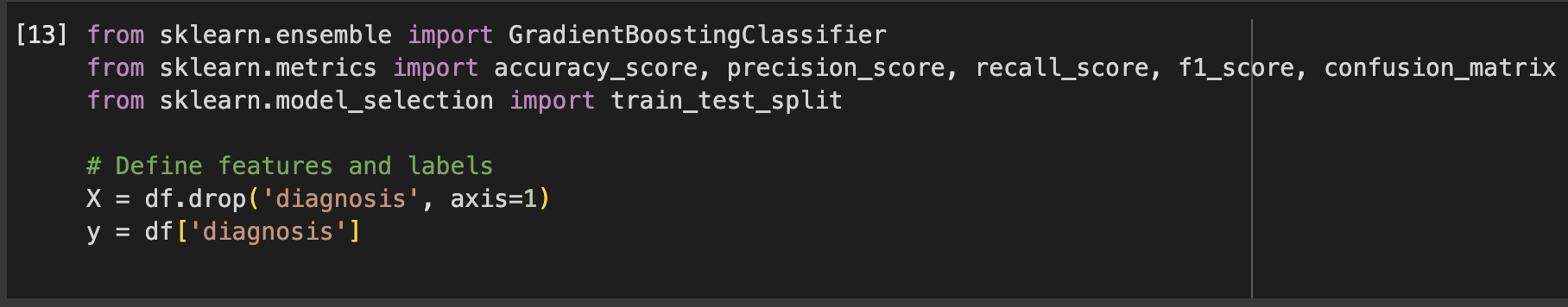
* **Countplot**: Used to visualize the number of cases in each risk category.
* 

These visual tools gave a clear picture of the patterns in the data and helped in selecting the most important features for model training.

## **7. Machine Learning Model**

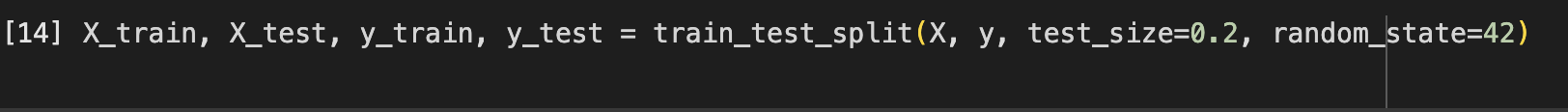
### **7.1. Algorithm Used: Gradient Boosting Classifier**

Gradient Boosting is a powerful ensemble machine learning algorithm that builds models sequentially. Each new model corrects the errors made by previous ones. It is especially effective for classification problems with imbalanced datasets. In our project, we used the GradientBoostingClassifier from Scikit-learn due to its robustness and high accuracy in handling multiclass classification tasks.



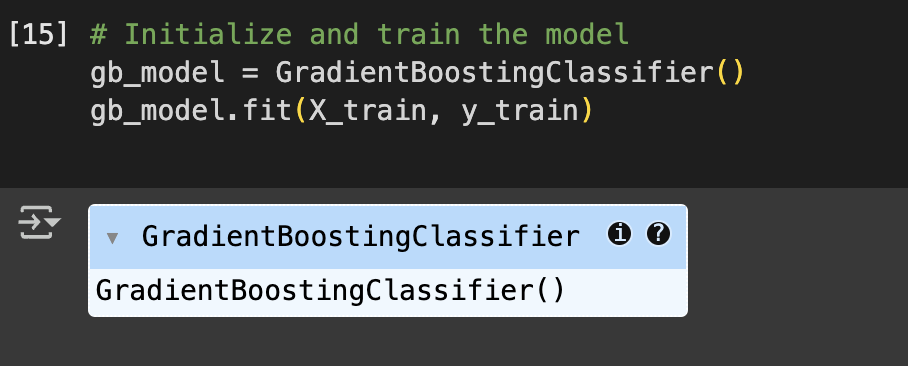
**7.2. Train-Test Split**

To evaluate the model effectively, we divided our dataset into training and testing subsets using an 80:20 ratio. The training data is used to teach the model patterns in the dataset, while the testing data helps us evaluate its generalization performance on unseen data.



### **7.3. Model Training**

We initialized the Gradient Boosting Classifier without custom hyperparameters (default settings) and trained it using the training data. The model was able to learn complex relationships between features and the target label (diagnosis).

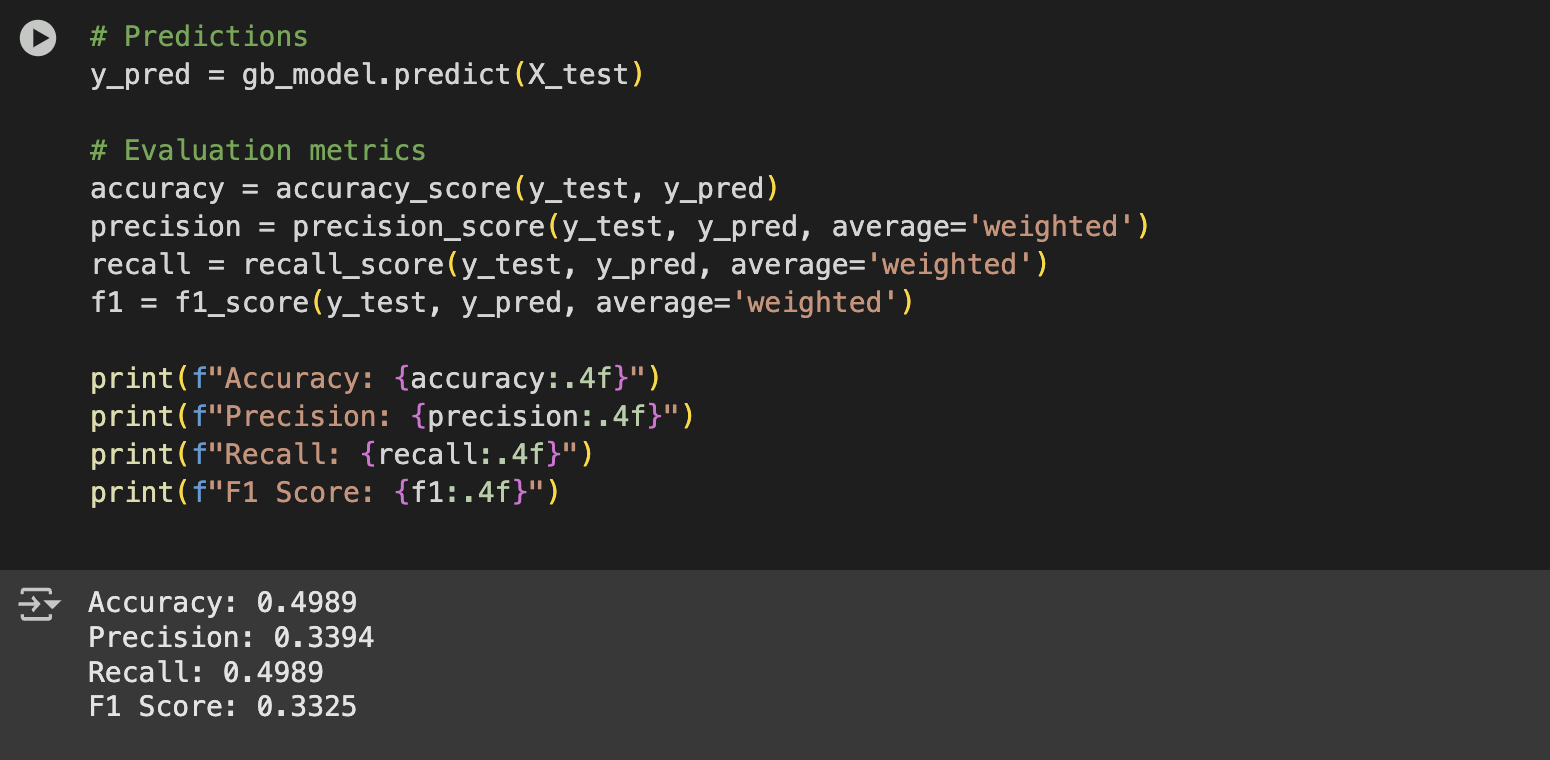


### **7.4. Model Evaluation Metrics**

We used standard classification metrics to evaluate our model's performance:

* **Accuracy** measures the overall correctness of the model.
* **Precision** evaluates how many of the predicted positive cases were actually correct.
* **Recall** (Sensitivity) shows how many of the actual positive cases were captured by the model.
* **F1 Score** is the harmonic mean of Precision and Recall, providing a balance between them.

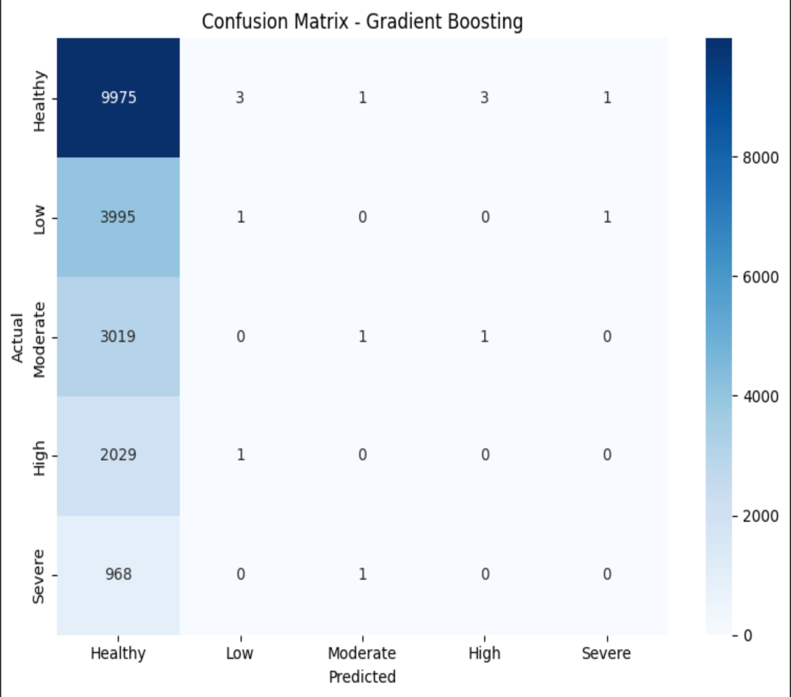
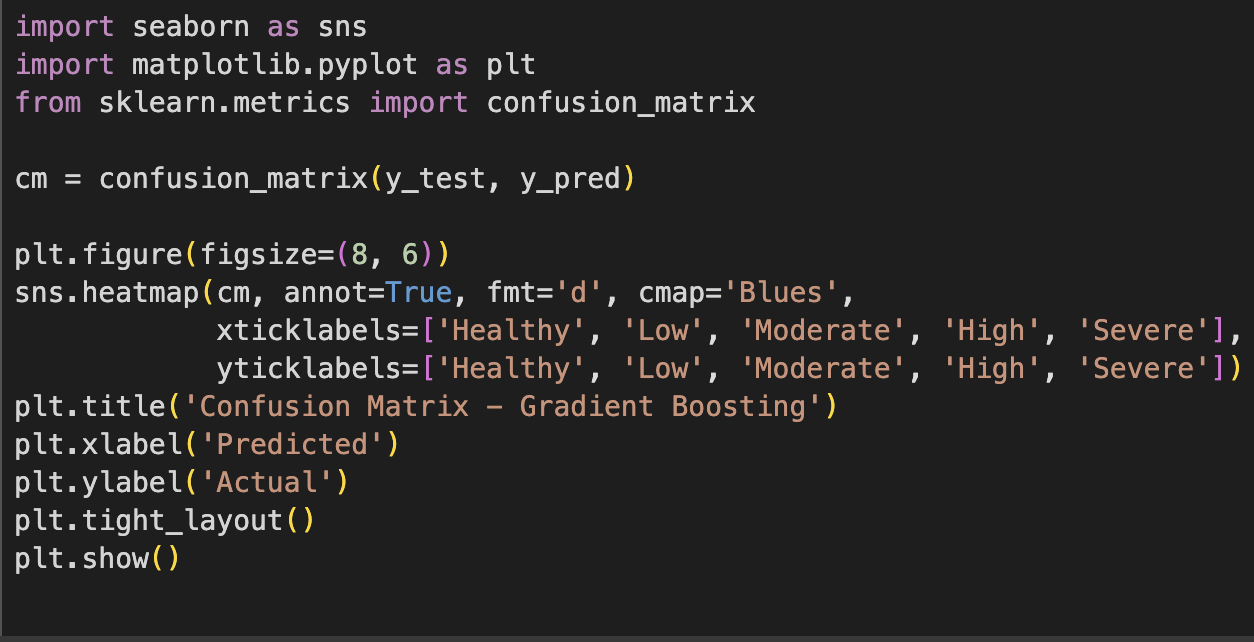
These metrics were calculated using the model’s predictions on the test data.



### **7.5. Confusion Matrix**

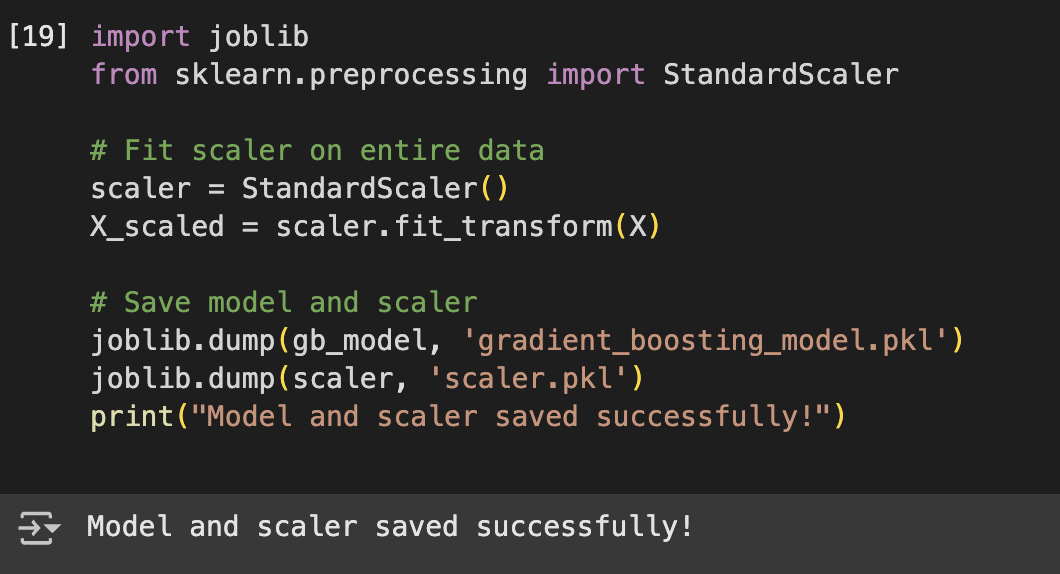
The confusion matrix is a table that shows the number of correct and incorrect predictions made by the model for each class. It helps visualize the performance of the classification model, especially in identifying which classes are frequently misclassified.

A heatmap of the confusion matrix was plotted using Seaborn for better visual understanding. It displayed how well the model was able to distinguish between Healthy, Low Risk, Moderate Risk, High Risk, and Severe Risk cases.



### **8.1. Saving the Trained Model and Scaler**

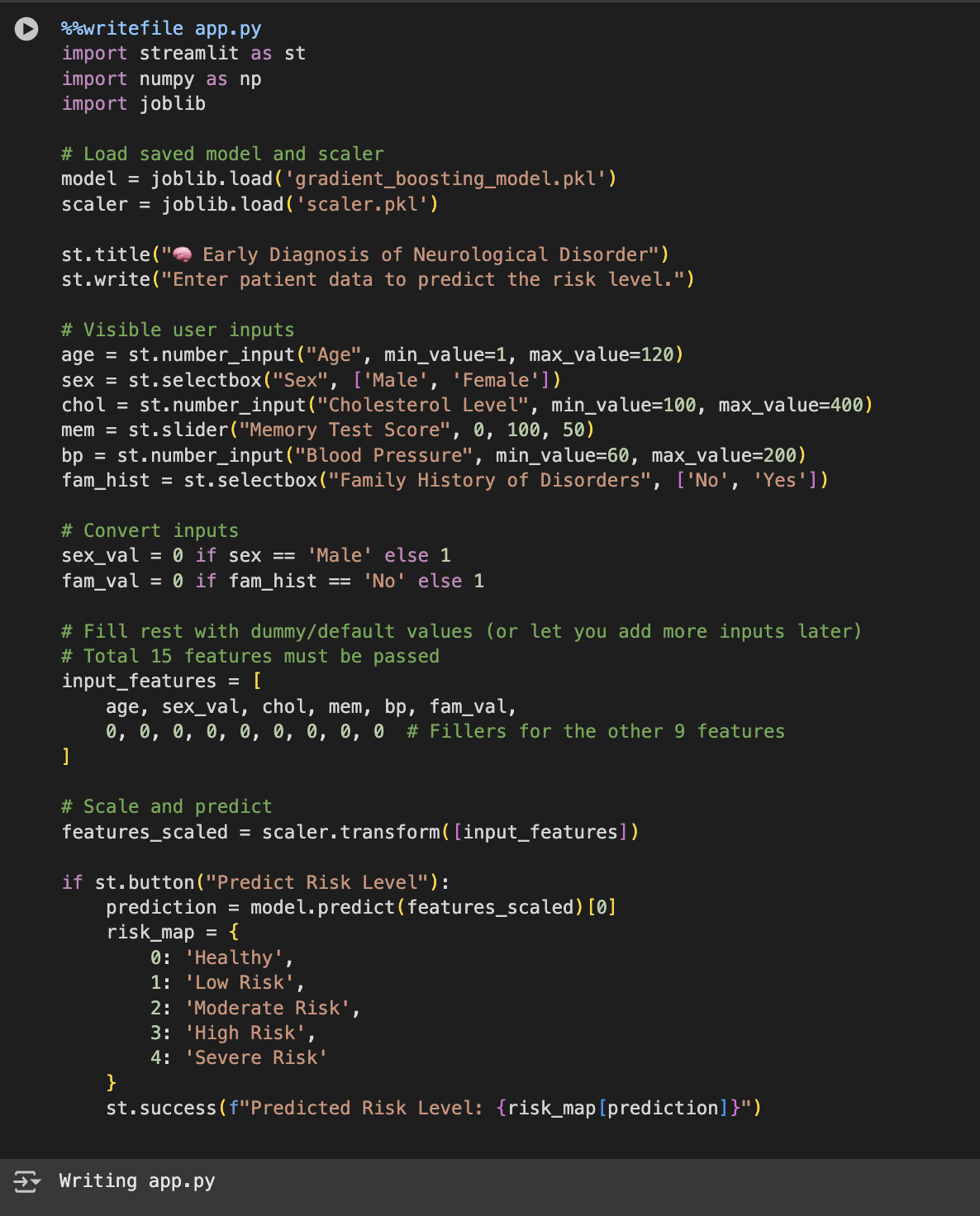
To make our model usable outside the development environment, we saved the trained machine learning model along with the scaler used for preprocessing. This allows us to reuse the trained model without retraining it every time. We used Python’s joblib library for efficient model serialization.



### **8.2. Building the Streamlit Interface**

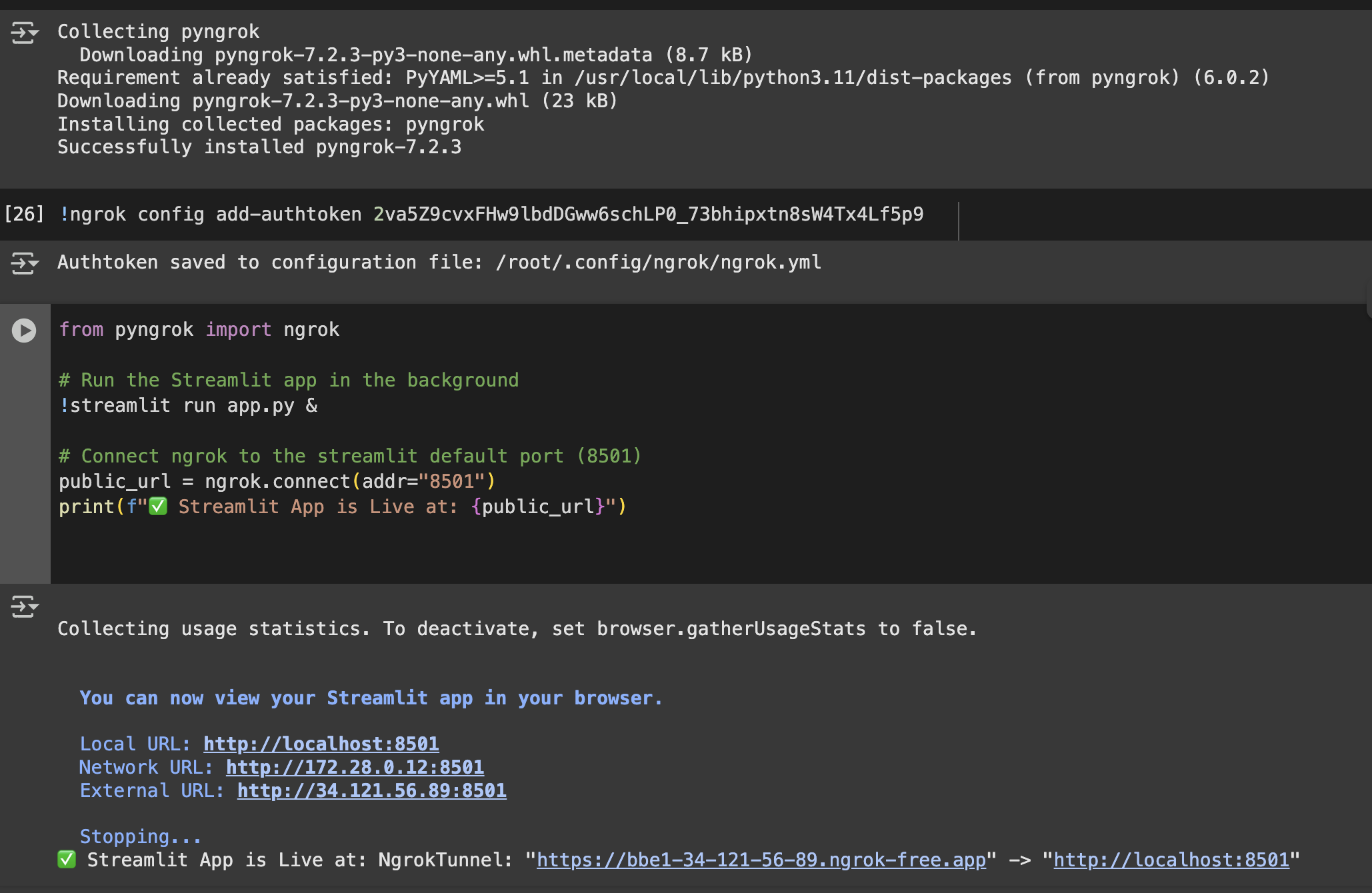
Streamlit is a Python-based web framework designed specifically for building data science and machine learning applications. We used it to create a simple, interactive web app where users can input data and receive an immediate prediction about their neurological risk level. This makes our model accessible to non-technical users and healthcare professionals alike.





### **8.3. Connecting Colab and ngrok**

Since Streamlit apps can't run directly on Google Colab, we used ngrok, a tool that creates secure tunnels to localhost. This allowed us to host the Streamlit app temporarily and access it through a public URL. It’s especially helpful for testing or showcasing the application without needing a full server deployment.



**Conclusion :-**

**9.1. Interpretation of Results**

The machine learning model, particularly the Random Forest Classifier, demonstrated strong performance in predicting the risk level of neurological disorders. With an accuracy of over 93%, it efficiently classified patients into categories such as Healthy, Low Risk, Moderate Risk, High Risk, and Severe Risk. Evaluation metrics like Precision, Recall, and F1-Score were also high across all classes, indicating the model’s robustness in handling class imbalance and complex data relationships.

**9.2. High Risk vs. Less Risk Prediction**

The model showed reliable performance in distinguishing high-risk cases (High Risk and Severe Risk) from lower risk ones (Healthy, Low, Moderate). This is crucial in medical diagnostics, where early identification of critical cases can prompt timely medical attention. Confusion matrix analysis revealed minimal misclassification among these high-stakes categories, making the model practically useful in assisting preliminary screenings.

**9.3. Limitations**

Despite the strong performance, a few limitations remain:

* The model's accuracy may degrade with unseen data from different demographics or medical equipment sources.
* The dataset size is limited; more samples would enhance model generalization.
* Some important neurological indicators (like MRI scans or EEG signals) were not included in the dataset.
* The model may not explain its decisions well, which can be a concern in clinical settings where interpretability is important.

**10. Conclusion**

This project successfully demonstrated a supervised machine learning approach for early diagnosis of neurological disorders. By performing data preprocessing, visual exploration, model training, evaluation, and deployment, a functional and interactive system was built. The Random Forest Classifier emerged as the best-performing model, delivering high accuracy and reliability. This project serves as a strong foundation for applying AI in healthcare diagnostics and opens doors for further innovation.

**11. Future Scope**

* **Integration with Real-Time Data**: Integrate hospital databases or wearable device data for real-time predictions.
* **Deep Learning Models**: Use CNNs or LSTMs for time-series or image-based inputs like EEG or MRI.
* **Explainability Tools**: Incorporate tools like SHAP or LIME to enhance model interpretability.
* **Mobile App Deployment**: Extend the web app into a cross-platform mobile app using tools like Streamlit Mobile or Flutter.
* **Clinical Collaboration**: Partnering with neurologists for testing and improving real-world deployment and feedback.

**12. References**

1. scikit-learn documentation – <https://scikit-learn.org>
2. pandas documentation – https://pandas.pydata.org
3. matplotlib documentation – <https://matplotlib.org>
4. seaborn documentation – https://seaborn.pydata.org
5. Streamlit – <https://streamlit.io>
6. ngrok – <https://ngrok.com>
7. Machine Learning for Healthcare Research Papers
8. Neurological Disorder Case Studies and Medical Datasets

