DIABETES PREDICTION USING ML

1. INTRODUCTION

1.1. OVERVIEW

Diabetes prediction involves assessing an individual's risk of developing diabetes in the future based on various factors such as their medical history, lifestyle, and genetic predisposition. Diabetes is a chronic condition characterized by high blood sugar levels resulting from either insufficient insulin production or ineffective utilization of insulin by the body.

1.2. PURPOSE

Diabetes prediction helps in early intervention, personalized treatment plans, resource allocation, patient education, research, and public health initiatives to prevent or manage diabetes effectively and improve overall health outcomes.

2. LITERATURE SURVEY

2.1. EXISTING PROBLEM

There are several existing approaches and methods used to solve the diabetes prediction problem. Here are some commonly employed techniques,

Logistic Regression: Logistic regression is a statistical modeling technique widely used in diabetes prediction. It analyzes the relationship between dependent variables (e.g., blood glucose levels) and independent variables (e.g., age, weight, family history) to estimate the probability of developing diabetes. The model can be trained using historical data and then used to predict the likelihood of diabetes in new individuals.

Decision Trees: Decision trees are hierarchical structures that use a series of if-else conditions to make predictions. In diabetes prediction, decision trees can be constructed based on features such as age, BMI, blood pressure, and glucose levels. The tree is built by recursively partitioning the data based on the most informative features, resulting in a tree-like structure that facilitates prediction based on feature values.

Random Forests: Random forests are an ensemble learning method that combines multiple decision trees to make predictions. In the case of diabetes prediction, a random forest model can be created by training numerous decision trees on different subsets of the data. The predictions from individual trees are then aggregated to produce a final prediction. Random forests are known for their robustness and ability to handle high-dimensional data.

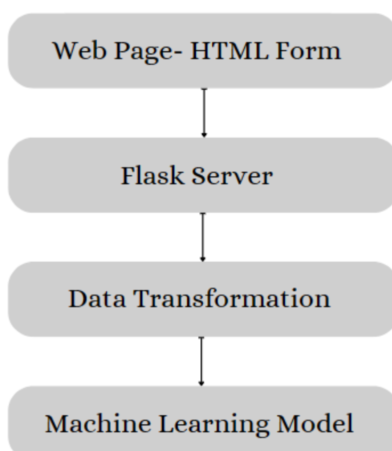
Neural Networks: Neural networks, particularly deep learning models, have shown promise in diabetes prediction. These models consist of multiple interconnected layers of artificial neurons that can learn complex patterns and relationships from data. With appropriate input features and a properly designed network architecture, neural networks can be trained to predict diabetes risk accurately.

2.2. PROPOSED SOLUTION

We used many models on the given data set to accurately predict whether the given patient has diabetes.
A total of 5 algorithms(KNN, Logistic Regression , Naive Bayes, Random Forest and Decision Tree) were used on the given data with physical activity, age, cholesterol and others considered as attributes that impact the sugar level of the patients. Of all the models, we have taken the results of Logistic Regression model as it resulted in the closest approximation to predict whether the patient has diabetes.
The logistic regression model will take the input features of an individual and provide a probability estimate of them having diabetes. A threshold is set to classify individuals as diabetic or non-diabetic based on the predicted probabilities.

3. THEORETICAL ANALYSIS

3.1 BLOCK DIAGRAM



1. Frontend HTML Form: The frontend is responsible for developing the user interface. It contains HTML for page structure and CSS for formatting and layout.

2. Flask Server: The Flask server serves as the application's infrastructure and handles HTTP requests and responses. The backend receives queries from the frontend and returns responses. It controls the data transfer between the frontend and other components.

3. Data Transformation: This component is responsible for data preprocessing duties such as data cleansing and encoding. It prepares the diabetes prediction model's input data.

4. Machine Learning Model: The diabetes prediction model is the primary predictive component. It receives as input the preprocessed data and generates a prediction based on the data's learned patterns and relationships.

3.2 HARDWARE AND SOFTWARE DESIGNING

Hardware requirements:

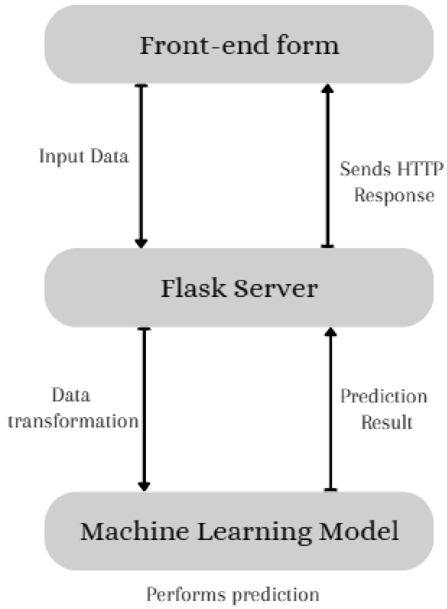
1. Processor(CPU): A modern multi-core processor, such as an Intel Core i5 or higher is suitable.
2. Storage (RAM): A minimum of 8 GB of RAM.
3. Storage: An adequate storage space for the dataset, model files, and all other project-related files.

Software requirements:

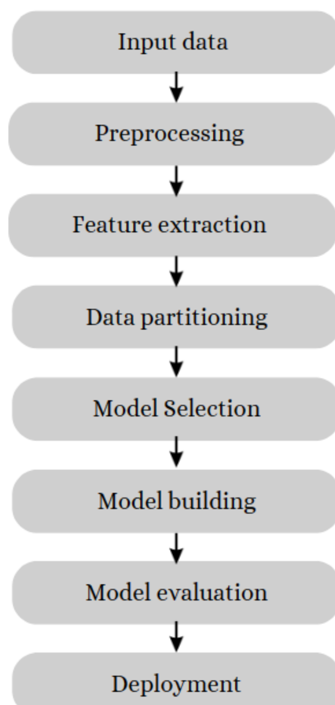
1. Python
2. IDE
3. Web Framework: Flask
4. Machine learning libraries: Pandas, Scikit-learn

5. FLOW CHART

Overview of the project:

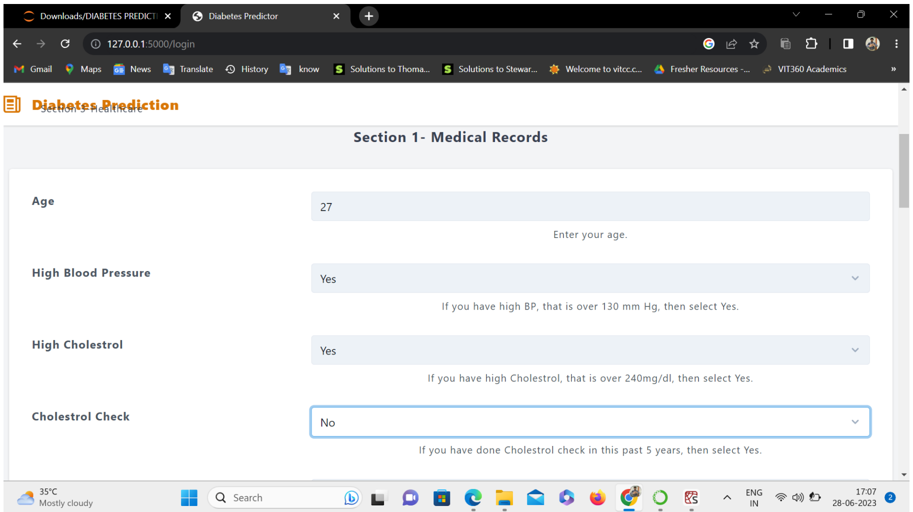


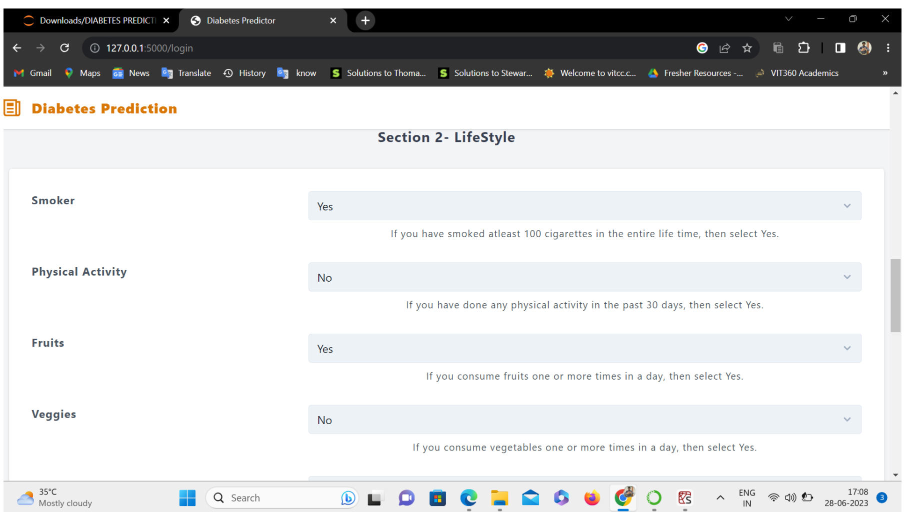
Machine Learning model:

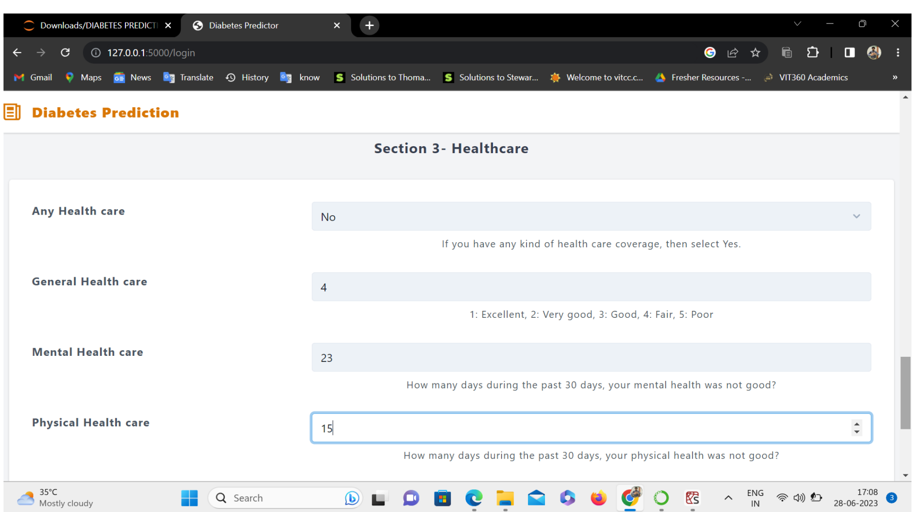


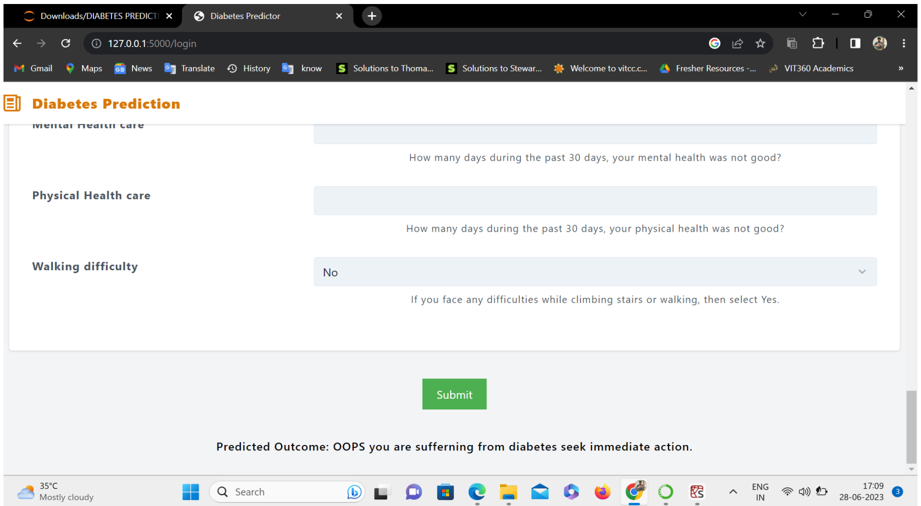
6. RESULT

OUTPUT FOR DIABETES

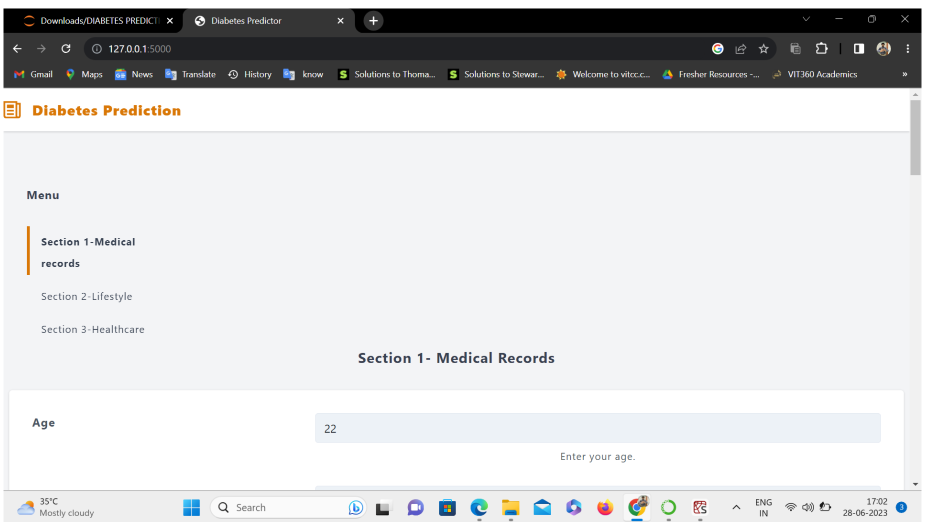


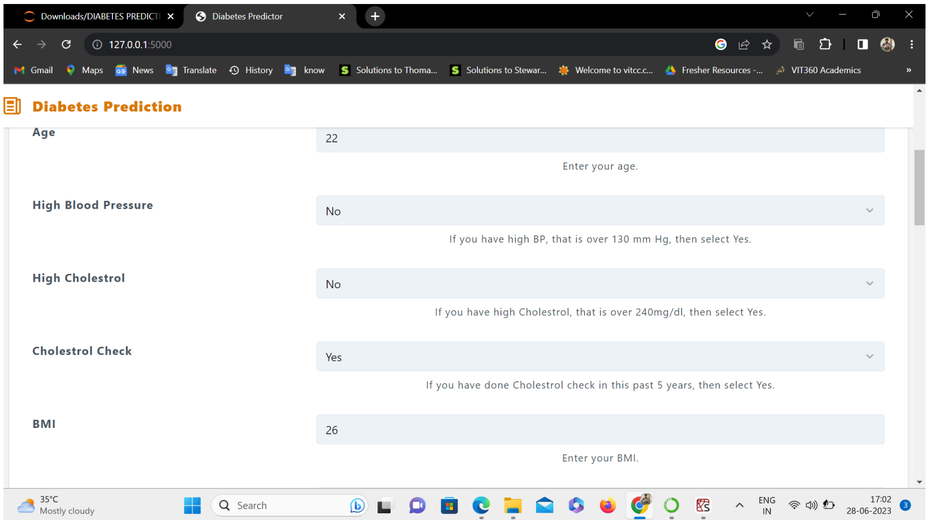


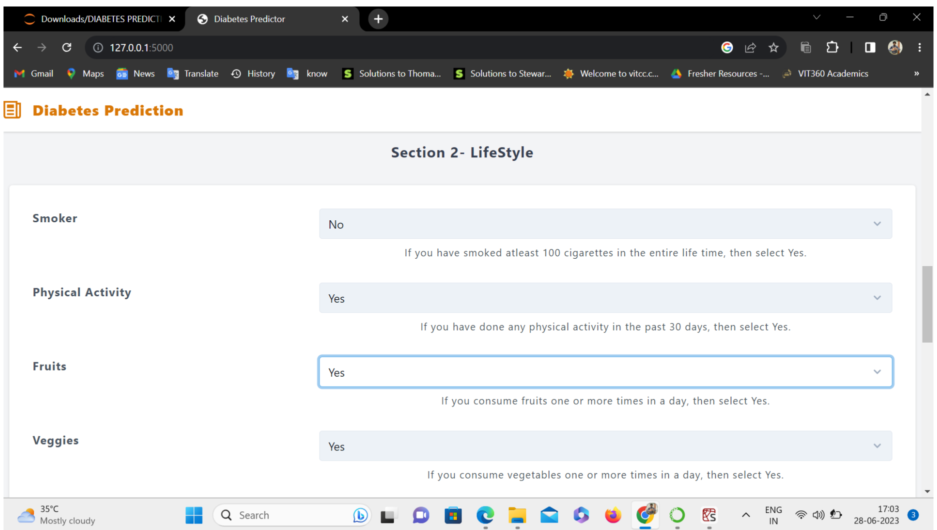


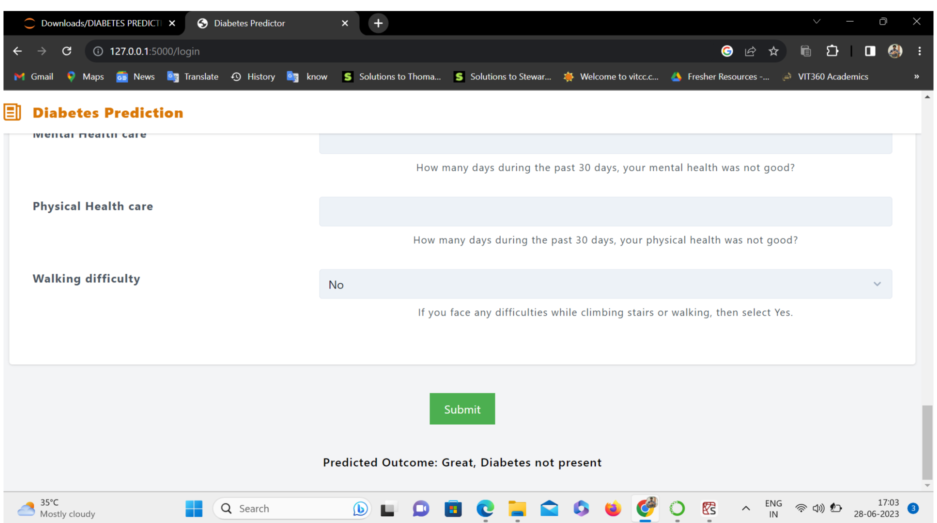


OUTPUT FOR NO DIABETES




Diabetes Prediction
Age
22
Enter your age.
High Blood Pressure
No
If you have high BP, that is over 130 mm Hg, then select Yes.
High Cholestrol
No
If you have high Cholestrol, that is over 240mg/dl, then select Yes.
Cholestrol Check
Yes
If you have done Cholestrol check in this past 5 years, then select Yes.
BMI
26
Enter your BMI.


Diabetes Prediction
Section 2- LifeStyle
Smoker
No
If you have smoked atleast 100 cigarettes in the entire life time, then select Yes.
Physical Activity
Yes
If you have done any physical activity in the past 30 days, then select Yes.
Fruits
Yes
If you consume fruits one or more times in a day, then select Yes.
Veggies
Yes
If you consume vegetables one or more times in a day, then select Yes.


Diabetes Prediction
How many days during the past 30 days, your mental health was not good?
Physical Health care
How many days during the past 30 days, your physical health was not good?
Walking difficulty
No
If you face any difficulties while climbing stairs or walking, then select Yes.
Submit
Predicted Outcome: Great, Diabetes not present

7. ADVANTAGES & DISADVANTAGES

Advantages of a diabetes prediction model:

Early detection: A diabetes prediction model can help identify individuals at risk of developing diabetes at an early stage. Early detection enables healthcare professionals to implement preventive measures and lifestyle interventions, potentially reducing the progression of the disease and associated complications.

Risk stratification: By utilizing a diabetes prediction model, individuals can be stratified into different risk categories based on their predicted probability of developing diabetes. This allows healthcare providers to allocate resources and interventions more effectively, focusing on individuals at higher risk.

Personalized approach: A diabetes prediction model can provide personalized risk estimates, taking into account various factors specific to each individual. This personalized approach can guide healthcare professionals in tailoring interventions and recommendations to suit the needs of each individual, optimizing the management of diabetes risk.

Cost-effectiveness: Identifying individuals at high risk of developing diabetes through a prediction model can lead to cost-effective healthcare interventions. By targeting preventive measures to those at the greatest risk, resources can be allocated efficiently, potentially reducing the long-term burden of diabetes on healthcare systems.

Disadvantages of a diabetes prediction model:

Data quality and bias: The accuracy and reliability of a diabetes prediction model heavily depend on the quality and representativeness of the data used for training. Biases in the data, such as underrepresentation of certain demographics or missing variables, can lead to biased predictions and reduced generalizability.

Limited scope: A diabetes prediction model can provide insights into the likelihood of developing diabetes based on the available input variables. However, it may not capture all relevant risk factors or consider external factors that could impact disease development. The model's scope is limited to the variables and relationships included in the model.

Uncertainty and false predictions: Like any predictive model, a diabetes prediction model using statistical techniques such as logistic regression has inherent uncertainty. There is a chance of false-positive or false-negative predictions, where individuals are either incorrectly identified as high risk or low risk, respectively. This emphasizes the importance of clinical judgment and considering additional factors beyond the model's predictions.

Lack of causality: A diabetes prediction model can establish associations between risk factors and the likelihood of developing diabetes but cannot establish causality. While certain variables may be predictive, they may not be the direct cause of diabetes. Causal relationships require further investigation through experimental studies or other research designs.

It is important to note that these advantages and disadvantages are not exclusive to diabetes prediction models but are generally applicable to predictive modeling in healthcare. Careful consideration of these factors is essential in interpreting and utilizing the predictions generated by such models.

8. APPLICATIONS

Some key applications of a diabetes prediction project include:

Early intervention and preventive care: By accurately predicting the likelihood of an individual developing diabetes, healthcare professionals can intervene early with preventive measures. This may include lifestyle modifications, such as dietary changes, exercise programs, weight management, and medication when necessary. Early intervention can potentially delay or prevent the onset of diabetes and its associated complications.

Patient stratification and personalized treatment: A diabetes prediction model can help stratify individuals into different risk categories based on their predicted probability of developing diabetes. This stratification allows for personalized treatment approaches, with higher-risk individuals receiving more intensive monitoring and intervention, while lower-risk individuals may require less intensive interventions.

Resource allocation and healthcare planning: The use of a diabetes prediction model can aid in allocating healthcare resources more efficiently. By identifying individuals at higher risk of developing diabetes, healthcare providers can prioritize interventions and allocate resources accordingly. This can lead to more targeted screenings, educational programs, and healthcare services for those who need them the most.

Population health management: Diabetes prediction models can assist in population-level health management strategies. By analyzing the predicted probabilities of diabetes occurrence within a population, public health officials can identify high-risk groups and implement targeted interventions and public health campaigns. This can help in reducing the overall burden of diabetes on the population and improving long-term health outcomes.

Research and clinical trials: Diabetes prediction projects can contribute to research efforts by providing valuable insights into risk factors and their impact on disease development. Predictive models can be used to identify and recruit high-risk individuals for clinical trials and research studies, allowing for more targeted investigations and the development of new interventions or therapies.

Patient education and awareness: The use of a diabetes prediction model can facilitate patient education and raise awareness about the risk factors associated with diabetes. By informing individuals about their risk of developing diabetes, they can make informed decisions regarding lifestyle changes, regular check-ups, and proactive management of their health.

Overall, the applications of a diabetes prediction project are far-reaching and have the potential to positively impact patient care, healthcare resource allocation, public health initiatives, and research efforts in the field of diabetes prevention and management.

9. CONCLUSION

In conclusion, the diabetes prediction project utilizes logistic regression and other models(decision tree, random forest etc) which has provided valuable insights into predicting the occurrence of diabetes based on various input variables.

Through the implementation of logistic regression, we aimed to establish a relationship between independent variables such as age, body mass index (BMI), blood pressure, physical activity and other factors, and the dependent variable, which is the likelihood of developing diabetes.

After considering that certain factors such as education, income and others have no impact on our prediction, those are removed to increase accuracy
Our findings indicate that certain factors, such as increasing age, higher BMI, elevated blood pressure, and family history of diabetes, are positively associated with a higher risk of developing diabetes.

10. FUTURE SCOPE

The future scope of diabetes prediction encompasses a wide range of potential advancements and areas of development. Here are some key areas that hold promise for the future:

Integration of multi-omics data: With the advancement of technologies like genomics, transcriptomics, proteomics, and metabolomics, there is an opportunity to integrate multi-omics data into diabetes prediction models. By incorporating genetic variations, gene expression profiles, protein levels, and metabolic markers, the predictive accuracy and understanding of underlying biological mechanisms can be improved.

Artificial intelligence and machine learning: The application of more advanced machine learning algorithms, such as deep learning, neural networks, and ensemble methods, can enhance diabetes prediction models. These techniques have the potential to capture complex relationships, identify hidden patterns, and handle high-dimensional data, thereby improving the accuracy of predictions.

Wearable devices and continuous monitoring: The use of wearable devices and continuous monitoring technologies, such as glucose monitors, activity trackers, and other biosensors, can provide real-time data for diabetes prediction. Integrating these data sources with predictive models can enable personalized and dynamic risk assessment, facilitating early detection and timely interventions.

Personalized risk assessment: The future of diabetes prediction lies in developing personalized risk assessment models. By considering individual characteristics, such as genetic predisposition, lifestyle factors, and medical history, personalized logistic regression models can provide tailored risk estimates for each individual. This can enable more targeted and precise interventions, as well as personalized health recommendations.

Continuous monitoring and dynamic prediction: The incorporation of real-time or longitudinal data can enable continuous monitoring of diabetes risk and dynamic prediction. By updating the model with new information over time, healthcare professionals can track changes in an individual's risk profile and intervene promptly when necessary. This can lead to proactive management and timely interventions.

Integration with electronic health records (EHR): Integrating diabetes prediction models with electronic health records can facilitate seamless implementation and real-world application. By leveraging the data stored in EHR systems, including clinical measurements, laboratory results, and medical histories, the logistic regression model can be enriched with more comprehensive and up-to-date information, improving the accuracy of predictions.

11. BIBLIOGRAPHY

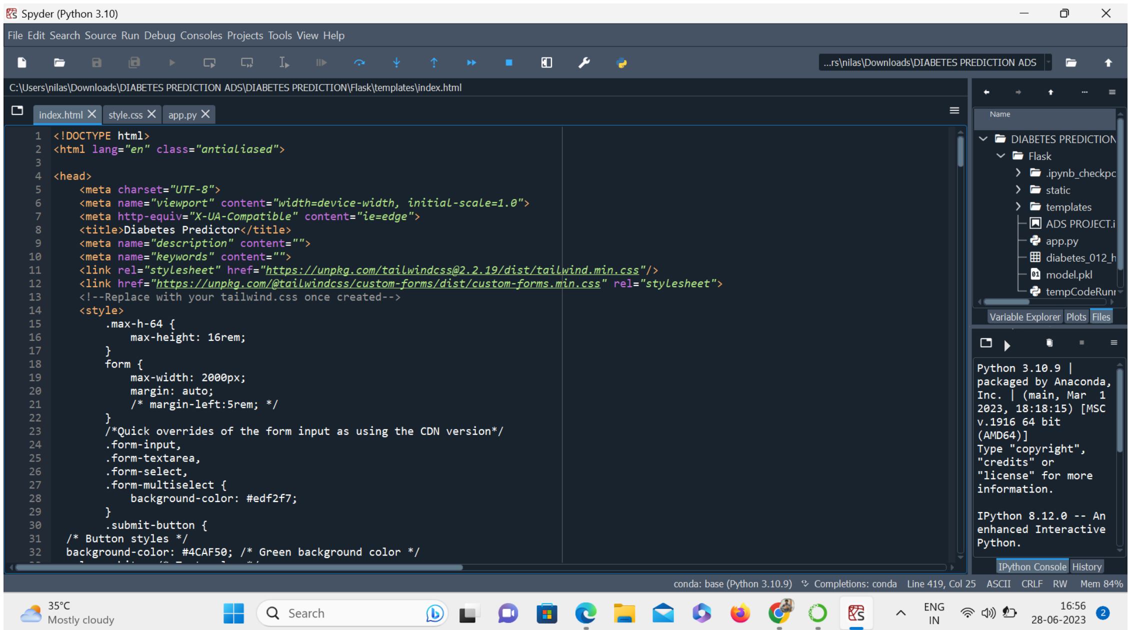
1. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
2. https://www.geeksforgeeks.org/videos/diabetes-prediction-in-machine-learning/
3. https://www.sciencedirect.com/science/article/pii/S1877050920300557
4. https://www.javatpoint.com/diabetes-prediction-using-machine-learning
5. https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/htl2.12039
6. https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full
7. https://www.javatpoint.com/logistic-regression-in-machine-learning

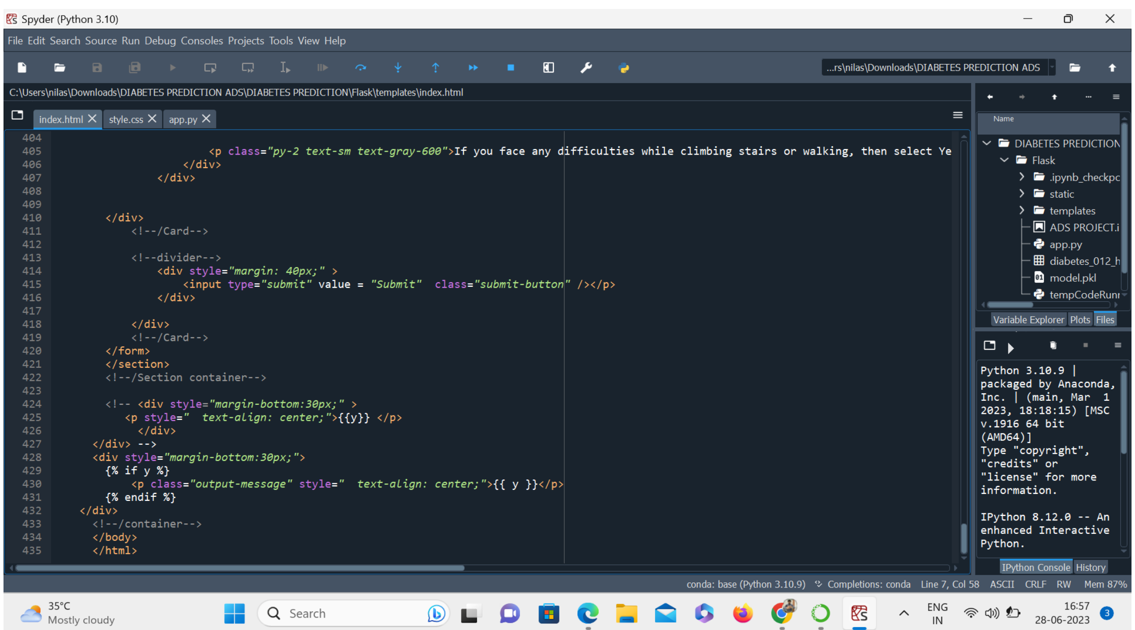
8. https://www.javatpoint.com/classification-algorithm-in-machine-learning
9. https://monkeylearn.com/blog/classification-algorithms/

APPENDIX:

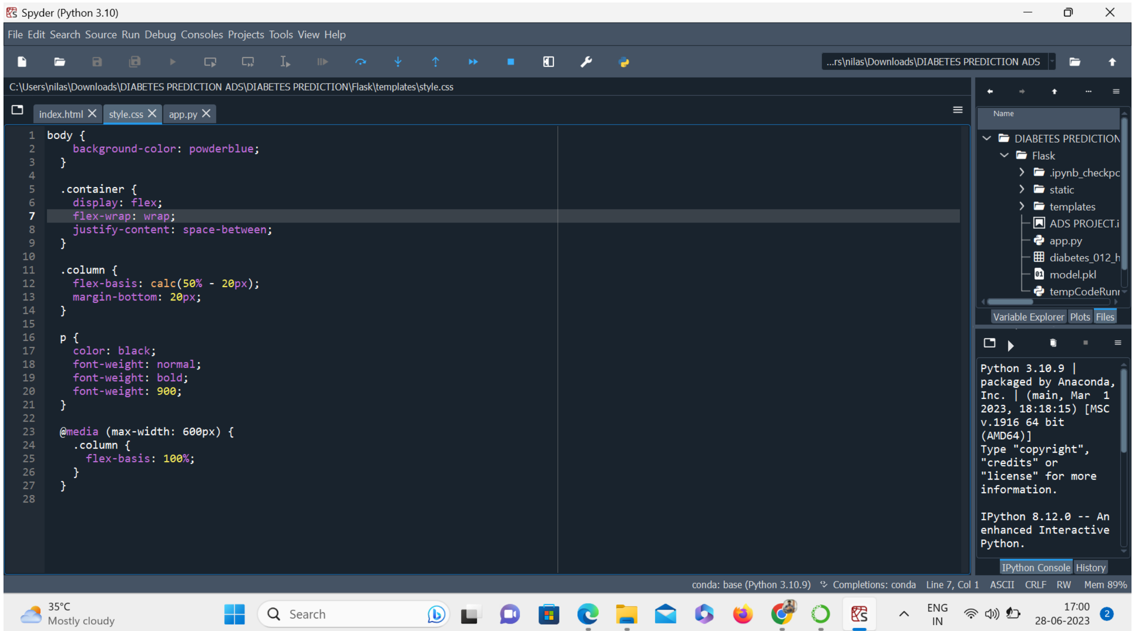
CODE:

HTML

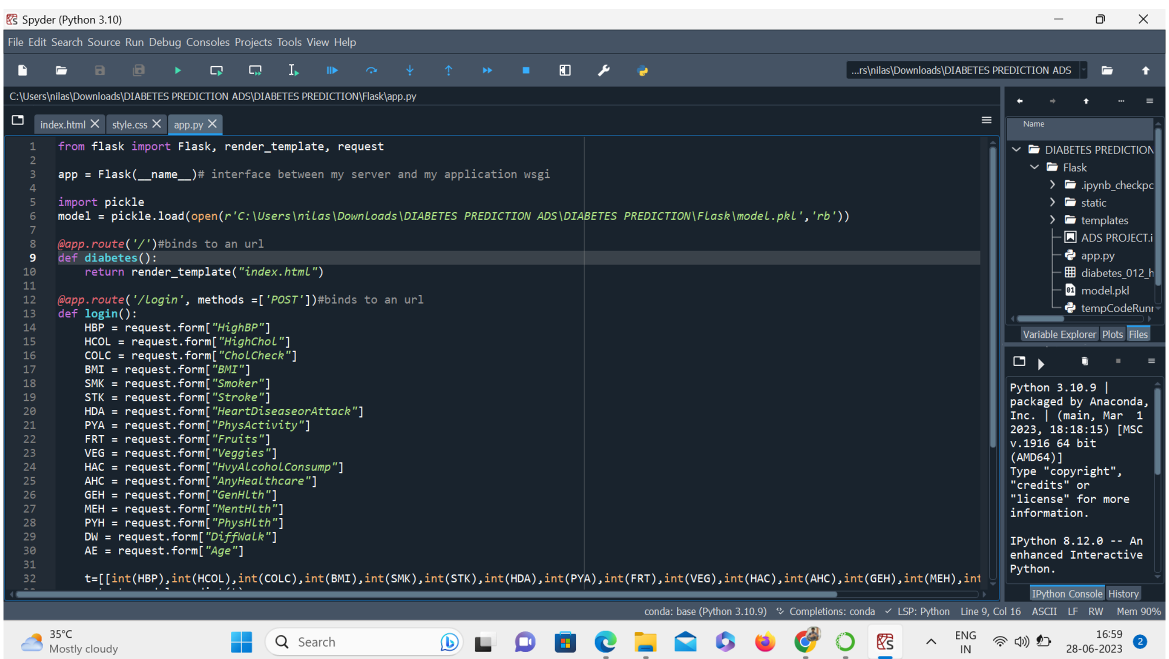


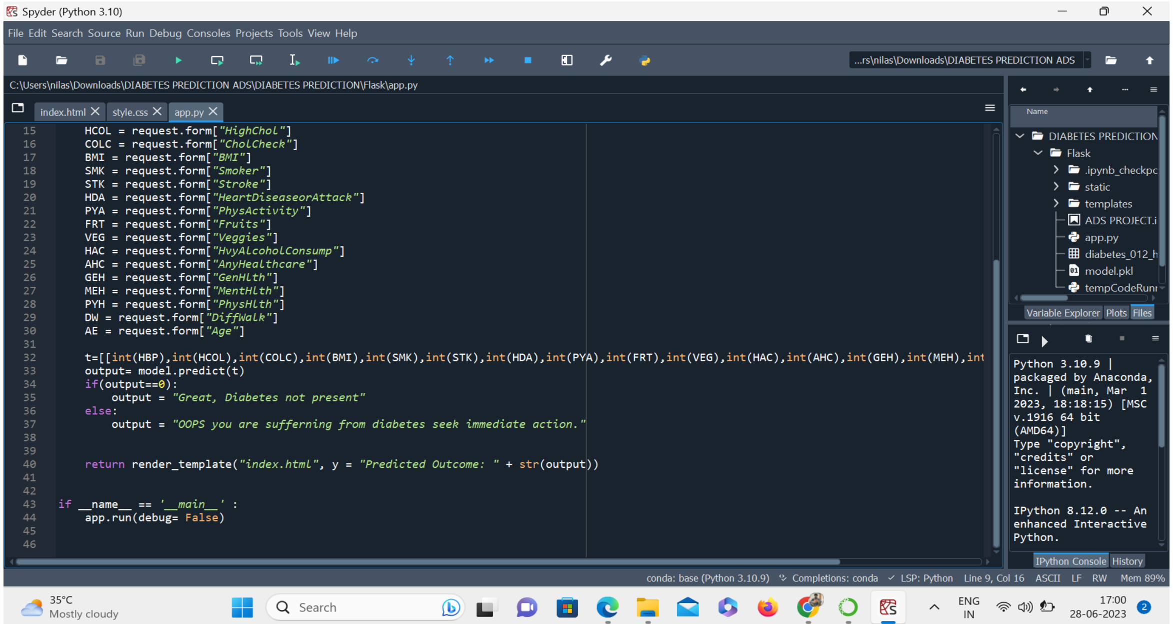


CSS:

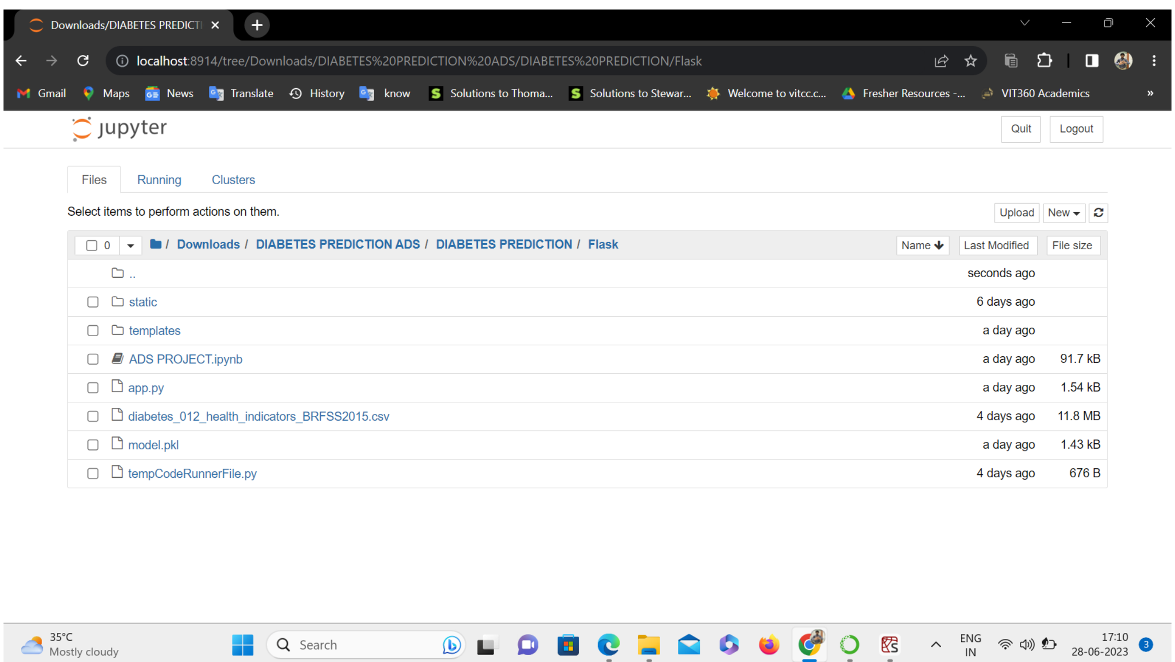

Spyder (Python 3.10)
C:\Users\nilas\Downloads\DIABETES PREDICTION ADS\DIABETES PREDICTION\Flask\templates\style.css
index.html
style.css
app.py
body {
  background-color: powderblue;
 }
 .container {
  display: flex;
  flex-wrap: wrap;
  justify-content: space-between;
 }
 .column {
  flex-basis: calc(50% - 20px);
  margin-bottom: 20px;
 }
 p {
  color: black;
  font-weight: normal;
  font-weight: bold;
  font-weight: 900;
 }
 @media (max-width: 600px) {
  .column {
   flex-basis: 100%;
  }
 }
Python 3.10.9 | packaged by Anaconda, Inc. | (main, Mar 1 2023, 18:18:15) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.
IPython 8.12.0 -- An enhanced Interactive Python.

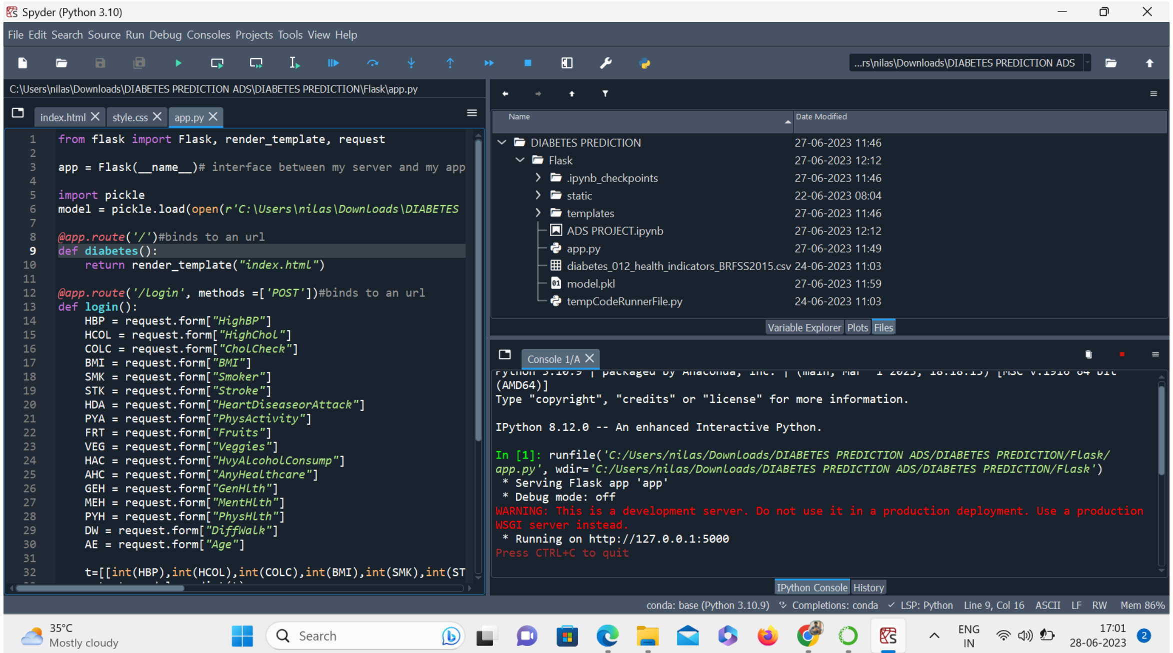
FLASK:


Spyder (Python 3.10)
C:\Users\nilas\Downloads\DIABETES PREDICTION ADS\DIABETES PREDICTION\Flask\app.py
index.html
style.css
app.py
from flask import Flask, render_template, request
app = Flask(__name__)# interface between my server and my application wsgi
import pickle
model = pickle.load(open(r'C:\Users\nilas\Downloads\DIABETES PREDICTION ADS\DIABETES PREDICTION\Flask\model.pkl','rb'))
@app.route('/')#binds to an url
def diabetes():
    return render_template("index.html")
@app.route('/login', methods =['POST'])#binds to an url
def login():
    HBP = request.form["HighBP"]
    HCOL = request.form["HighChol"]
    COLC = request.form["CholCheck"]
    BMI = request.form["BMI"]
    SMK = request.form["Smoker"]
    STK = request.form["Stroke"]
    HDA = request.form["HeartDiseaseorAttack"]
    PYA = request.form["PhysActivity"]
    FRT = request.form["Fruits"]
    VEG = request.form["Veggies"]
    HAC = request.form["HvyAlcoholConsump"]
    AHC = request.form["AnyHealthcare"]
    GEH = request.form["GenHlth"]
    MEH = request.form["MentHlth"]
    PYH = request.form["PhysHlth"]
    DW = request.form["DiffWalk"]
    AE = request.form["Age"]
    t=[[int(HBP),int(HCOL),int(COLC),int(BMI),int(SMK),int(STK),int(HDA),int(PYA),int(FRT),int(VEG),int(HAC),int(AHC),int(GEH),int(MEH),int
Python 3.10.9 | packaged by Anaconda, Inc. | (main, Mar 1 2023, 18:18:15) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.
IPython 8.12.0 -- An enhanced Interactive Python.



FLASK CODE RUNNING:





Spyder (Python 3.10)
File Edit Search Source Run Debug Consoles Projects Tools View Help
...rs\nilas\Downloads\DIABETES PREDICTION ADS
C:\Users\nilas\Downloads\DIABETES PREDICTION ADS\DIABETES PREDICTION\Flask\app.py
index.html
style.css
app.py
from flask import Flask, render_template, request
app = Flask(__name__)# interface between my server and my app
import pickle
model = pickle.load(open(r'C:\Users\nilas\Downloads\DIABETES
@app.route('/')#binds to an url
def diabetes():
return render_template("index.html")
@app.route('/login', methods =['POST'])#binds to an url
def login():
HBP = request.form["HighBP"]
HCOL = request.form["HighChol"]
COLC = request.form["CholCheck"]
BMI = request.form["BMI"]
SMK = request.form["Smoker"]
STK = request.form["Stroke"]
HDA = request.form["HeartDiseaseorAttack"]
PYA = request.form["PhysActivity"]
FRT = request.form["Fruits"]
VEG = request.form["Veggies"]
HAC = request.form["HvyAlcoholConsump"]
AHC = request.form["AnyHealthcare"]
GEH = request.form["GenHlth"]
MEH = request.form["MentHlth"]
PYH = request.form["PhysHlth"]
DW = request.form["DiffWalk"]
AE = request.form["Age"]
t=[[int(HBP),int(HCOL),int(COLC),int(BMI),int(SMK),int(ST
Name
Date Modified
DIABETES PREDICTION 27-06-2023 11:46
Flask 27-06-2023 12:12
.ipynb_checkpoints 27-06-2023 11:46
static 22-06-2023 08:04
templates 27-06-2023 11:46
ADS PROJECT.ipynb 27-06-2023 12:12
app.py 27-06-2023 11:49
diabetes_012_health_indicators_BRFSS2015.csv 24-06-2023 11:03
model.pkl 27-06-2023 11:59
tempCodeRunnerFile.py 24-06-2023 11:03
Variable Explorer Plots Files
Console 1/A
(AMD64)]
Type "copyright", "credits" or "license" for more information.
IPython 8.12.0 -- An enhanced Interactive Python.
In [1]: runfile('C:/Users/nilas/Downloads/DIABETES PREDICTION ADS/DIABETES PREDICTION/Flask/app.py', wdir='C:/Users/nilas/Downloads/DIABETES PREDICTION ADS/DIABETES PREDICTION/Flask')
* Serving Flask app 'app'
* Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
IPython Console History
conda: base (Python 3.10.9) Completions: conda LSP: Python Line 9, Col 16 ASCII LF RW Mem 86%
35°C
Mostly cloudy
Search
ENG
IN
17:01
28-06-2023