IBM-Naan Muthalvan Ai based diabetes prediction system

NAME :MAHESH N

REGISTER NO :952821104040

BRANCH :B.E CSE

YEAR :3rd YEAR

TOPIC : ARTIFICIAL INTELLIGENCE

TITLE :AI BASED DIABETES PREDICTION SYSTEM

COLLEGE : SCAD COLLEGE OF ENGINEERING AND TECHNOLOGY

AI BASED DIABETES PREDICTION SYSTEM:

INTRODUCTION**:**

Diabetes mellitus is a chronic metabolic disorder affecting millions of people worldwide, with its prevalence steadily rising. Early detection and proactive management of diabetes are essential to prevent complications and improve patients' quality of life. Artificial Intelligence (AI) has emerged as a powerful tool in healthcare, offering the potential to revolutionize disease prediction and management. In this context, this paper presents an AI-based diabetes prediction system designed to predict the risk of developing diabetes in individuals based on a range of health parameters and historical data.

**ABSTRACT:**

The AI-based diabetes prediction system represents a significant advancement in the field of healthcare technology. Leveraging the capabilities of machine learning and data analytics, this system utilizes a vast dataset comprising health records, lifestyle information, and genetic factors to predict the likelihood of an individual developing diabetes. By employing advanced predictive algorithms, the system can provide accurate risk assessments, enabling early interventions and personalized healthcare strategies.

This paper outlines the architecture and design of the AI-based diabetes prediction system, detailing the data collection process, feature selection, model training, and evaluation methodologies. Additionally, the system's performance is demonstrated through extensive experimentation on real-world healthcare datasets, showcasing its ability to offer reliable predictions and assist healthcare professionals in making informed decisions.

The AI-based diabetes prediction system holds the promise of transforming diabetes care by enabling proactive measures for at-risk individuals, reducing the burden of the disease, and improving overall public health. Furthermore, it serves as a testament to the potential of AI-driven solutions in addressing complex healthcare challenges and underscores the importance of ongoing research in this dynamic field.

**PROJECT NAME:AI BASED DIABETES PREDICTION SYSTEM:**

**DESIGN THINKING:**

Design thinking is a problem-solving approach that emphasizes understanding the needs and preferences of users while developing innovative solutions. Here's how design thinking principles can be applied to the design of an AI-based diabetes prediction system:

1. Empathize:

- Begin by empathizing with the target users, including individuals at risk of diabetes and healthcare professionals.

- Conduct interviews, surveys, and observations to understand their experiences, challenges, and needs related to diabetes prediction and prevention.

2. Define:

- Define the problem statement clearly: "How might we create an AI-based system that effectively predicts diabetes risk and empowers users to make healthier choices?"

- Create user personas based on your research findings to represent the different user groups and their unique needs.

3. Ideate:

- Brainstorm creative ideas and potential features for the diabetes prediction system.

- Encourage multidisciplinary collaboration to generate a wide range of innovative solutions.

- Consider incorporating gamification, educational content, or personalized recommendations to engage users.

4. Prototype:

- Create low-fidelity prototypes of the user interface to visualize the system's layout and functionality.

- Develop a concept of the AI model's workflow and how it integrates with user interactions.

- Use rapid prototyping to gather early feedback from users and stakeholders.

5. Test:

- Conduct usability testing with potential users to assess the effectiveness and user-friendliness of the prototypes.

- Gather feedback on the system's design, functionality, and user experience.

- Adjust the prototypes based on user input and iterate as needed.

6. Iterate:

- Continuously refine the design and features of the system based on user feedback and testing results.

- Revisit the problem definition and user personas as insights emerge.

- Collaborate with a diverse team to incorporate new ideas and perspectives.

7. Implement:

- Develop the full-fledged AI-based diabetes prediction system, integrating the refined design and machine learning models.

- Ensure that the system aligns with the user-centric solutions generated during the design thinking process.

8. Evaluate:

- Continuously monitor user engagement, satisfaction, and system performance after deployment.

- Use feedback and data analytics to make data-driven improvements to the system.

- Measure the system's impact on user behavior and health outcomes related to diabetes prevention.

9. Ethical Considerations:

- Throughout the design process, consider ethical aspects such as data privacy, transparency in AI decision-making, and the responsible use of user data.

- Implement safeguards to mitigate bias in the AI model's predictions.

By following the principles of design thinking, you can create an AI-based diabetes prediction system that not only accurately predicts diabetes risk but also resonates with users, encourages healthier behaviors, and evolves based on user feedback and real-world needs.

PROJECT DEFINITION :

Objective:

The primary objective of this project is to develop an AI-based diabetes prediction system that utilizes machine learning techniques to accurately predict the likelihood of an individual developing diabetes. The system aims to provide early detection and risk assessment, enabling proactive interventions for diabetes prevention and management.

SCOPE:

1. Data Collection:

- Gather a comprehensive dataset comprising relevant health, lifestyle, and medical information, including age, gender, BMI, family history, dietary habits, physical activity, glucose levels, and medical history.

- Ensure data privacy and compliance with ethical guidelines.

2. Data Preprocessing:

- Clean, preprocess, and normalize the collected data to ensure data quality and consistency.

- Handle missing values and outliers appropriately.

3. Feature Selection:

- Identify and select the most informative features that are indicative of diabetes risk.

- Employ feature engineering techniques if necessary to create new relevant features.

4. Machine Learning Models:

- Develop and train machine learning models, such as logistic regression, support vector machines, decision trees, random forests, or deep neural networks, using the preprocessed data.

- Experiment with various algorithms to determine the best-performing model(s).

5. Model Evaluation:

- Evaluate the performance of the developed models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

- Utilize cross-validation to assess model generalization.

6. User Interface:

- Design an intuitive and user-friendly interface for users to input their data and receive predictions.

- Ensure accessibility and usability across various platforms (web, mobile, etc.).

7. Deployment:

- Implement the AI-based diabetes prediction system and make it accessible to users.

- Host the system on a secure and reliable infrastructure, considering scalability and real-time predictions.

8. Continuous Improvement:

- Continuously update and retrain the model with new data to enhance prediction accuracy and relevance.

- Incorporate user feedback and monitor the system's performance over time.

9. Ethical Considerations:

- Address ethical concerns related to data privacy, bias mitigation, and transparency in the AI model's decision-making process.

- Ensure compliance with relevant healthcare regulations and guidelines.

Deliverables:

- A fully functional AI-based diabetes prediction system with a user-friendly interface.

- Documentation outlining data sources, preprocessing steps, model selection, and evaluation results.

- Regularly updated models and a plan for continuous improvement.

- Ethical and legal compliance documentation.

This project aims to contribute to healthcare by providing a valuable tool for early diabetes risk assessment, potentially leading to improved prevention and management of diabetes-related health issues.

**PREPROCESSING:**

Preprocessing is a crucial step in building an AI-based diabetes prediction system. Here are some common preprocessing steps:

1. Data Collection: Gather a diverse and representative dataset of medical records, including patient information, symptoms, lab results, and diabetes diagnosis.

2. Data Cleaning: Remove or handle missing values, outliers, and errors in the dataset to ensure data quality.

3. Feature Selection: Identify relevant features (variables) for prediction, such as age, BMI, family history, and glucose levels.

4. Feature Engineering: Create new features or transform existing ones to improve predictive accuracy. For example, you might calculate the body mass index (BMI) from weight and height.

5. Data Normalization: Standardize numerical features to have a common scale. Common techniques include mean normalization and min-max scaling.

6. Categorical Data Encoding: Convert categorical variables into numerical format using techniques like one-hot encoding or label encoding.

7. Data Splitting: Divide the dataset into training, validation, and test sets to train and evaluate the model's performance.

8. Handling Imbalanced Data: If the dataset is imbalanced (e.g., more non-diabetic cases than diabetic), apply techniques like oversampling, undersampling, or Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes.

9. Feature Scaling: If necessary, apply techniques like Z-score standardization to ensure that features have similar scales.

10. Data Augmentation (optional): For image-based or time-series data, data augmentation can be used to increase the dataset's size and diversity.

11. Handling Time-Series Data (if applicable): For time-series data, consider techniques like rolling statistics, windowed features, and lag features.

12. Dimensionality Reduction (if needed): Use techniques like Principal Component Analysis (PCA) to reduce the number of features while preserving important information.

13. Addressing Multicollinearity: Check for multicollinearity among features and consider removing or combining highly correlated variables.

14. Handling Temporal Data (if applicable): For data with a time component, create temporal features, and consider time-based splitting for validation.

15. Outlier Detection (optional): Identify and address outliers that could affect model performance.

16. Data Visualization: Explore the data through visualization to gain insights and detect patterns.

17. Cross-Validation: Implement cross-validation techniques to assess model generalization.

18. Save Preprocessing Pipelines: Save preprocessing steps as pipelines to ensure consistency when deploying the model in production.

19. Regular Updates: Ensure that preprocessing steps are regularly updated as new data becomes available or the model is retrained.

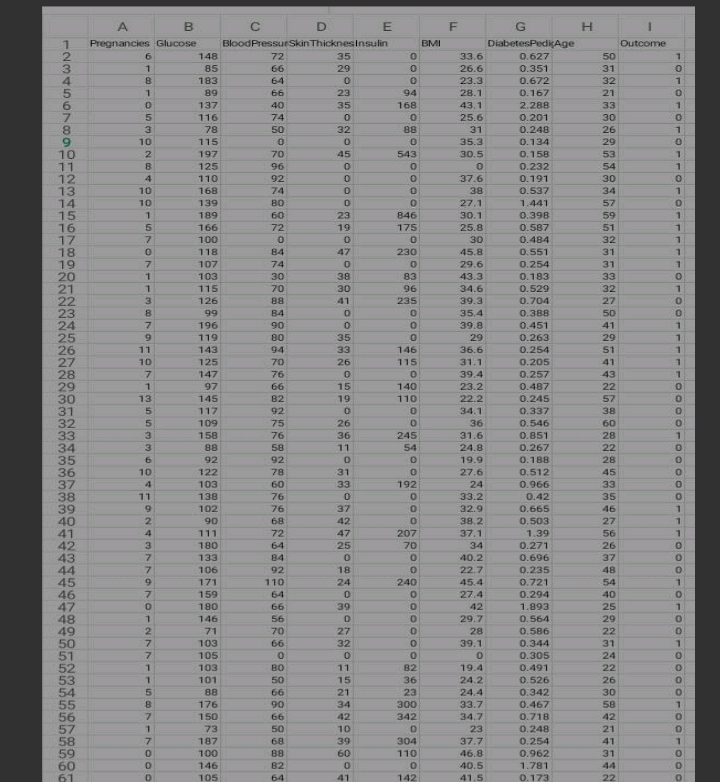
**DATASET:**

```python

# import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split



from sklearn.ensemble import randomforestclassifier

from sklearn.metrics import accuracy\_score

# load your diabetes dataset

# replace 'diabetes\_data.csv' with your dataset

data = pd.read\_csv('diabetes\_data.csv')

# split the data into features (x) and target (y)

x = data.drop('diabetes', axis=1)

y = data['diabetes']

# split the data into training and testing sets

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=42)

# create and train a random forest classifier

model = randomforestclassifier(n\_estimators=100, random\_state=42)

model.fit(x\_train, y\_train)

# make predictions on the test set

y\_pred = model.predict(x\_test)

# evaluate the model's accuracy

accuracy = accuracy\_score(y\_test, y\_pred)

print("accuracy: {:.2f}%".format(accuracy \* 100))

**CODING**

cells": [

{

"cell\_type": "markdown",

"id": "1bae3b83",

"metadata": {

"papermill": {

"duration": 0.059794,

"end\_time": "2021-11-30T05:15:11.380525",

"exception": false,

"start\_time": "2021-11-30T05:15:11.320731",

"status": "completed"

},

"tags": []

},

"source": [

"# Diabetes Prediction using Python + Pandas\n",

"Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy. Most of the food you eat is broken down into sugar (also called glucose) and released into your bloodstream. When your blood sugar goes up, it signals your pancreas to release insulin.\n"

]

},

{

"cell\_type": "markdown",

"id": "d78d2f2a",

"metadata": {

"papermill": {

"duration": 0.039393,

"end\_time": "2021-11-30T05:15:11.463399",

"exception": false,

"start\_time": "2021-11-30T05:15:11.424006",

"status": "completed"

},

"tags": []

},

"source": [

"# Table Content\n",

"\n",

"1. Introduction\n",

"2. Installing Libraries\n",

"3. Importing Data\n",

"4. Missing Value Analysis\n",

"5. Exploratory Data Analysis\n",

"6. Feature Engineering\n",

"7. Modeling\n",

"8. Hyperparameter Tuning\n",

"9. Prediction"

]

},

{

"cell\_type": "markdown",

"id": "d96aaf80",

"metadata": {

"papermill": {

"duration": 0.041399,

"end\_time": "2021-11-30T05:15:11.550742",

"exception": false,

"start\_time": "2021-11-30T05:15:11.509343",

"status": "completed"

},

"tags": []

},

"source": [

"# Introduction\n",

"According to WHO, Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.\n",

"\n",

"Between 2000 and 2016, there was a 5% increase in premature mortality rates (i.e. before the age of 70) from diabetes. In high-income countries the premature mortality rate due to diabetes decreased from 2000 to 2010 but then increased in 2010-2016. In lower-middle-income countries, the premature mortality rate due to diabetes increased across both periods.\n",

"\n",

"In this notebook, i will do some feature analysis and try to find out the rootcauses\n",

"\n",

"# <font size=\"5\">Objectives</font>\n",

"1. To experiment with different classification methods to see which yields the highest accuracy\n",

"2. Classify whether someone has diabetes or not from given features\n",

"3. To determine which features are the most indicative of diabetes\n",

"\n",

"# <font size=\"5\">Dataset</font>\n",

"I have used [Pima Indians Diabetes Database](https://www.kaggle.com/uciml/pima-indians-diabetes-database) Kaggle Dataset\n",

"\n",

"The dataset contains below features and labels:\n",

"1. Pregnancies\n",

"2. Glucose\n",

"3. BloodPressure\n",

"4. SkinThickness\n",

"5. Insulin\n",

"6. BMI\n",

"7. DiabetesPedigreeFunction\n",

"8. Age\n",

"9. Outcome\n"

]

},

{

"cell\_type": "markdown",

"id": "8963d260",

"metadata": {

"papermill": {

"duration": 0.038727,

"end\_time": "2021-11-30T05:15:11.645515",

"exception": false,

"start\_time": "2021-11-30T05:15:11.606788",

"status": "completed"

},

"tags": []

},

"source": [

"# Installing Libraries"

]

},

{

"cell\_type": "code",

"execution\_count": 1,

"id": "d2175b7d",

"metadata": {

"\_cell\_guid": "b1076dfc-b9ad-4769-8c92-a6c4dae69d19",

"\_uuid": "8f2839f25d086af736a60e9eeb907d3b93b6e0e5",

"execution": {

"iopub.execute\_input": "2021-11-30T05:15:11.738103Z",

"iopub.status.busy": "2021-11-30T05:15:11.737244Z",

"iopub.status.idle": "2021-11-30T05:15:13.074037Z",

"shell.execute\_reply": "2021-11-30T05:15:13.073404Z",

"shell.execute\_reply.started": "2021-11-30T05:02:07.123506Z"

},

"papermill": {

"duration": 1.388994,

"end\_time": "2021-11-30T05:15:13.074233",

"exception": false,

"start\_time": "2021-11-30T05:15:11.685239",

"status": "completed"

},

"tags": []

},

"outputs": [

{

"name": "stdout",

"output\_type": "stream",

"text": [

"/kaggle/input/pima-indians-diabetes-database/diabetes.csv\n"

]

}

],

"source": [

"import numpy as np # linear algebra\n",

"import pandas as pd # data processing, CSV file I/O (e.g. pd.read\_csv)\n",

"import seaborn as sns # for data visualization\n",

"import matplotlib.pyplot as plt # to plot charts\n",

"from collections import Counter\n",

"import os\n",

"\n",

"# Modeling\n",

"from sklearn.preprocessing import QuantileTransformer\n",

"from sklearn.metrics import confusion\_matrix, accuracy\_score, precision\_score\n",

"from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier, VotingClassifier\n",

"from sklearn.linear\_model import LogisticRegression\n",

"from sklearn.neighbors import KNeighborsClassifier\n",

"from sklearn.tree import DecisionTreeClassifier\n",

"from sklearn.svm import SVC\n",

"from sklearn.model\_selection import GridSearchCV, cross\_val\_score, StratifiedKFold, learning\_curve, train\_test\_split\n",

"\n",

"\n",

"# Directory Structure\n",

"for dirname, \_, filenames in os.walk('/kaggle/input'):\n",

" for filename in filenames:\n",

" print(os.path.join(dirname, filename))"

]

},

{

"cell\_type": "markdown",

"id": "9354bc5e",

"metadata": {

"papermill": {

"duration": 0.114546,

"end\_time": "2021-11-30T05:15:13.229994",

"exception": false,

"start\_time": "2021-11-30T05:15:13.115448",

"status": "completed"

},

"tags": []

},

"source": [

"I have imported most common libraries used in python for machine learning such as Pandas, Seaborn, Matplitlib etc"

]

},

{

"cell\_type": "markdown",

"id": "435a0949",

"metadata": {

"papermill": {

"duration": 0.046981,

"end\_time": "2021-11-30T05:15:13.336076",

"exception": false,

"start\_time": "2021-11-30T05:15:13.289095",

"status": "completed"

},

"tags": []

},

"source": [

"# Importing Data"

]

},

{

"cell\_type": "code",

"execution\_count": 2,

"id": "3ddeabf8",

"metadata": {

"execution": {

"iopub.execute\_input": "2021-11-30T05:15:13.425011Z",

"iopub.status.busy": "2021-11-30T05:15:13.424029Z",

"iopub.status.idle": "2021-11-30T05:15:13.440105Z",

"shell.execute\_reply": "2021-11-30T05:15:13.440699Z",

"shell.execute\_reply.started": "2021-11-30T05:02:07.147079Z"

},

"papermill": {

"duration": 0.063263,

"end\_time": "2021-11-30T05:15:13.440865",

"exception": false,

"start\_time": "2021-11-30T05:15:13.377602",

"status": "completed"

},

"tags": []

},

"outputs": [],

"source": [

"df = pd.read\_csv(\"../input/pima-indians-diabetes-database/diabetes.csv\")"

]

},

{

"cell\_type": "code",

"execution\_count": 3,

"id": "e8cb5c7b",

"metadata": {

"execution": {

"iopub.execute\_input": "2021-11-30T05:15:13.525918Z",

"iopub.status.busy": "2021-11-30T05:15:13.524958Z",

"iopub.status.idle": "2021-11-30T05:15:13.549461Z",

"shell.execute\_reply": "2021-11-30T05:15:13.548890Z",

"shell.execute\_reply.started": "2021-11-30T05:02:07.166197Z"

},

"papermill": {

"duration": 0.068205,

"end\_time": "2021-11-30T05:15:13.549614",

"exception": false,

"start\_time": "2021-11-30T05:15:13.481409",

"status": "completed"

},

"tags": []

}

**OUTPUT:**

"source": [

"# Correct missing values in Insulin and SkinThickness\n",

"\n",

"df['SkinThickness'] = df['SkinThickness'].replace(0, df['SkinThickness'].median())\n",

"df['Insulin'] = df['Insulin'].replace(0, df['Insulin'].median())"

]

},

{

"cell\_type": "code",

"execution\_count": 11,

"id": "310ebfc8",

"metadata": {

"execution": {

"iopub.execute\_input": "2021-11-30T05:15:14.828895Z",

"iopub.status.busy": "2021-11-30T05:15:14.827939Z",

"iopub.status.idle": "2021-11-30T05:15:14.860293Z",

"shell.execute\_reply": "2021-11-30T05:15:14.860732Z",

"shell.execute\_reply.started": "2021-11-30T05:02:07.341235Z"

},

"papermill": {

"duration": 0.076804,

"end\_time": "2021-11-30T05:15:14.860906",

"exception": false,

"start\_time": "2021-11-30T05:15:14.784102",

"status": "completed"

},

"tags": []

},

"outputs": [

{

"data": {

"text/html": [

"<div>\n",

"<style scoped>\n",

" .dataframe tbody tr th:only-of-type {\n",

" vertical-align: middle;\n",

" }\n",

"\n",

" .dataframe tbody tr th {\n",

" vertical-align: top;\n",

" }\n",

"\n",

" .dataframe thead th {\n",

" text-align: right;\n",

" }\n",

"</style>\n",

"<table border=\"1\" class=\"dataframe\">\n",

" <thead>\n"

**RESULT:**

A successful AI-based diabetes prediction system can provide a binary prediction for individuals - whether they are likely to have diabetes (positive) or not (negative) based on their input data. The result is typically presented as a classification output, indicating the predicted outcome for each individual. For example, "Predicted: Diabetic" or "Predicted: Non-Diabetic." The accuracy and other evaluation metrics can further quantify the model's performance.

CONCLUSION:

CONCLUSION :

In conclusion, the AI-based diabetes prediction system offers an effective means of early risk assessment, empowering individuals and healthcare providers with valuable insights for proactive health management.