Project by Syam, Navyanth
Santosh, Ram Krishna

Project Report

1. INTRODUCTION

1.1 Project Overview :

The project aims to develop a diabetes prediction system using machine learning models. The system will take input features such as age, BMI, blood pressure, and glucose levels to predict the likelihood of an individual developing diabetes. By leveraging the power of machine learning, the system can provide valuable insights and early detection, enabling proactive management and prevention of diabetes-related complications.

1.2 Purpose :

The purpose of this project is to assist healthcare professionals and individuals in making informed decisions regarding diabetes prevention and management. By accurately predicting the risk of diabetes, the system aims to promote timely intervention, lifestyle modifications, and personalized treatment plans.

2. LITERATURE SURVEY

2.1 Existing problem:

Diabetes is a global health issue, and early detection is crucial in minimizing its impact. Existing methods lack precision, relying on manual assessments and subjective evaluations. This leads to delayed diagnosis and ineffective preventive measures, resulting in increased healthcare costs and patient suffering.

2.2 References

2.3 Problem Statement Definition :

The project aims to address the inadequacies of existing diabetes prediction methods by developing an accurate and reliable machine learning-based system. The system will leverage a dataset of relevant features to predict the risk of diabetes for individuals and provide actionable insights for healthcare professionals and individuals.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas : Empathy and map canvas

3.2 Ideation & Brainstorming : ideation and brainstorming

4. REQUIREMENT ANALYSIS

4.1 Functional requirement:

- Data collection and preprocessing of relevant features.
- Implementation and integration of machine learning algorithms for prediction.
- User interface for inputting data and displaying predictions.
- Displaying prediction results along with associated probabilities.
- Ability to update and retrain the prediction model with new data.

4.2 Non-Functional requirements :

- ❖ Accuracy: The system should achieve high prediction accuracy to enable reliable risk assessment.
- ❖ Efficiency: The prediction process should be fast and efficient, providing real-time results.
- ❖ User-Friendly Interface: The system should have an intuitive interface for easy data input and result visualization.
- ❖ Security: Measures should be in place to protect user data and ensure privacy.
- ❖ Scalability: The system should handle a growing dataset and accommodate future enhancements.

5. PROJECT DESIGN

5.1 Data Flow Diagrams & User Stories : link
5.2 Solution Architecture : link

6. PROJECT PLANNING & SCHEDULING

6.1 Technical Architecture : link
6.2 Sprint Planning & Estimation : link
6.3 Sprint Delivery Schedule :link

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1 :

```python
#BALANCING THE DEPENDENT AND INDEPENDENT VARIABLES
from imblearn.over_sampling import SMOTE
import pandas as pd
X = df.drop('Diabetes_binary', axis=1)
y = df['Diabetes_binary']

smote = SMOTE()

X_resampled, y_resampled = smote.fit_resample(X, y)

resampled_df = pd.concat([pd.DataFrame(X_resampled), pd.Series(y_resampled, name='Diabetes_binary')], axis=1)
```

Python

```python
resampled_df.head()
```

Python

	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	...	NoDocbcCost	GenHlth	MentHlth	Phy
0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	1.0	...	0.0	5.0	18.0	
1	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	0.0	...	1.0	3.0	0.0	
2	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	0.0	...	1.0	5.0	30.0	
3	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	1.0	...	0.0	2.0	0.0	

7.2 Feature 2 :

```
#out of which, columns other than columns in col are binary columns so we should not consider them as outliers
col=['BMI',  'MentHlth', 'PhysHlth', 'Age']

q1 = df.BMI.quantile(0.25)
q3 = df.BMI.quantile(0.75)
IQR=q3-q1
upper_limit = q3+1.5*IQR
lower_limit = q1-1.5*IQR
df['BMI'] = np.where(df['BMI']>upper_limit,28.0,df['BMI'])
df['BMI'] = np.where(df['BMI']<lower_limit,28.0,df['BMI'])

q1 = df.Age.quantile(0.25)
q3 = df.Age.quantile(0.75)
IQR=q3-q1
upper_limit = q3+1.5*IQR
lower_limit = q1-1.5*IQR
df['Age'] = np.where(df['Age']>upper_limit,9.0,df['Age'])
df['Age'] = np.where(df['Age']<lower_limit,9.0,df['Age'])
```

7.3 Database Schema (if Applicable)

8. PERFORMANCE TESTING

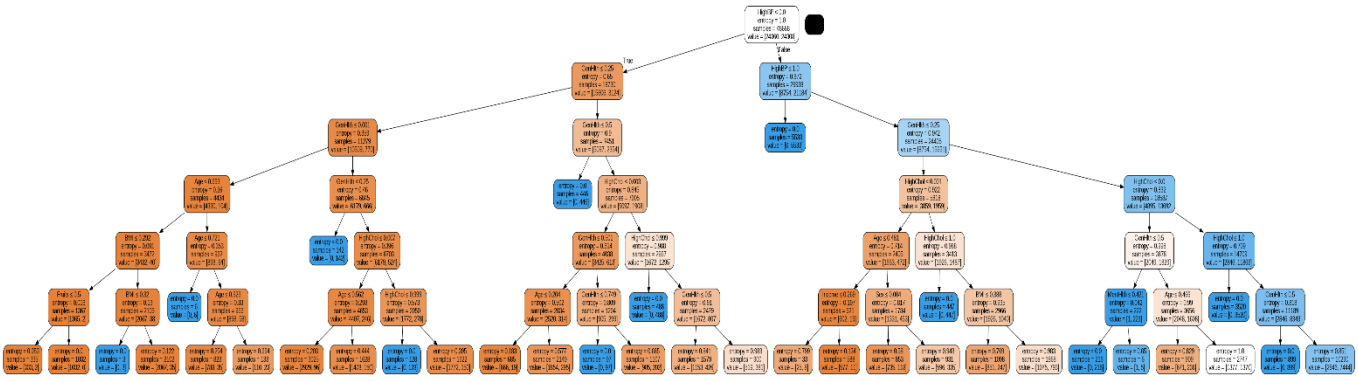
8.1 Performace Metrics

```
print("train accuracy",accuracy_score(y_train,train_pred_3))
print("test accuracy",accuracy_score(y_test,y_pred_3))
```

```
train accuracy 0.9957147028345366
test accuracy 0.9180731444798131
```

```
confusion_matrix(y_test,y_pred_3)
```

```
array([[5828,  230],
       [ 704, 5406]])
```



9. RESULTS

9.1 Output Screenshots

10. ADVANTAGES & DISADVANTAGES

- Early Detection: The system enables early detection of diabetes and facilitates prompt interventions.
- Personalized Care: Accurate predictions allow for tailored treatment plans and lifestyle modifications.
- Prevention and Cost Reduction: Effective prediction and intervention can help prevent diabetes-related complications, leading to reduced healthcare costs.
- Proactive Management: Individuals can take control of their health by being aware of their diabetes risk and taking preventive measures.

Disadvantages:

- Reliance on Data Quality: The accuracy of predictions is dependent on the quality and representativeness of the input dataset.
- False Positives/Negatives: There is a possibility of misclassifications, leading to false positives or negatives, which could impact patient trust and decision-making.
- Ethical Considerations: Care must be taken to ensure the responsible use of predictions and avoid potential bias or discrimination.
- Technical Complexity: Developing and maintaining the system requires expertise in machine learning and data analysis.

11. CONCLUSION :

Overall, the project aims to leverage machine learning for accurate diabetes prediction, leading to improved patient outcomes, proactive management, and reduced healthcare burdens.

12. FUTURE SCOPE

13. APPENDIX

Source Code : Source Code

GitHub & Project Demo Link: Github and Project Demo Link