**DIABETES PREDICTION**

Nhi Bao Hoang

DATA-6150-02

Professor [Mehmet Ergezer](mailto:ergezerm@wit.edu)

December 10, 2023

**DIABETES PREDICTION**

**Introduction:**

In this project, my central focus was dedicated to the investigation of diabetes prediction, propelled by a keen interest in health analytics. The impetus behind this exploration originates from a genuine desire to comprehend and visually represent key health parameters that contribute to the manifestation of diabetes. After engaging with the dataset, my objective was to employ advanced machine learning techniques, specifically decision tree and random forest classifiers, to predict the likelihood of an individual having diabetes based on their health attributes. This discourse invites you to accompany me as we uncover nuanced insights from the dataset, meticulously analyze underlying patterns, and navigate the sophisticated application of machine learning for diabetes prediction.

**Selection of Data:**

The dataset chosen for the diabetes prediction project was meticulously curated from Kaggle, representing a comprehensive repository of health-related information. Its pivotal role in unveiling discernible patterns and facilitating diabetes prediction through machine learning is underscored. The dataset is strategically designed to focus on pertinent parameters, including but not limited to blood glucose levels, BMI, age, and various other health indicators. During the initial data preprocessing phase, a systematic approach was undertaken to address null values by their elimination, ensuring the preservation of data integrity. Extraneous columns were judiciously excluded, and any instances of NaN values were methodically replaced with appropriate placeholders. This meticulous curation endeavors to provide a coherent and complete dataset, forming the foundation for subsequent in-depth analysis.

In alignment with the overarching goals of the project, numerical and categorical features in the dataset were accorded priority. Categorical features, such as gender and smoking history, underwent encoding to facilitate seamless integration with machine learning algorithms. An exploratory analysis of the numerical features within the dataset imparted valuable insights into their distributions and characteristics. This groundwork serves as a precursor to sophisticated visualizations, further exploration, and the subsequent implementation of machine learning models to predict diabetes based on the intricacies of the health attributes present in the dataset.

**Methods:**

The execution of this project involved the adept utilization of industry-standard tools and libraries, including NumPy, Pandas, Seaborn, Matplotlib, and SciKit. The implementation of a Random Forest Classifier emerged as a judicious choice for predictive modeling, leveraging the algorithm's efficacy in handling complex datasets and delivering accurate predictions.

**Tools Used:**

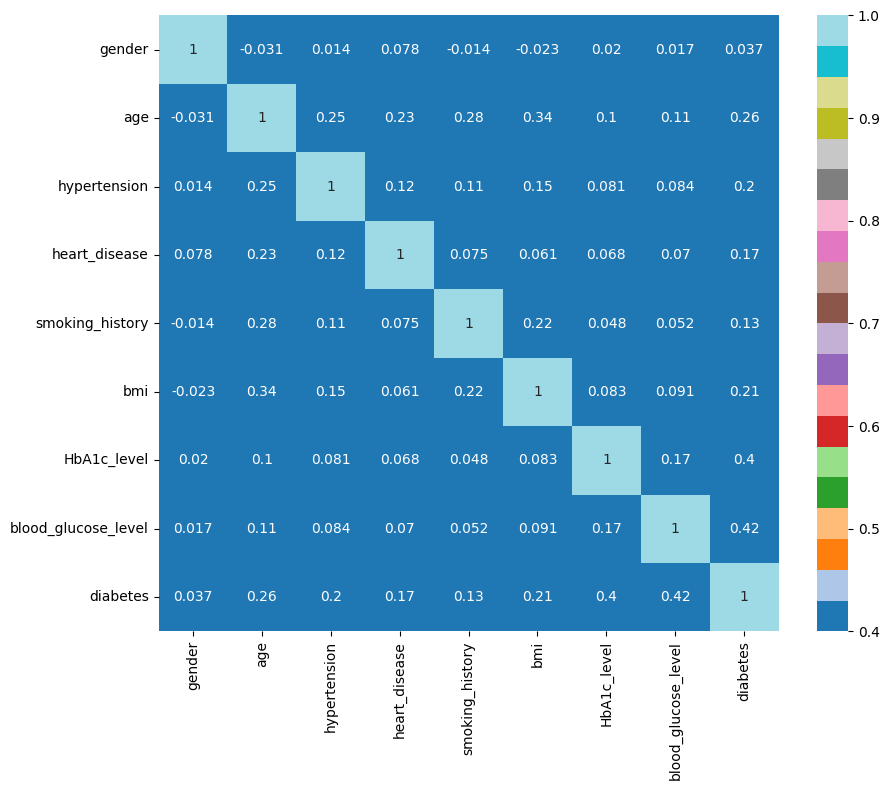
The coding and developmental aspects were proficiently carried out in Google Colab, leveraging its advanced features to ensure a streamlined and efficient workflow. The platform's capabilities were harnessed judiciously to enhance the overall development process, optimizing productivity, and facilitating a comprehensive analysis.

**Resources:**

Gained insights from Datacamp and Oracle blogs, combining structured learning with industry perspectives. Also, appreciated the support from friends and inanimate objects during explanations.

**Graphs:**

To investigate potential correlations between health indicators, a Seaborn-generated heatmap was employed. This visual aid aimed to uncover relationships between features, shedding light on potential patterns and dependencies.

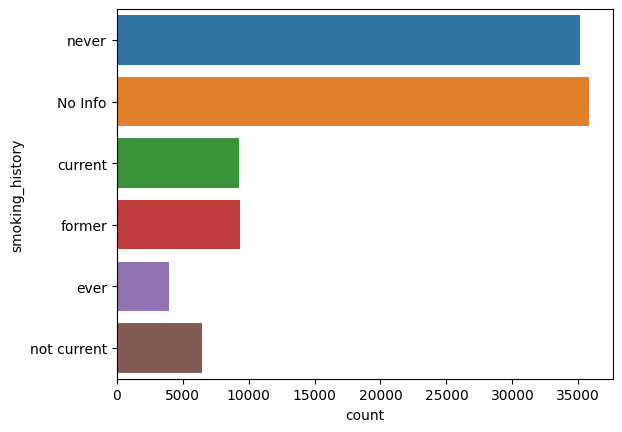


Subsequently, I delved into the distribution of all patients' health stats, creating a histogram to assess the equality of stat distribution across the entire dataset.

A graph of different types of blood sugar levels

Description automatically generated with medium confidence

Exploring the categorical features of the diabetes dataset, two key countplots shed light on significant aspects. The countplot for "smoking\_history" delineates the distribution of smoking histories, encompassing categories like "No Info," "Never," "Former," "Current," "Not Current," and "Ever." This visual provides a swift overview of the prevalence of each smoking category, offering context for potential associations between smoking history and diabetes.

Simultaneously, the countplot for "gender" offers insights into the gender composition, detailing categories such as "Female," "Male," and "Other." This visualization is pivotal for comprehending the dataset's gender-specific patterns, a crucial factor in the analysis of health-related data.  


A graph with a number of numbers

Description automatically generated with medium confidence

**Random Forest Results:**

The application of the Random Forest Classifier to predict diabetes yielded promising results, achieving an accuracy of approximately 97 % after training, this was after training the decision tree that gave 95% accuracy. This successful outcome, especially for a first-time engagement with such methodologies, underscores the effectiveness of the machine learning model. Precision, recall, and f1\_score also gave greater results.

**Random Forest Results:**

0.9722333333333333

[[27465 826]

[ 7 1702]]

precision recall f1-score support

0 1.00 0.97 0.99 28291

1 0.67 1.00 0.80 1709

accuracy 0.97 30000

macro avg 0.84 0.98 0.89 30000

weighted avg 0.98 0.97 0.97 30000

A user-friendly console input system was implemented to enhance practicality, enabling users to input six different health stats. The system then predicts whether the corresponding patient is diabetic or non-diabetic.

Enter HbA1c level: 5

Enter blood glucose level: 5

Enter age: 23

Enter BMI: 7

NO diabetes detected  
**Discussion:**

Examining the heatmap for the diabetes dataset indicated limited correlations among most health indicators. To explore potential influences, weight was considered as a factor affecting health stats. The histograms revealed a generally even distribution of health stats across all patients. However, when focusing on individuals with diabetes, some unevenness was observed, particularly in features like blood glucose level.

The scatter plots provided valuable insights into potential relationships. Contrary to expectations, the scatter plot comparing diabetes instances to heart disease, hypertension, and diabetes showcased complexities in these connections. This challenges assumptions and underscores the multifaceted nature of health conditions.

The Random Forest model's success in predicting diabetes with an accuracy of 95-97% is noteworthy, especially for a first-time application. The user input system further enhances practicality, enabling individuals to assess their diabetes risk based on provided health stats.

**Summary:**

Key findings emphasize the intricate nature of health conditions in the diabetes dataset. The lack of strong correlations among health indicators highlights the need for comprehensive analyses. The even distribution of health stats across patients underscores the diverse nature of diabetes cases.

The success of the Random Forest model in predicting diabetes showcases its potential utility in health predictions. The user-friendly input system adds a practical dimension, allowing individuals to assess their diabetes risk based on provided health indicators. Despite challenges, the project's outcomes contribute to a more nuanced understanding of health conditions and the potential for machine learning in predictive healthcare.

**References**

Janert, P. K. (2010). *Data analysis with open source tools: a hands-on guide for programmers and data scientists*. " O'Reilly Media, Inc.".

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learn-ing in Python, Journal of Machine Learning Re-search, 12.

Smith, A. (2023, February 14). Demystifying Random Forests in Python. Datacamp blog. <https://campus.datacamp.com/courses/kaggle-python-tutorial-on-machine-learning/improving-your-predictions-through-random-forests?ex=1>