Predictive Modelling For Diabetes Risk Assessment Using Logistic Regression: A SAS Analysis

Data set from: Diabetes Prediction (kaggle.com)

1 INTRODUCTION

- **Background:**

Diabetes is a chronic health condition characterized by elevated levels of blood sugar, which can lead to various complications if left untreated. Early detection and management of diabetes are crucial for preventing complications and improving patient outcomes. Machine learning models, such as logistic regression, can play a significant role in predicting the likelihood of diabetes based on risk factors.

- **Dataset Description:**

The dataset used for this analysis contains a comprehensive set of features associated with diabetes risk factors. These features include variables such as blood sugar levels, body mass index (BMI), age, family history, and other relevant health indicators. Each set of feature values is accompanied by a diagnosis label indicating whether the individual has diabetes or not.

- **Objectives:**

The primary objective of this analysis is to explore the relationship between various risk factors and the likelihood of developing diabetes. Specifically, we aim to:

- Apply logistic regression analysis to predict the likelihood of diabetes based on the provided risk factors.

- Interpret the coefficients of the logistic regression model to understand the impact of each risk factor on diabetes risk.

- Evaluate the performance of the logistic regression model in predicting diabetes diagnosis.

2 ANALYSIS (EDA)

- **Gender Distribution:**

- The dataset comprises 59.66% females and 40.34% males, indicating a slight majority of females.

- Understanding gender distribution is essential for assessing potential gender-based patterns in health-related variables.

- **Age Distribution:**

- The average age of individuals is approximately 41.56 years, with a standard deviation of around 22.57.

- The age distribution ranges from infancy (0.08 years) to elderly (80 years), with a median age of 43 years.

- The majority of individuals fall within the age range of early adulthood to middle age, with a peak at 43 years.

- **Blood Glucose Level:**

- The mean blood glucose level is approximately 137.31 mg/dL, with a standard deviation of around 40.63.

- Blood glucose levels range from 80 mg/dL to 300 mg/dL, with a median of 140 mg/dL.

- The distribution of blood glucose levels exhibits a slightly right-skewed pattern, indicating that the majority of individuals have blood glucose levels within the normal range.

- **BMI (Body Mass Index):**

- The average BMI is approximately 27.26, with a standard deviation of around 6.74.

- BMI values range from 10.69 to 79.46, with a median of 27.32.

- The distribution of BMI values appears to be slightly right-skewed, suggesting that a significant portion of individuals may be overweight or obese.

- **HbA1c Level:**

- The mean HbA1c level is about 5.53, with a standard deviation of approximately 1.06.

- HbA1c levels range from 3.5% to 9%, with a median of 5.8%.

- The distribution shows a nearly symmetric pattern, with a mode at 6.6%, indicating a common HbA1c level among the population.

3 LOGISTIC REGRESSION ANALYSIS

- **Model Interpretation:**

- Model Fit: The model with covariates (age, blood glucose level, BMI, and HbA1c level) fits significantly better than the intercept-only model, as evidenced by the likelihood ratio, score, and Wald tests ($p < 0.0001$ for all).

- Parameter Estimates: Each independent variable (age, blood glucose level, BMI, and HbA1c level) has a statistically significant effect on the likelihood of having diabetes, as indicated by their Wald chi-square tests ($p < 0.0001$ for all).

- Odds Ratios: The odds ratios indicate the multiplicative change in the odds of having diabetes associated with a one-unit increase in each independent variable. For example, for every one-unit increase in age, the odds of having diabetes increase by a factor of approximately 1.061, holding other variables constant.

- Goodness-of-Fit: The Hosmer and Lemeshow test, with a p-value of 0.2768, suggests that the model fits the data well, indicating no significant difference between observed and expected values across the ten groups.

4 HYPOTHESIS TESTING FOR PREDICTOR VARIABLES

- **Age:**

- Null Hypothesis (H0): The coefficient of age in the logistic regression model is equal to zero, implying that age has no effect on the log odds of diabetes.

- Alternative Hypothesis (H1): The coefficient of age is not equal to zero, indicating that age has a significant effect on the log odds of diabetes.

- Wald Chi-square Statistic: 111.89

- p-value: < 0.0001

- Conclusion: With a p-value < 0.0001, we reject the null hypothesis. Therefore, age has a statistically significant effect on the log odds of diabetes.

- **Blood Glucose Level:**

- Null Hypothesis (H0): The coefficient of blood glucose level is equal to zero, suggesting no association with the log odds of diabetes.

- Alternative Hypothesis (H1): The coefficient of blood glucose level is not equal to zero, indicating a significant effect on the log odds of diabetes.

- Wald Chi-square Statistic: 242.58

- p-value: < 0.0001

- Conclusion: The p-value < 0.0001, leading us to reject the null hypothesis. Thus, blood glucose level significantly influences the log odds of diabetes.

- **BMI:**

- Null Hypothesis (H0): The coefficient of BMI is zero, implying no impact on the log odds of diabetes.

- Alternative Hypothesis (H1): The coefficient of BMI is not zero, suggesting a significant effect on the log odds of diabetes.

- Wald Chi-square Statistic: 93.07

- p-value: < 0.0001

- Conclusion: Since the p-value < 0.0001, we reject the null hypothesis. Therefore, BMI significantly affects the log odds of diabetes.

- **HbA1c Level:**

- Null Hypothesis (H0): The coefficient of HbA1c level is zero, indicating no association with the log odds of diabetes.

- Alternative Hypothesis (H1): The coefficient of HbA1c level is not zero, suggesting a significant effect on the log odds of diabetes.

- Wald Chi-square Statistic: 227.32

- p-value: < 0.0001

- Conclusion: With a p-value < 0.0001, we reject the null hypothesis. Hence, HbA1c level significantly influences the log odds of diabetes.

5 CONCLUSION:

All four predictor variables (age, blood glucose level, BMI, and HbA1c level) exhibit statistically significant effects on the log odds of diabetes, as indicated by their respective Wald chi-square statistics and p-values.

This comprehensive analysis provides valuable insights into diabetes risk assessment, highlighting the importance of age, blood glucose level, BMI, and HbA1c level as predictors of diabetes likelihood.

6.FUTURE DIRECTIONS AND LIMITATIONS:

FUTURE DIRECTIONS:

1. Enhanced Predictive Models: Moving forward, refining our predictive models by incorporating advanced machine learning techniques could bolster their accuracy and reliability. Exploring algorithms like ensemble methods or deep learning architectures may unveil hidden patterns within the data, potentially leading to more precise risk assessments for diabetes.

2. Integration of Additional Data Sources: Expanding our analysis to incorporate data from diverse sources beyond the provided dataset could enrich our understanding of diabetes risk factors. Integration with electronic health records, wearable device data, or genomic information might offer supplementary insights into individual health profiles.

3. Clinical Implementation and Validation: Validating our predictive models in clinical settings through prospective studies is imperative for their real-world utility. Collaborating with healthcare institutions to deploy and assess the performance of our models in practice settings would validate their efficacy and inform future improvements.

4. Precision Medicine Approaches: Embracing a precision medicine paradigm, where interventions are tailored to individual characteristics, could revolutionize diabetes prevention and management. By leveraging predictive models alongside personalized health data, clinicians can deliver targeted interventions that address each patient's unique risk profile.

5. Patient Engagement and Education: Empowering individuals to take initiative-taking steps in managing their health through education and engagement initiatives is essential. Developing user-friendly tools and educational resources based on our predictive models can help raise awareness about diabetes risk factors and encourage healthy lifestyle choices.

LIMITATIONS:

1. Data Quality and Completeness: Given that our analysis relies on existing datasets, we must acknowledge potential limitations in data quality and completeness. Incomplete or inaccurate data entries could introduce biases or hinder the robustness of our findings, necessitating cautious interpretation.

2. Generalizability Concerns: While our analysis provides valuable insights, its generalizability to broader populations may be limited. The characteristics of the dataset, such as its demographic composition and geographic origin, could influence the applicability of our findings to other populations.

3. Model Interpretability: The complexity of predictive models, particularly advanced machine learning algorithms, may pose challenges in terms of interpretability. Ensuring that stakeholders, including clinicians and patients, can understand and trust the outputs of our models is critical for their adoption and implementation in healthcare settings.

4. Ethical Considerations: Ethical considerations surrounding data privacy, informed consent, and potential biases in algorithmic decision-making demand careful attention. Striving for transparency, fairness, and accountability in our analysis processes is paramount to uphold ethical standards and safeguard patient interests.

5. Continuous Evaluation and Iteration: Recognizing the dynamic nature of healthcare data and evolving clinical practices, our analysis should be subject to continuous evaluation and iteration. Regular updates and refinements to our predictive models based on new evidence and feedback from stakeholders can ensure their relevance and effectiveness over time.

By addressing these limitations and embracing future research directions, we can advance the field of diabetes risk assessment and contribute to improved health outcomes for individuals at risk of developing diabetes.