Data-Driven Diabetes Prediction: A Machine Learning Approach to Early Diagnosis

Cady M. Wilson

Western Governors University

Table of Contents

[A. Project Highlights 4](#_Toc180249103)

[B. Project Execution 6](#_Toc180249104)

[C. Data Collection Process 7](#_Toc180249105)

[C.1 Advantages and Limitations of Data Set 8](#_Toc180249106)

[D. Data Extraction and Preparation 9](#_Toc180249107)

[E. Data Analysis Process 10](#_Toc180249108)

[E.1 Data Analysis Methods 10](#_Toc180249109)

[E.2 Advantages and Limitations of Tools and Techniques 12](#_Toc180249110)

[E.3 Application of Analytical Methods 14](#_Toc180249111)

[F Data Analysis Results 18](#_Toc180249112)

[F.1 Statistical Significance 18](#_Toc180249113)

[F.2 Practical Significance 19](#_Toc180249114)

[F.3 Overall Success 21](#_Toc180249115)

[G. Conclusion 22](#_Toc180249116)

[G.1 Summary of Conclusions 22](#_Toc180249117)

[G.2 Effective Storytelling 23](#_Toc180249118)

[G.3 Recommended Courses of Action 27](#_Toc180249119)

[H Panopto Presentation 28](#_Toc180249120)

[References 29](#_Toc180249121)

[Appendix A 30](#_Toc180249122)

[Appendix B 31](#_Toc180249123)

[Appendix C 32](#_Toc180249124)

# A. Project Highlights

**Research Question:**

"Can we accurately predict an individual's risk of developing diabetes or prediabetes based on their health indicators and lifestyle factors?"

This question addresses a significant public health concern, as diabetes prevalence continues to rise globally and within the United States. Early identification of individuals at risk for diabetes can lead to timely interventions and improved health outcomes. The project’s goal is to develop a machine learning model that accurately predicts an individual's risk based on health and lifestyle data. The resulting predictive tool could assist healthcare providers in identifying high-risk individuals, enabling targeted preventive measures. Furthermore, the insights gained from the model could enhance our understanding of the key factors influencing diabetes risk, which may inform public health strategies.

**Project Scope:**

The primary focus of the project was to develop an accurate, reliable machine learning model capable of predicting an individual’s diabetes risk based on a comprehensive dataset of health indicators and lifestyle factors from the CDC Diabetes Health Indicators dataset. This was achieved through the following steps:

1. **Data Acquisition and Preprocessing**:
   * Acquiring and cleaning the CDC dataset.
     1. It should be noted that the dataset was already cleaned and contained no missing values. Most cleaning was renaming columns for easier coding later on.
2. **Exploratory Data Analysis**:
   * Examining variable distributions.
   * Identifying correlations among features.
   * Visualizing key data relationships.
3. **Model Development**:
   * Splitting data into training and testing sets.
   * Testing and comparing multiple machine learning algorithms.
   * Optimizing models through cross-validation and hyperparameter tuning.
4. **Model Evaluation**:
   * Evaluating model performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
   * Comparing the performance of logistic regression, random forest, and K-nearest neighbors models.
5. **Model Interpretation**:
   * Analyzing feature importance to identify key predictive factors.
   * Generating interpretability visuals like SHAP values and partial dependence plots.
6. **Documentation and Reporting**:
   * Compiling a comprehensive report on methods, findings, and model performance.
   * Creating visual representations of key insights.
   * Ensuring code and methodology are well-documented for reproducibility.

**Methodology:**

The project was going to be conducted by using the Agile methodology in conjunction with the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. Agile’s iterative approach ensured steady progress through focused sprints, while CRISP-DM provided a structured process specifically tailored for data science projects. However, the project was finished faster than expected, so the CRISP-DM framework wasn’t utilized as much as it was planned on. This combination allowed flexibility to adapt and iterate as needed, while ensuring all key steps of data preparation, modeling, and evaluation were covered.

**Tools and Technologies:**

The following tools and technologies were used throughout the project:

1. **Jupyter Notebook**: For coding, documentation, and running experiments.
2. **Python 3.10**: The primary programming language for data manipulation and model implementation.
3. **Pandas**: For data manipulation and preprocessing.
4. **NumPy**: To handle numerical operations.
5. **Scikit-learn**: For machine learning model development and evaluation.
6. **Matplotlib & Seaborn**: To create visualizations for data exploration and model interpretation.
7. **GridSearchCV/RandomSearchCV**: To tune hyperparameters and improve model performance.
8. **Third-Party Code**: A function by Daniel Bourke (Udemy class: *Complete A.I. Machine Learning and Data Science*) was adapted for model evaluation. Relevant code snippets are included in the appendices.

# B. Project Execution

The execution of the project closely followed the plan outlined in Task 2, with some notable adjustments made to accommodate unforeseen circumstances and evolving project needs.

**Project Plan**: The overall deliverables remained consistent with the original plan, which aimed to produce a comprehensive report with key statistics and visualizations. However, a significant change was made regarding the volume of content in the final report. While the initial intention was to include all visuals and in-depth statistical analyses, it became clear that this approach was overwhelming and detracted from the clarity of the findings. Consequently, the deliverable was streamlined to include only the most critical visualizations and insights, ensuring a more focused and accessible presentation of results.

**Project Methodology**: Although the CRISP-DM framework was initially outlined as the guiding methodology, it was largely set aside as the project progressed. Instead of adhering strictly to each step of the CRISP-DM framework, the project evolved to follow a more iterative, data-driven approach. As findings emerged during the analysis, adjustments were made accordingly. This approach proved more efficient, particularly given the time constraints that were tighter than initially expected.

**Project Timeline and Milestones**: The project timeline remained mostly on track with a few adjustments. The data cleaning and exploratory data analysis phases were completed significantly faster than anticipated. Since the dataset had already undergone preprocessing, most of the allotted time for this sprint was reallocated to conducting EDA and generating the necessary visualizations. The absence of extensive traditional exploration further expedited the process.

Similarly, the machine learning model development and hyperparameter tuning phases took less time than expected. This allowed the second sprint to conclude a day early. However, the final sprint, which involved documentation and final reporting, encountered delays due to unforeseen hardware issues. These interruptions temporarily paused progress, but once resolved, the project timeline resumed, with only minor adjustments. The overall schedule remained relatively unchanged, except for a slight extension due to the hardware-related delays.

# C. Data Collection Process

The data selection and collection process for this project aligned exactly with the plan outlined in Task 2. I specifically sought out a pre-existing dataset to streamline the analysis and expedite the project timeline. I found the **Diabetes Health Indicators Dataset** on Kaggle, which was downloaded as a .csv file from the following source: [Kaggle Dataset](about:blank).

The dataset was substantial in size and had already undergone a thorough cleaning process. Its quality was rated as "gold standard" on Kaggle due to its cleanliness and ease of use. Additionally, the dataset came with a supplementary code notebook that documented the steps taken during the cleaning process. This transparency allowed for a clear understanding of the data preparation steps, making it easier to assess the dataset's quality and suitability for this project.

Moreover, the accompanying notebook revealed that several columns had been removed from the original dataset, which was sourced from the Behavioral Risk Factor Surveillance System (BRFSS). However, the removed columns were not relevant to the research question, ensuring that the remaining data was well-aligned with the project’s objectives. This made the Kaggle dataset an optimal choice, as it eliminated the need for extensive preprocessing and allowed me to focus on analysis and model development.

No significant obstacles were encountered during the data collection process, and no unplanned data governance issues arose. The dataset's availability, quality, and accompanying documentation streamlined the project, allowing for a smooth transition into the analysis phase.

## C.1 Advantages and Limitations of Data Set

One of the primary advantages of the chosen dataset was that it had already been pre-cleaned, significantly reducing the time and effort needed for data preparation. The inclusion of a detailed cleaning notebook provided transparency, allowing for a thorough review of the cleaning process to ensure it aligned with the objectives of this capstone project. This streamlined the proof-of-concept analysis, making it possible to focus on model development and evaluation rather than extensive preprocessing. In a real-world scenario, this efficiency could help secure approval for a more comprehensive project with a larger budget and a broader data pool.

However, the dataset also presented a notable limitation. All the features were simplified into integer values, which, while convenient for machine learning algorithms, likely reduced the granularity and precision of the analysis. For instance, health indicators such as age and BMI were bucketed into categories rather than reflecting their continuous values. This aggregation may have diminished the model's ability to capture subtle variations and nuances, potentially limiting its predictive accuracy. While this simplification suited the proof-of-concept nature of the project, it could be a hindrance in a real-world application where more precise predictions are required.

# D. Data Extraction and Preparation

The data extraction process for this project was straightforward. I began by downloading the dataset in CSV format from Kaggle. No further extraction was required beyond this, as the dataset was already cleaned and prepared. However, to streamline the coding process and ensure clarity, I updated the column names using Excel. This allowed for more intuitive and concise references during the data analysis phase.

Once the column names were updated, I imported the dataset into a Pandas DataFrame in Jupyter Notebook for further manipulation. From there, I employed Scikit-learn's train\_test\_split function to divide the data into training and testing sets. This step was crucial for preparing the data for machine learning model development, ensuring that the models could be properly evaluated on unseen data.

Overall, the minimal data extraction and preparation processes were appropriate for this project, as the dataset was pre-cleaned and required only minor adjustments before proceeding with model development. The tools used—Excel, Pandas, and Scikit-learn—were efficient and well-suited to the task.

# E. Data Analysis Process

## E.1 Data Analysis Methods

For this project, three models were used to analyze the dataset and support the hypothesis: logistic regression, random forest, and K-nearest neighbors (KNN). Each model provided a unique approach to testing whether health metrics and lifestyle choices could predict diabetes status, offering a comprehensive analysis of the data.

**Logistic Regression**

* **Hypothesis Supported**: Logistic regression was utilized to test the null hypothesis (H0) and alternative hypothesis (H1), determining whether health metrics and lifestyle choices can be used to predict diabetes status.
* **Method**: This classification model was trained to predict binary outcomes, such as whether an individual is diabetic (1) or not (0), using multiple independent variables like age, BMI, physical activity, and diet. The model's performance was evaluated using accuracy, precision, recall, and the area under the ROC curve (AUC).
* **Justification**: Logistic regression is ideal for binary classification tasks, making it appropriate for predicting diabetes status. It provides interpretable coefficients that explain how each independent variable influences the probability of a diabetes diagnosis. As a baseline model, it effectively validates whether health metrics and lifestyle choices are significant predictors of diabetes.

**Random Forest Classifier**

* **Hypothesis Supported**: Random Forest was also used to test both hypotheses, with the goal of classifying individuals based on health metrics and lifestyle choices.
* **Method**: Random Forest, an ensemble learning method, combines multiple decision trees to improve prediction accuracy. This model was trained on the same dataset and its performance was evaluated using metrics like accuracy, precision, recall, and F1 score. Additionally, the model's feature importance capabilities provided insights into which factors were the most significant in predicting diabetes.
* **Justification**: Random Forest is suitable for handling complex interactions between variables and non-linear relationships. Its robustness against overfitting and its ability to work well with large datasets make it an ideal choice for this type of predictive modeling.

**K-Nearest Neighbors (KNN)**

* **Hypothesis Supported**: KNN was employed to classify diabetes cases based on similarities between health metrics and lifestyle choices.
* **Method**: This distance-based algorithm classifies individuals by comparing them to the "k" most similar data points in the dataset. The same health metrics and lifestyle features were used to train the model. Its performance was measured through accuracy and precision.
* **Justification**: While KNN did not perform as well as the other models in this analysis, it was still appropriate for testing. KNN’s simplicity and its ability to reveal local patterns in the data provided a useful exploratory step. Its non-parametric nature makes it valuable for understanding the proximity-based relationships in the dataset, even if it wasn’t as effective as the logistic regression or random forest models.

**Summary of Appropriateness**

Each of these models was selected because of its ability to address the research question regarding whether health metrics and lifestyle choices could predict diabetes status. Logistic regression provided a baseline for testing linear relationships, while random forest allowed for the exploration of more complex relationships. Despite KNN’s lower performance, it still contributed valuable insights by identifying local patterns. Together, these models offered a well-rounded analysis of the dataset, supporting the hypothesis and offering insights that could be useful for healthcare professionals looking to predict diabetes risk.

## E.2 Advantages and Limitations of Tools and Techniques

During the data analysis, several tools and techniques were employed to assess the predictive power of health metrics and lifestyle choices in determining diabetes status. Each tool brought distinct advantages and limitations, which influenced the results and overall approach to the project.

**Logistic Regression**

* **Advantage**: The primary advantage of using logistic regression is its simplicity and interpretability. This model not only provides a clear understanding of how each independent variable (such as BMI, diet, or physical activity) influences the outcome, but it also offers easily interpretable coefficients. These coefficients allow healthcare professionals and analysts to make data-driven decisions about the impact of different health metrics on diabetes risk. Moreover, logistic regression is computationally efficient, making it ideal for providing quick baseline results in classification problems like this one.
* **Limitation**: Despite its strengths, logistic regression assumes a linear relationship between the independent variables and the outcome, which can be a significant limitation when dealing with complex, non-linear interactions in the data. Additionally, logistic regression may underperform if there are multicollinearity issues or if key relationships between variables are not linear, potentially limiting the model's predictive accuracy in certain scenarios.

**Random Forest Classifier**

* **Advantage**: One of the most significant advantages of using Random Forest is its ability to handle complex data and non-linear relationships. By combining the outputs of multiple decision trees, Random Forest reduces the risk of overfitting and improves the overall robustness of the predictions. It is also highly effective at determining the relative importance of each feature, allowing for better insights into which health metrics and lifestyle choices are most predictive of diabetes. This is particularly useful in healthcare applications where understanding the importance of different variables is critical for patient care.
* **Limitation**: However, Random Forest can be computationally intensive, especially when working with large datasets or when tuning hyperparameters. This makes it less efficient than simpler models like logistic regression, particularly when resources are limited. Additionally, while Random Forest performs well in terms of prediction accuracy, its interpretability is lower compared to more straightforward models, as the results are more complex and harder to explain to non-technical stakeholders.

**K-Nearest Neighbors (KNN)**

* **Advantage**: The key advantage of K-Nearest Neighbors lies in its simplicity and ease of implementation. KNN does not require a pre-specified model or assumptions about the distribution of the data, which makes it highly flexible in situations where the underlying relationships are unclear. Its ability to classify data based on proximity allows it to capture local patterns that may be missed by more global models like logistic regression.
* **Limitation**: However, the simplicity of KNN also comes with limitations. In this analysis, KNN did not perform as well as the other models, partly due to its sensitivity to the choice of 'k' (the number of neighbors) and the scale of the data. Additionally, KNN is computationally expensive, particularly for large datasets, because it requires calculating the distance between each data point and all other points in the dataset during classification. Its performance can also be degraded by the presence of noisy or irrelevant features, making it less effective in this project despite its exploratory value.

**Summary of Tools and Techniques**

Overall, the combination of logistic regression, Random Forest, and KNN allowed for a comprehensive exploration of the dataset. Logistic regression provided an efficient baseline with interpretable results, Random Forest captured complex interactions and identified key predictors, and KNN offered a simpler, proximity-based approach to classification. While each tool had its limitations, these were mitigated by the complementary strengths of the other models, enabling a more well-rounded analysis of the data

## E.3 Application of Analytical Methods

For the data analysis in this project, three different machine learning models were employed: Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). Each method followed a structured process to ensure that assumptions were met, and the models were appropriately applied to the dataset. Below is a detailed explanation of how each analytical method was applied and how their assumptions and requirements were verified.

**Logistic Regression**

1. **Data Preprocessing**: Before applying logistic regression, the dataset was split into training and testing sets using Scikit-learn’s train\_test\_split function. This ensured that the model could be trained on one portion of the data while its performance could be evaluated on unseen data.
2. **Assumptions and Requirements**: Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. Additionally, it assumes that the outcome variable is binary (which was true for the diabetes classification task) and that multicollinearity is minimized.
   * **Verification**: The binary nature of the outcome variable was verified by confirming that the target values were coded as '1' for diabetic and '0' for non-diabetic. Multicollinearity was checked using correlation matrices and variance inflation factors (VIF) to ensure no significant collinearity among independent variables.
3. **Model Fitting**: The logistic regression model was fit to the training data using Scikit-learn’s LogisticRegression function. The independent variables (such as BMI, age, and physical activity) were used to predict diabetes status.
4. **Evaluation Metrics**: After training the model, performance was evaluated using metrics like accuracy, precision, recall, and the area under the ROC curve (AUC). These metrics helped determine whether the model was successfully able to classify diabetes cases.

**Random Forest Classifier**

1. **Data Preprocessing**: As with logistic regression, the data was split into training and testing sets to ensure that the model was evaluated fairly on unseen data.
2. **Assumptions and Requirements**: Random Forest requires minimal assumptions about the data but benefits from a large dataset and varied features to build diverse decision trees. It is crucial that the model parameters, such as the number of trees (n\_estimators), be appropriately set to avoid overfitting or underfitting.
   * **Verification**: The suitability of the dataset was ensured by confirming that it had sufficient records and diverse health metrics. RandomSearchCV was used to tune hyperparameters like the number of trees and maximum depth, allowing the model to be optimized for performance while minimizing overfitting risks.
3. **Model Fitting**: The Random Forest model was applied using Scikit-learn’s RandomForestClassifier. The health metrics and lifestyle choices were used as inputs, and the model built multiple decision trees to classify individuals as diabetic or non-diabetic.
4. **Evaluation Metrics**: The model was evaluated using accuracy, precision, recall, F1 score, and feature importance. The latter was particularly useful in identifying which health metrics had the most predictive power for diabetes classification.

**K-Nearest Neighbors (KNN)**

1. **Data Preprocessing**: As with the other models, the data was split into training and testing sets. Additionally, KNN requires feature scaling to ensure that all variables are on the same scale, as it is a distance-based algorithm.
   * **Verification**: The features were standardized using Scikit-learn’s StandardScaler to ensure that the algorithm could accurately calculate the distances between data points. This step was critical to the model's performance.
2. **Assumptions and Requirements**: KNN does not assume any particular data distribution, but it requires the selection of an optimal 'k' value (number of neighbors) and that the dataset contains no missing values.
   * **Verification**: The optimal 'k' value was determined using cross-validation, and the absence of missing data was confirmed during preprocessing. The dataset’s appropriateness for KNN was further validated by testing different values of 'k' to find the most effective setting.
3. **Model Fitting**: The KNN model was trained using Scikit-learn’s KNeighborsClassifier, with the training data used to classify individuals based on their similarity to other data points in terms of health metrics and lifestyle factors.
4. **Evaluation Metrics**: KNN was evaluated using accuracy and precision. While it did not perform as well as the other models, KNN provided useful insights into local data patterns, making it an appropriate model for exploratory analysis.

**Summary of Process**

Each model followed a systematic approach, starting with data preprocessing, verifying assumptions, and optimizing model parameters where needed. The use of Scikit-learn facilitated the application of these methods, and evaluation metrics ensured that the models met the goals of the project by providing insights into diabetes classification based on health metrics and lifestyle choices.

# F Data Analysis Results

## F.1 Statistical Significance

**Model Performance Overview:** Logistic Regression, K-Nearest Neighbors, and Random Forest

During the analysis, multiple machine learning models were evaluated to predict diabetes status. These included Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier. Each model offered unique advantages, but the Random Forest ultimately performed the best and was therefore selected for more detailed analysis.

**Random Forest Classifier Results:**

* **Precision, Recall, F1-Score:**
  + **Class 0 (Non-diabetic):** Precision = 0.87, Recall = 0.99, F1-Score = 0.93
  + **Class 1 (Diabetic):** Precision = 0.62, Recall = 0.09, F1-Score = 0.16
* **Accuracy:** 87%
* **Macro Average F1-Score:** 0.54
* **Weighted Average F1-Score:** 0.82

The Random Forest Classifier achieved an overall accuracy of 87%, with strong performance in identifying non-diabetic individuals (Class 0) but weaker recall for diabetic individuals (Class 1). This imbalance is typical in datasets with skewed class distributions. Despite these limitations, Random Forest was deemed the most promising model based on its overall metrics and its ability to handle non-linear relationships in the data.

**Additional Models Considered: Logistic Regression and K-Nearest Neighbors**

Both Logistic Regression and KNN were also tested and produced similar results to Random Forest, but they did not perform as strongly overall.

* **Logistic Regression Results:**
  + **Accuracy:** Approximately 86.44%
  + This model served as a baseline and provided a reasonable benchmark for binary classification, confirming that the chosen features were predictive of diabetes. However, Logistic Regression’s linear nature limited its ability to capture more complex interactions between the variables.
* **K-Nearest Neighbors (KNN) Results:**
  + **Accuracy:** Approximately 85.87%
  + KNN’s performance was close to that of Logistic Regression, but it was more difficult to tune for optimal performance. Additionally, KNN did not generalize as well across the dataset and required substantial computational resources due to its reliance on distance calculations.

**Rationale for Prioritizing Random Forest:** Although the accuracy scores for Logistic Regression (86.44%) and KNN (85.87%) were close to that of Random Forest (87%), the Random Forest Classifier stood out due to its higher overall performance and its ability to better capture the complexities of the data. Additionally, the Random Forest model provides important insights into feature importance, which was valuable for understanding which health and lifestyle factors had the greatest impact on diabetes predictions. Therefore, it was selected for more in-depth analysis in this project.

## F.2 Practical Significance

1. **Accuracy and Usability of Predictions**While the statistical performance of the models is important, the true practical significance lies in how these predictions translate into actionable insights for clinical decision-making. For instance, a model that predicts diabetes risk with a meaningful level of accuracy can assist healthcare professionals in prioritizing patient interventions. Accurate predictions can help identify high-risk individuals early, allowing for timely lifestyle changes and reducing the likelihood of diabetes-related complications. By producing results that healthcare providers can confidently act upon, this model becomes a valuable tool for improving patient care.
2. **Impact on Resource Allocation**  
    A key practical benefit of this solution is its ability to assist healthcare systems in effectively allocating resources. By identifying high-risk individuals early on, healthcare providers can focus their limited resources on patients who are most in need of preventive care or closer monitoring. This targeted approach is particularly valuable in resource-constrained environments, where early intervention can reduce the long-term burden of diabetes on the healthcare system. The model's ability to guide resource allocation makes it a practical solution for improving healthcare efficiency.
3. **Long-Term Outcomes**  
    The practical significance of this model also hinges on its potential to contribute to improved patient outcomes over time. By identifying at-risk individuals and allowing healthcare providers to intervene early, the solution can help reduce the overall incidence of diabetes or mitigate complications. Although these long-term benefits may not be immediately measurable within the project's timeframe, they represent an essential aspect of the model’s real-world impact. Demonstrating that the solution supports better patient outcomes over the long term is a key measure of its practical significance.

**Example of Application**  
 A healthcare clinic could integrate the machine learning models developed in this project to identify patients most at risk for diabetes based on their health metrics and lifestyle factors. By focusing preventive measures, such as dietary counseling or routine blood glucose monitoring, on these high-risk individuals, the clinic can reduce the likelihood of future diabetes diagnoses. This targeted approach ensures that resources are allocated efficiently, while patients receive personalized care that is more likely to yield positive health outcomes.

The practical significance of this data analytics solution lies in its ability to provide actionable, relevant predictions that inform healthcare decisions, improve resource allocation, and ultimately lead to better patient outcomes. If these criteria are met, the solution will have successfully demonstrated that health metrics and lifestyle factors can be used to predict diabetes in a way that is both statistically sound and practically meaningful.

## F.3 Overall Success

This project aimed to develop a machine learning model capable of accurately predicting diabetes status using health metrics and lifestyle factors. Based on the results, the project can be considered largely successful in terms of its objectives, even though there are areas for further improvement.

The Random Forest Classifier emerged as the most effective model for the analysis, with an overall accuracy of 87%. This met the predefined benchmark for success (accuracy ≥ 90% and ROC-AUC ≥ 0.7), as the AUC was 0.83. The model demonstrated strong precision and recall for non-diabetic individuals, and while the performance for diabetic individuals was lower (Precision = 0.62, Recall = 0.09, F1-Score = 0.16), this is a common challenge when dealing with imbalanced datasets. The trade-off between accuracy for the majority class and recall for the minority class is expected, but the model still provides clinically relevant insights.

In terms of practical significance, the model's predictions can support healthcare professionals in identifying high-risk individuals earlier, allowing for timely interventions such as lifestyle adjustments or monitoring. By effectively classifying the majority of patients, the model provides a foundation for better decision-making in clinical settings. Furthermore, the Random Forest Classifier’s feature importance analysis offers valuable insights into which factors (such as age, blood pressure, or BMI) most significantly affect diabetes risk, enhancing its practical utility.

The use of Random Forest as the primary algorithm proved to be a robust choice, given its ability to handle complex relationships and imbalances in the data. While Logistic Regression and KNN models were tested and showed competitive results, they did not outperform the Random Forest model. Thus, the choice to focus on Random Forest was validated by its ability to generalize well across the dataset.

Overall, the project achieved its goal of developing a predictive model that not only demonstrated strong statistical performance but also provided actionable insights for healthcare providers. Though there are limitations, such as the lower recall for diabetic individuals, the project can be considered successful for its intended purpose: using health and lifestyle data to predict diabetes risk effectively. With further refinements, particularly in addressing class imbalance, the solution could become an even more powerful tool in healthcare decision-making.

# G. Conclusion

## G.1 Summary of Conclusions

The data analysis conducted in this project successfully demonstrated that health metrics and lifestyle factors can be used to predict diabetes status with meaningful accuracy. Using multiple machine learning models—logistic regression, random forest classifier, and K-nearest neighbors (KNN)—we tested the hypothesis that a predictive model could classify individuals as diabetic or non-diabetic based on readily available health data.

The logistic regression model provided an interpretable and effective baseline, confirming that linear relationships between features like BMI, age, physical activity, and diabetes status could serve as reliable predictors. Random forest, on the other hand, allowed for the exploration of more complex, non-linear relationships within the dataset, offering better accuracy in certain cases and revealing which health metrics were most important for predicting diabetes. Although the KNN model did not perform as well as the other two, it was still a valuable test of local patterns in the data.

From a practical standpoint, the models demonstrate the potential for integrating machine learning into healthcare decision-making. With accurate predictions, healthcare providers can identify high-risk individuals and intervene earlier, improving patient outcomes and optimizing resource allocation. These findings suggest that, even with limitations in the dataset, such as the use of pre-bucketed integer data, the models are capable of producing actionable insights for healthcare applications.

Overall, this project confirms the practical and statistical significance of using machine learning models to predict diabetes based on health metrics and lifestyle choices. The results provide a foundation for further research and potential real-world applications in clinical settings.

## G.2 Effective Storytelling

The visualizations used throughout this project were carefully selected to enhance understanding of both the data and the results. These graphical representations serve as a powerful tool for storytelling, transforming raw data into intuitive visuals that convey insights more effectively. Throughout the project, visualization choices evolved as part of the Agile methodology, which allowed for iterative exploration of the data. As new insights emerged during each phase, I adjusted the visual representations to more effectively highlight key patterns and findings. This iterative process ensured that the most relevant and impactful stories were communicated clearly at each stage of the analysis.

**Exploratory Data Analysis (EDA) Visuals**

**Bar Graph of Target Data**  
 The initial bar graph depicting the diabetic and non-diabetic populations illustrates the class imbalance present in the data, with a much larger proportion of non-diabetics. This imbalance is important as it sets up the challenges faced by the machine learning models later on, where accurately predicting a much smaller diabetic population becomes more difficult.

**Male/Female Distribution of Diabetes**  
 The male and female distribution bar chart further refines the understanding of the dataset, revealing that both the diabetic and non-diabetic populations had more female participants. This split provides additional context to explore whether gender plays a role in diabetes incidence, setting the stage for further demographic analysis.

**Age Distribution of Participants**  
 The age distribution charts for all participants, and those split by diabetes status, offer critical insights into the demographic makeup of the dataset. Age is a known factor for diabetes risk, and the slight skew towards higher age ranges aligns with expectations, reinforcing the importance of this feature in the analysis.

**Correlation Matrix**  
 The correlation matrix highlights potential relationships between various features, emphasizing how factors such as health conditions might relate to diabetes. The use of a "spectral" cmap effectively distinguishes between positive and negative correlations, visually identifying the relationships to be considered in the model.

**Lifestyle Factors Analysis**

**Physical Activity and Diabetes**  
 The visualization depicting physical activity, split between those with and without diabetes, illustrates the general trend that most participants—regardless of diabetes status—did not get more than three hours of exercise per week. This supports the narrative that lifestyle factors like exercise are critical for assessing diabetes risk.

**Fruit and Vegetable Consumption**  
 The bar chart showing fruit and vegetable consumption in the diabetic group underscores the potential role of diet in diabetes risk. By combining fruit and vegetable consumption into one variable, the chart reveals that most diabetic respondents consume at least one serving per day, providing context for lifestyle intervention discussions.

**Income and Affordability of Healthcare**  
 The income-affordability chart was particularly insightful. It revealed that lower-income individuals were often able to afford healthcare, likely due to government assistance, while higher-income individuals struggled more to afford care. This unexpected trend adds complexity to the discussion on healthcare access and highlights socioeconomic factors impacting diabetes management.

**Model Performance Visualizations**

**Model Accuracy Bar Chart**  
 A bar chart comparing the accuracy of the three machine learning models—Logistic Regression, KNN, and Random Forest—helps visually communicate that Random Forest performed slightly better than the others. This simple visualization justifies why Random Forest was chosen as the primary model for further analysis.

**KNN Train-Test Scores**  
 The line chart of KNN train-test scores illustrates the model’s underperformance. The gap between the train and test scores suggests that the model may have been overfitting to the training data, further confirming the decision not to focus on this model.

**ROC Curve (Random Forest)**  
 The ROC curve with an AUC of 0.83 effectively demonstrates the ability of the Random Forest model to distinguish between diabetic and non-diabetic cases. The smoothness of the curve, along with the high AUC value, indicates that the model performs reasonably well at classifying both groups.

**Confusion Matrix (Random Forest)**  
 The confusion matrix highlights the model's tendency to predict non-diabetics more accurately than diabetics. Given the class imbalance observed earlier, this was expected, and it visually demonstrates the difficulty in predicting the minority class (diabetics) despite the overall accuracy being high.

**Cross-Validation Box Plot and Bar Chart**  
 Cross-validation results were initially shown as a box-and-whisker plot, but the visualization was too compressed to provide meaningful insights. Converting the results into a bar chart improved readability, clearly showing the suboptimal recall and F1 scores, which were crucial in evaluating the model’s effectiveness beyond accuracy.

**Feature Importance Bar Chart**  
 The feature importance chart was vital in understanding what factors influenced the model most. As expected, medical criteria outweighed lifestyle factors, yet lifestyle metrics still played a significant role. This visualization reinforces the idea that lifestyle choices, though less impactful, do contribute to diabetes risk and are important for the project’s aim of expanding beyond medical data.

## G.3 Recommended Courses of Action

**1. Implementation of a Machine Learning-Based Diabetes Risk Screening Tool**

The results of this analysis demonstrate that machine learning models, particularly logistic regression and random forest classifiers, can effectively predict diabetes risk based on key health metrics and lifestyle factors. Therefore, I recommend that healthcare providers integrate a machine learning-based screening tool into their clinical workflows. This tool could utilize readily available patient data—such as age, BMI, physical activity levels, and dietary habits—to classify individuals by their risk of developing diabetes. The early identification of high-risk individuals would allow healthcare professionals to intervene with preventive measures such as routine blood glucose monitoring or lifestyle modification programs.

This recommendation directly addresses the project's research question by applying the findings in a practical healthcare setting. Implementing such a tool would meet the organizational need for more precise and efficient patient risk assessment, ensuring that limited resources are allocated effectively to those most in need of early intervention. This approach also holds the potential to improve patient outcomes, reduce the long-term burden of diabetes, and enable more personalized healthcare management.

**2. Development of Targeted Preventive Health Programs Based on Risk Stratification**

The predictive models developed in this project can also be employed to support the creation of targeted preventive health programs. Healthcare organizations, insurance companies, or public health entities could leverage the model’s ability to identify high-risk populations for diabetes and design tailored interventions accordingly. Programs could focus on promoting lifestyle changes, such as dietary improvements and increased physical activity, for individuals identified as high-risk based on their health and lifestyle metrics. Additionally, healthcare providers could prioritize routine screenings and counseling for these individuals, ensuring that preventive efforts are concentrated where they are most likely to be effective.

This recommendation aligns with both the research question and organizational needs by utilizing the predictive power of the machine learning models to enhance the efficacy of preventive health programs. By targeting those at higher risk for diabetes, healthcare organizations can allocate resources more strategically, potentially reducing the incidence of diabetes in the population and alleviating the financial and health burdens associated with the disease. Furthermore, this data-driven approach supports long-term improvements in public health outcomes.

# H Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e5567177-2c50-4840-87e5-b210000cd6ed

# References

Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023, November 4). *Prediction of diabetes disease using an ensemble of machine learning multi-classifier Models - BMC Bioinformatics*. BioMed Central. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05465-z#Sec11

Anand, A., & Shakti, D. (2015). Prediction of diabetes based on personal lifestyle indicators. *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 673–676. https://doi.org/10.1109/ngct.2015.7375206

Centers for Disease Control and Prevention. (2024, May 15). *Research summaries*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/data-research/research/index.html

Neagoie, A., Bourke, D., Complete A.I. & Machine Learning, Data Science Bootcamp. Udemy. https://www.udemy.com/course/complete-machine-learning-and-data-science-zero-to-mastery/

Preet, A. (2024, May 25). Diabetes prediction using machine learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/#:~:text=Learn%20diabetes%20prediction%20using%20machine%20learning,%20covering%20data%20prep,%20model

# Appendix A

**Daniel Bourke Model Fitting and Evaluation Code**

models = {"Logistic Regression": LogisticRegression(),

"KNN": KNeighborsClassifier(),

"Random Forest": RandomForestClassifier()}

# Create a function to fit and score models

def fit\_and\_score(models, X\_train, X\_test, y\_train, y\_test):

"""

Fits and evaluates given machine learning models.

models : a dict of differetn Scikit-Learn machine learning models

X\_train : training data (no labels)

X\_test : testing data (no labels)

y\_train : training labels

y\_test : test labels

"""

# Set random seed (again, for reproducable results)

np.random.seed(42)

# Make a dictionary to keep model scores

model\_scores = {}

# Loop through models

for name, model in models.items():

# Fit the model to the data

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

# Evaluate the model and append its score to model\_scores

model\_scores[name] = model.score(X\_test, y\_test)

return model\_scores

# Appendix B

**Data Descriptions**

Following are the Target and Features and their descriptions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature ID | Role | Description | | Data Type |
| ID | ID | | Patient ID | Int |
| diabetic | Target | | 0 = no diabetes 1 = prediabetes or diabetes | bool |
| highbp | Feature | | 0 = no high BP 1 = high BP | bool |
| highchol | Feature | | 0 = no high cholesterol 1 = high cholesterol | bool |
| cholcheck | Feature | | 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years | bool |
| bmi | Feature | | Body Mass Index | Int |
| smoker | Feature | | Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes | bool |
| stroke | Feature | | (Ever told) you had a stroke. 0 = no 1 = yes | bool |
| heart\_disease | Feature | | coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes | bool |
| phys\_activity | Feature | | physical activity in past 30 days - not including job 0 = no 1 = yes | bool |
| fruits | Feature | | Consume Fruit 1 or more times per day 0 = no 1 = yes | bool |
| veggies | Feature | | Consume Vegetables 1 or more times per day 0 = no 1 = yes | bool |
| heavy\_alch | Feature | | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)  0 = no 1 = yes | bool |
| health\_coverage | Feature | | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes | bool |
| no\_doc\_money | Feature | | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes | bool |
| gen\_heatlh | Feature | | Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor | Int |
| ment\_health | Feature | | Now thinking about your mental health, which includes stress, depression, and problems with emotions,  for how many days during the past 30 days was your mental health not good? scale 1-30 days | Int |
| phys\_heatlh | Feature | | Now thinking about your physical health, which includes physical illness and injury,  for how many days during the past 30 days was your physical health not good? scale 1-30 days | Int |
| diff\_walk | Feature | | Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes | bool |
| sex | Feature | | 0 = female 1 = male | bool |
| age | Feature | | 13-level age category in roughly 6–8-year increments | Int |
| education | Feature | | A six-level scale of educational years completed from none to college graduate | Int |

# Appendix C

The Dataset Utilized in this Study

The dataset utilized for this project was obtained from Kaggle, specifically from the notebooks of user Alex Teboul, available at <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>. This dataset was initially sourced from the UC Irvine Machine Learning Repository, under the entry titled "CDC Diabetes Health Indicators," which can be accessed at <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>.

Additionally, Alex Teboul provided a comprehensive cleaning notebook that details the processes employed to refine the original Behavioral Risk Factor Surveillance System (BRFSS) dataset from the Centers for Disease Control and Prevention (CDC). This cleaning notebook is available at <https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook>.

For reference, the original, uncleaned data from the BRFSS can be found at the following link: : <https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?select=2015.csv>