NAME: FELIX ASIBOR

TECHNICAL REPORT FOR DIABETES PREDICTION

Introduction

Objective of the Project

The objective of this project is to explore a comprehensive dataset focused on diabetes prediction, analyzing how demographic factors, lifestyle choices, and clinical parameters contribute to diabetes risk. The goal is to uncover meaningful correlations and trends that can inform early diagnosis and public health strategies.

Problem Being Addressed

This analysis aims to answer the following key questions:

- Does BMI or Waist Circumference correlate more strongly with Fasting Blood Glucose?
- Is there a difference in HbA1c levels across different ethnicities?
- How does Smoking Status impact LDL and HDL cholesterol levels?
- What is the relationship between Physical Activity Level and Blood Pressure?

Key Datasets and Methodologies

Dataset Used: A Comprehensive Dataset for Diabetes Prediction with various parameters.

Methods in Excel

- **Pivot Tables:** To group, filter, and summarize data across variables like ethnicity, smoking status, and physical activity levels
- **Charts (Bar charts, Column, and Line Charts):** To visualize trends and correlations among key clinical indicators and risk factors

Story of Data

Data Source

The dataset used in this analysis is publicly available on Kaggle, titled “Diabetes Prediction Dataset” by Marshal Patel. It is intended for use in predictive modeling, health research, and educational purposes.

Data Collection Process

The dataset aggregates health-related data through clinical records and survey responses, combining objective medical measurements with self-reported lifestyle and demographic information. The exact collection methodology is not fully detailed but appears to simulate real-world patient data.

Data Structure

The dataset is structured in a flat tabular format where:

- Rows represent individual patient records, and
- Columns represent variables such as age, BMI, glucose level, cholesterol, smoking status, and physical activity level.

Important Features and Their Significance

1. **BMI & Waist Circumference:** Measures of body fat, linked to metabolic disorders like diabetes.
2. **Fasting Blood Glucose & HbA1c:** Primary indicators for assessing blood sugar control and diagnosing diabetes.
3. **LDL & HDL Cholesterol:** Cardiovascular markers affected by metabolic health and lifestyle choices.

4. **Physical Activity Level & Smoking Status:** Modifiable risk factors that directly impact health outcomes.
5. **Ethnicity & Age:** Demographic variables known to influence disease susceptibility and risk profiles.

Data Limitations or Biases

1. **Missing or Incomplete Values:** May affect the accuracy of some summary statistics and correlations.
2. **Simulated or Aggregated Data:** As it's not drawn from a live clinical population, real-world variability might be underrepresented.
3. **Reporting Bias:** Lifestyle factors are often self-reported and may be subject to inaccuracies.
4. **Lack of Time-Series Data:** The dataset captures a single point in time, limiting longitudinal insights.

Data Splitting and Preprocessing

Data Cleaning

Minimal cleaning was required, as the dataset was well-structured and consistent. The only data cleaning step taken was formatting the column headings in Excel for readability and standardization (e.g., adjusting capitalization and spacing).

Handling Missing Values

There were no missing values in the dataset, so no imputation or deletion was necessary. This ensured a smooth analysis process without data integrity concerns.

Data Transformations

Data transformation involved the use of IF and IFS statements in Excel to classify continuous health metrics into meaningful clinical categories. Specifically:

- BMI values were categorized into underweight, normal, overweight, and obese ranges.
- Fasting Blood Glucose levels were classified into hypoglycemia, normal, prediabetes, diabetes, and hyperglycemia.
- HbA1c values were grouped into normal, prediabetes and diabetes ranges for individuals with diabetes.

Data Splitting

- Dependent variables included clinical metrics such as Fasting Blood Glucose, HbA1c, Blood Pressure, and Cholesterol (LDL and HDL).
- Independent variables consisted of demographic and lifestyle indicators such as BMI, Waist Circumference, Ethnicity, Smoking Status, Physical Activity Level, and Age.

Industry Context

The dataset pertains to the healthcare industry, with relevance to clinical research, public health monitoring, and chronic disease prevention—particularly focused on diabetes.

Stakeholders

- **Patients** – Want to manage or prevent diabetes and related conditions.
- **Healthcare Providers** – Need data to assess risks and personalize treatments.
- **Public Health Agencies** – Use it for epidemiological trends and policy planning.
- **Insurance Companies** – Use this data to manage health risk and premiums.
- **Researchers** – Use it to develop models and identify risk factors.

Value to the Industry

By examining the relationship between lifestyle, demographic, and clinical factors and their influence on diabetes indicators, this analysis contributes to more informed healthcare decisions. It supports early identification of at-risk individuals and helps shape effective public health policies, ultimately reducing the burden of diabetes and improving population health outcomes.

Pre-Analysis

Key Trends

1. Individuals with low physical activity and high BMI consistently show elevated fasting glucose, indicating a strong interaction between lifestyle and metabolic health.
2. Heavy alcohol consumption and smoking are linked with higher GGT levels and elevated blood pressure, increasing cardiovascular risk.
3. A waist circumference above 100 cm appears to trigger a sharp rise in HbA1c, regardless of BMI, suggesting central obesity as a more critical factor.

Potential Correlations

1. Asian ethnicity combined with low physical activity results in disproportionately high fasting glucose, implying ethnicity may act as a genetic modifier.
2. Moderate BMI in Asian individuals may still be associated with elevated HbA1c, hinting at a potential genetic predisposition.
3. For the same BMI range, males tend to have higher LDL and lower HDL, suggesting gender-based metabolic differences.
4. High dietary intake (above 3000 calories) correlates with increased serum urate and LDL, indicating links to both gout and metabolic syndrome.

Initial Insights

1. Women with a history of gestational diabetes tend to maintain elevated HbA1c years later, emphasizing the need for long-term monitoring.
2. Serum urate levels are often elevated in males with high-calorie diets, flagging risk for gout and insulin resistance.
3. Low physical activity, excessive calorie intake and alcohol use can increase the risk of diabetes and cardiovascular complications.

In-Analysis

Unconfirmed Insights

These findings emerged from initial analysis but require further investigation and possibly statistical validation:

Waist Circumference vs. Fasting Glucose as HbA1c Predictors

A surprising observation is that across a wide range of HbA1c levels (from 4.4 to 13.9), waist circumference remains relatively constant (97–98 cm). However, fasting glucose values vary, though not consistently. This challenges the typical assumption that HbA1c closely tracks with waist size and suggests that fasting glucose may be a more sensitive immediate indicator of glucose regulation than waist circumference. Further regression analysis could clarify the predictive strength of each variable.

HbA1c by Ethnicity

While all groups showed elevated HbA1c, Hispanics and Blacks exhibited slightly higher average values (9.6), followed by Asians (9.5) and Whites (9.4). The differences are marginal but consistent, suggesting possible genetic or cultural dietary influences. However, due to the narrow range, statistical significance testing is needed to confirm this pattern.

Ethnicity, Calorie Intake, and Cholesterol

Despite modest dietary and cholesterol differences by ethnicity (e.g., Hispanics: 2749 kcal, 225.5 cholesterol vs. Asians: 2736.1 kcal, 224.3 cholesterol), the overall spread is minimal, implying ethnicity alone may not drive variation in these metrics. This opens the possibility that lifestyle and genetic predispositions have a stronger influence than diet alone.

Physical Activity and Blood Pressure

Systolic blood pressure slightly decreases as physical activity increases (from 134.5 to 133.9), though the change is subtle. This might indicate a small but consistent benefit of moderate exercise on cardiovascular health, warranting deeper time-series or longitudinal study.

Age vs. BMI

Across all age brackets, BMI remains fairly consistent, with a slight increase observed in the 60–64 age group. This implies that aging alone doesn't significantly change BMI, but rather that lifestyle changes, medication, or health conditions may become more influential over time.

BMI vs. Cholesterol

Contrary to expectations, cholesterol levels do not rise linearly with BMI. In fact, individuals with a low BMI (18.5) had surprisingly high cholesterol (241.9). This finding suggests that cholesterol is likely influenced by more than body mass alone, such as diet quality, genetics, or sedentary behavior. This challenges the oversimplified notion that higher BMI always equals higher cholesterol.

Age vs. Fasting Glucose

A modest decline in fasting glucose levels is observed in younger age groups. The 60–64 group had the highest average (136.5), while the 35–39 group had the lowest (132.3). Though the difference is not dramatic, it supports the idea that aging may impair insulin sensitivity or pancreatic function, contributing to increased fasting glucose.

Recommendations

Based on these insights, the following are preliminary recommendations for further exploration and policy consideration:

1. **Healthcare Screening Focus:** Older adults and individuals with consistently high HbA1c but average waist circumference should be prioritized for early diabetes screening, even if they appear "normal" by visual or anthropometric standards.
2. **Targeted Awareness Campaigns:** Ethnic groups with elevated HbA1c—particularly Hispanics and Blacks—may benefit from tailored nutritional and fitness interventions that respect cultural contexts.
3. **Refinement of Risk Models:** The non-linear relationship between BMI and cholesterol calls for multi-variable risk assessment models that include physical activity, genetics, and diet patterns, rather than BMI alone.
4. **Promotion of Moderate Physical Activity:** Even minor drops in blood pressure associated with increased activity levels indicate a worthwhile public health investment in encouraging daily movement.

Analysis Techniques Used in Excel

To conduct this analysis, several Excel functionalities were leveraged effectively:

- **Pivot Tables:** These were the backbone of our data slicing. They allowed the segmentation of data by demographic groups (e.g., age, gender, ethnicity) and facilitated rapid aggregation of values such as average HbA1c, BMI, or cholesterol.
- **Charts:** Bar graphs, column charts, stacked bar charts, area charts were used to visualize relationships and trends across key health indicators. These visuals made it easier to detect patterns and communicate findings clearly.
- **Data Filters and Sorting:** Excel’s built-in sort/filter tools were employed extensively to isolate specific groups or metric ranges, which enhanced clarity during exploratory analysis.

Post-Analysis and Insights

Key Findings

1. Fasting glucose levels have a stronger influence on HbA1c than waist circumference. Despite stable waist sizes, HbA1c fluctuated in tandem with fasting glucose, suggesting the latter is a more direct predictor of glycemic control.

2. Hispanic and Black individuals exhibit slightly higher HbA1c levels, indicating a potentially elevated long-term risk for type 2 diabetes within these groups. This points to a role of genetics, access to healthcare, or lifestyle factors.

3. Cholesterol levels are highest among Black individuals despite similar calorie and cholesterol intake across ethnicities. This suggests that genetics or metabolic factors may influence cholesterol more than diet alone.

4. Systolic blood pressure decreases modestly with higher physical activity. This highlights the cumulative cardiovascular benefits of regular, even moderate, exercise.

5. BMI tends to increase through midlife before slightly declining in older age groups, with an overall average of 29.4. This indicates a general trend toward overweight status, presenting long-term metabolic health risks.

6. Individuals with higher BMI sometimes display lower cholesterol levels, highlighting that BMI alone may not reliably predict cholesterol or diabetes risk.
7. Fasting glucose levels increase with age, peaking at ages 60–64. This group had the highest average glucose levels (136.5 mg/dL), emphasizing the need for early screening and intervention in older adults.

Comparison with Initial Findings

1. **HbA1c and Glucose vs. Waist Size:** Initially, it was expected that waist circumference would play a significant role in HbA1c variation. However, the data showed that fasting glucose had a more pronounced effect, shifting focus away from body measurements toward metabolic indicators.

2. **Ethnicity and Health Indicators:** The assumption was that diet might explain differences in cholesterol and HbA1c across ethnicities, but the findings suggest a stronger role for non-dietary factors like genetics or healthcare access, particularly for Black individuals who had higher cholesterol despite similar diets.

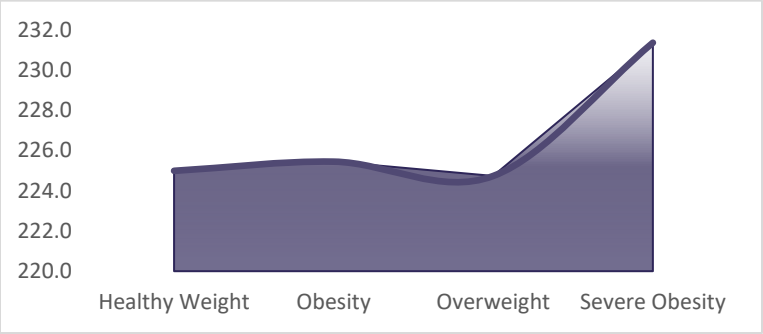
3. **BMI and Cholesterol:** Contrary to the common belief that higher BMI correlates with higher cholesterol, the analysis found a slightly inverse relationship, which was unexpected and points to the complexity of metabolic health.

4. **Physical Activity and Blood Pressure:** As expected, physical activity was associated with lower systolic pressure, but the effect was more modest than anticipated, suggesting the need for additional factors in hypertension control.

5. **Age and Fasting Glucose:** The steady rise in fasting glucose with age aligns with initial expectations. However, the notably higher levels in the 60–64 age group highlight a critical point for early screening efforts, more than initially assumed.

Data Visualizations & Charts

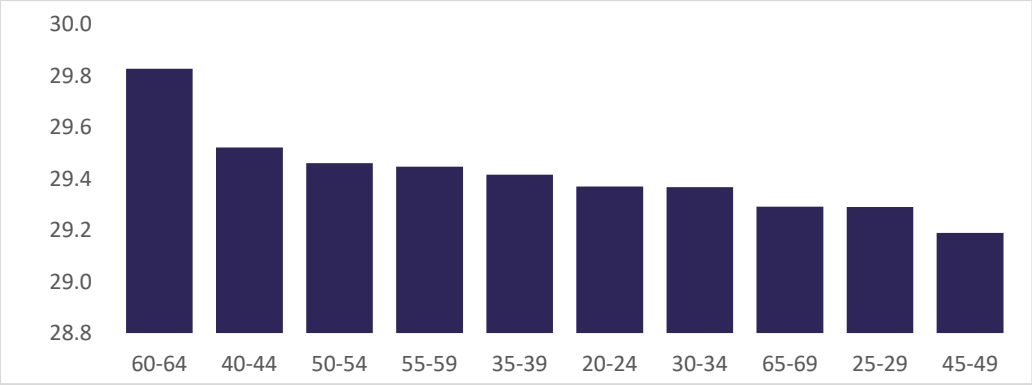
1. BMI By Average Cholesterol Levels



Explanation

Individuals with a healthy weight, overweight, and obesity have similar average cholesterol levels, around 225 mg/dL. However, those classified under severe obesity exhibit a notably higher average cholesterol level of 231.4 mg/dL. This suggests that while cholesterol levels remain stable across most BMI ranges, severe obesity may be linked to elevated cholesterol and increased cardiovascular risk.

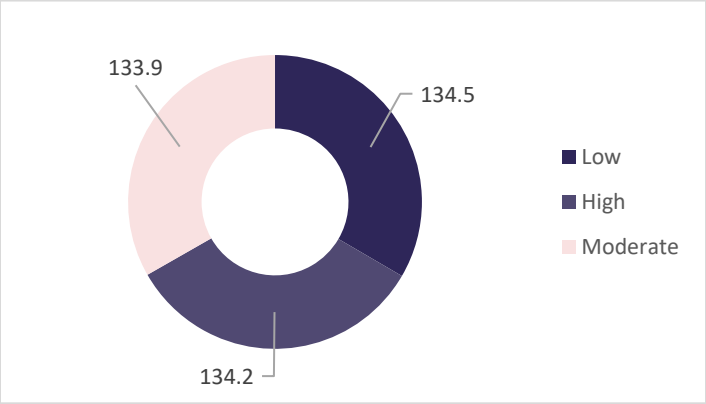
2. **Age Group By BMI Group**



Explanation

BMI tends to remain stable across ages but slightly increases in older adults, suggesting age-related weight gain and the need for age-targeted weight management strategies.

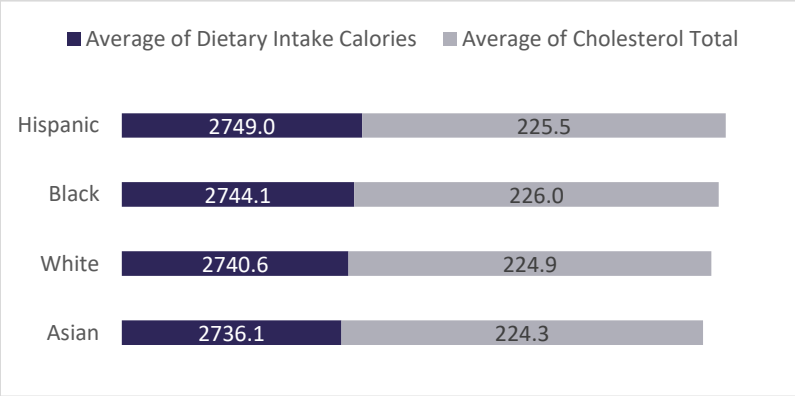
3. Physical activity on Blood Pressure



Explanation

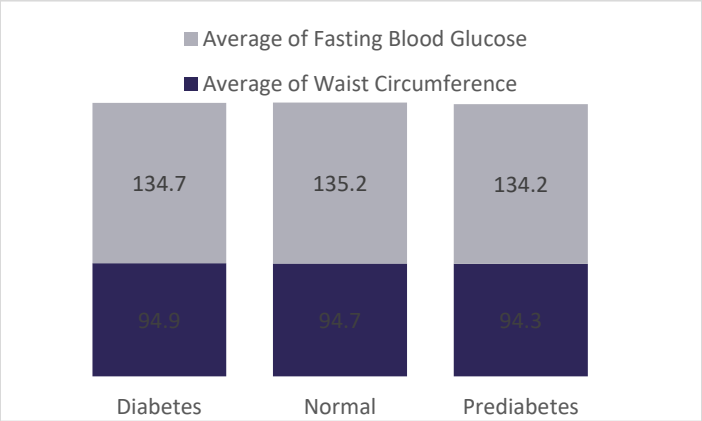
The chart shows the average systolic blood pressure for different levels of physical activity: low activity has an average of 134.5 mmHg, high activity is slightly lower at 134.2 mmHg, and moderate activity results in the lowest average at 133.9 mmHg.

4. Ethnicity Vs. Calorie Intake Vs. Cholesterol



The chart shows average dietary intake and cholesterol levels by ethnicity. Hispanics have the highest average calorie intake (2749), followed by Blacks (2744.1), Whites (2740.6), and Asians (2736.1). Cholesterol levels are similar across groups, ranging from 224.3 to 226.0 mg/dL.

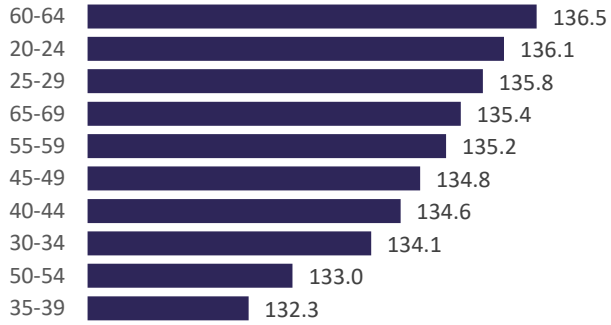
5. HbA1c Predictor: Waist Circumference or Fasting Glucose



Explanation

The average waist circumference and fasting blood glucose levels are quite similar across all three HbA1c classes, with slight differences in waist size being observed. Blood glucose levels hover around 134 mg/dL, regardless of the HbA1c class.

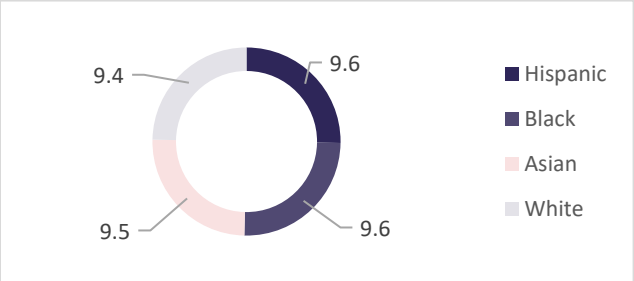
6. **Age Group By Fasting Glucose**



Explanation

Fasting blood glucose levels are relatively consistent across age groups, with the highest average observed in the 60-64 age group (136.5 mg/dL) and the lowest in the 35-39 age group (132.3 mg/dL). There is a slight trend of decreasing blood glucose levels as age decreases, but the differences are minimal.

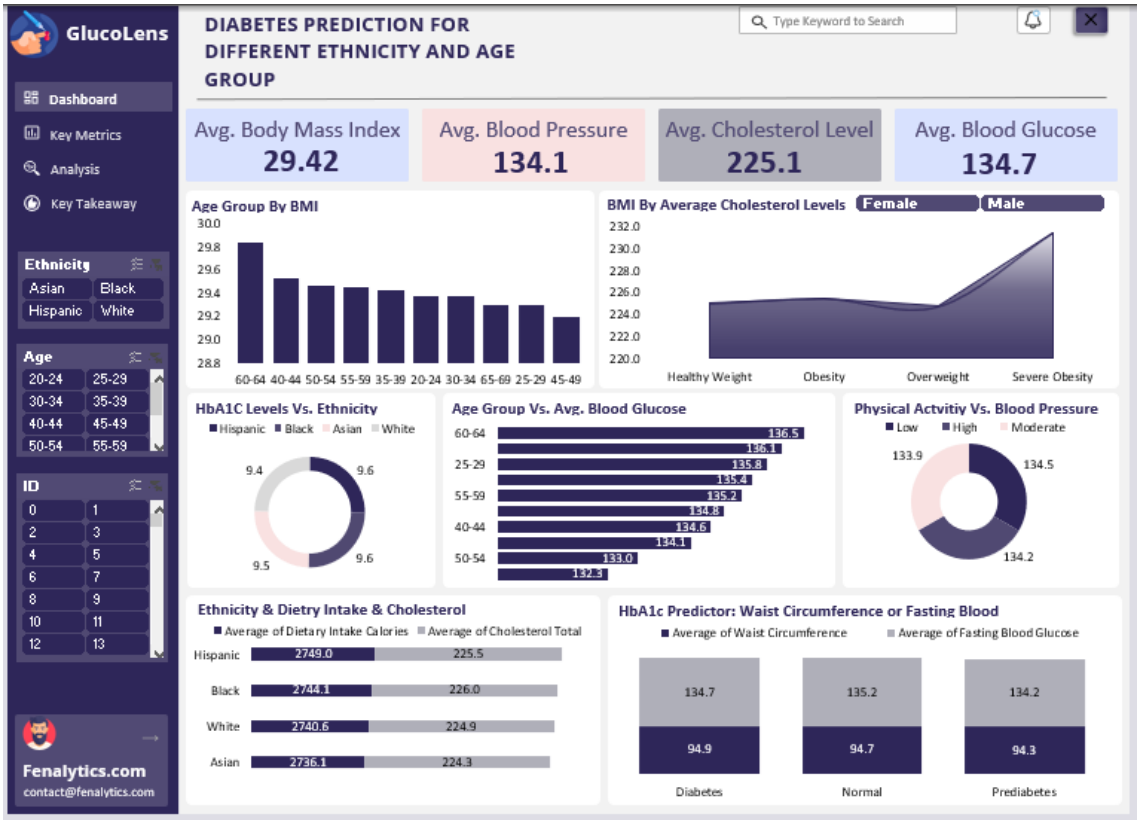
7. **HbA1c levels across different ethnicities**



Explanation

The average HbA1c levels are similar across ethnic groups, with Hispanics and Blacks having the highest average at 9.6, followed closely by Asians at 9.5 and Whites at 9.4. This suggests a slight variation in HbA1c levels among ethnicities, but the differences are minimal.

8. Final Dashboard



Recommendations and Observations

Actionable Insights

1. Focus predictive efforts on fasting glucose monitoring rather than waist circumference for early glycemic risk detection.

2. Introduce tailored diabetes prevention and management strategies for higher-risk ethnic groups through culturally relevant education and community outreach.

3. Develop personalized dietary interventions by integrating genetic, lifestyle, and metabolic profiling.

4. Promote moderate, consistent physical activity via accessible programs like walking groups or workplace challenges.

5. Implement early weight management interventions for adults in their 30s and 40s to reduce long-term metabolic risks.

6. Use a comprehensive risk assessment approach that includes cholesterol, diet, activity level, and family history instead of relying solely on BMI.

7. Initiate fasting glucose screenings starting at age 40, and introduce lifestyle improvement programs through local centers.

Optimizations For Business Decisions

1. Allocate resources to community-based screenings and educational initiatives in high-risk populations.

2. Invest in data infrastructure and AI tools to support personalized health profiling.

3. Shift program design from generic wellness to targeted behavior change strategies based on age and risk factors.

4. Integrate interdisciplinary health teams to deliver more holistic care.

Unexpected Outcomes

1. Despite assumptions, waist circumference and fasting glucose levels were nearly identical across HbA1c categories, suggesting it may not be a reliable early indicator alone.

2. The minor differences in HbA1c levels across ethnic groups were smaller than expected, which may reflect sampling bias or uniformity in clinical care access.

3. Surprisingly, fasting glucose levels peaked in the 60–64 age group but were also high among younger adults, highlighting the need for earlier screening.

Conclusion

Key Learnings

1. Fasting blood glucose is a more reliable indicator of glycemic status than waist circumference across HbA1c classes.

2. Ethnic disparities in HbA1c exist, with Hispanic and Black individuals showing slightly higher averages, indicating the need for tailored interventions.

3. Moderate physical activity is associated with the lowest systolic blood pressure, highlighting its cardiovascular benefit.

4. Adults in their 30s and 40s show lower fasting glucose, suggesting this is a critical window for preventive strategies.

5. Average dietary intake and cholesterol levels vary minimally across ethnicities, signaling similar nutritional exposures or reporting behaviors.

Limitations

1. The dataset lacks contextual variables such as medication use, socioeconomic status, or healthcare access, which could influence outcomes.

2. Sample size and representation across ethnic and age groups were not specified, which may affect generalizability.

3. No longitudinal data was available, limiting the ability to assess trends over time or causality.

Future Research

1. Include additional biomarkers (e.g., insulin levels, triglycerides) for a more comprehensive metabolic risk profile.

2. Explore longitudinal trends to understand how glycemic markers evolve with age, lifestyle, and intervention.

3. Segment data by gender, income level, and geographic region to identify more granular disparities and opportunities.

4. Incorporate behavioral and psychosocial factors to better understand barriers to effective health interventions.

References

National Health Service (2023). Obesity. [online] NHS. Available at: https://www.nhs.uk/conditions/obesity/.

The Scan Clinic - Private Ultrasound London. (2023). Your Helpful Guide To The HbA1c Test, And HbA1c Normal Range. [online] Available at: https://thescanclinic.co.uk/health/your-helpful-guide-to-the-hba1c-test-and-hba1c-normal-range/.

Seery, C. (2022). Normal and Diabetic Blood Sugar Level Ranges - Blood Sugar Levels for Diabetes. [online] Diabetes.co.uk. Available at: https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html.