Initial Transformation Plan:

To enhance the predictive strength and interpretability of the diabetes classification model, a series of well-reasoned data transformations were applied to the Diabetes Health Indicators dataset.

- Binary Variable Handling: Binary variables such as HighBP, HighChol, Smoker, and PhysActivity were retained as-is, given their native 0/1 structure. These were reviewed for clarity in labeling to ensure downstream interpretability in visualizations and modeling.
- Binning Age Groups: The original categorical age brackets (1–13) were collapsed into broader life stages: 18–34, 35–54, and 55+. This binning aligns with public health segmentation and simplifies model complexity while maintaining interpretive value.
- Scaling Continuous Variables: Features like BMI, MentHlth, and PhysHlth were standardized to accommodate algorithms sensitive to feature magnitude, such as logistic regression and support vector machines. This also prevents overweighting by high-variance variables.
- Re-labeling Ordinal Variables: Variables such as GenHlth, Education, and Income, though numerically encoded, were re-coded with descriptive labels to better reflect their ordinal nature and to avoid misinterpretation as continuous features.

Engineered Features:

Two engineered features were created to surface deeper insights relevant to diabetes risk, grounded in domain knowledge:

- **Chronic Risk Load**

  Logic: Aggregates metabolic and cardiovascular comorbidities including HighBP, HighChol, Stroke, and HeartDiseaseorAttack into a single composite risk score.

  Construction: Summed the presence of these conditions (binary indicators) for each respondent.

  Value: Offers a cumulative measure of chronic burden, supporting risk stratification in predictive modeling.
- **Healthcare Barrier Index**

  Logic: Identifies structural obstacles to preventive care, particularly cost and insurance coverage.

  Construction: Combined AnyHealthcare and NoDocbcCost into an index (range 0–2).

Value: Captures latent socioeconomic factors that impact access to diagnosis and care—factors often underrepresented in clinical datasets.

Feature Selection Strategy:

Feature refinement was guided by a balance of statistical relevance and public health logic:

- Correlation Analysis: Assessed potential redundancy, such as between MentHlth and PhysHlth. Although moderately correlated, both variables were retained due to their conceptual independence.
- Variance Filtering: Features with very low variance (e.g., CholCheck, which was overwhelmingly populated with '1') were flagged as candidates for exclusion based on low discriminatory power.
- Domain Relevance: Priority was given to features like BMI, Income, DiffWalk, and Education, which are known determinants of diabetes risk, even if their statistical correlation was modest.

AI Feedback:

I consulted ChatGPT to evaluate whether my transformations were statistically sound and clinically meaningful. The key takeaways were:

- Binning Health Days: Suggested binning MentHlth and PhysHlth to reduce right skew and enhance interpretation. I implemented binning into four categories (0, 1–10, 11–20, 21–30).
- Placeholder Detection: Recommended verifying values such as 77, 88, and 99 as potential encodings for missing data. Upon inspection, these were not present in the dataset.
- Disaggregating Risk Index: Encouraged experimenting with both the composite Chronic Risk Load and its individual components to evaluate which form yields stronger predictive performance. I chose to retain both forms for future comparison during modeling.

What I'd Like Feedback On:

- Does the Chronic Risk Load index adequately capture cumulative health burden, or would using individual indicators separately offer better transparency for stakeholders?
- Are there other potential interaction terms—perhaps involving Income, GenHlth, or DiffWalk—that might help uncover hidden subgroup patterns in diabetes risk?

- Would you recommend any dimensionality reduction techniques at this stage, or should I hold off until model performance testing begins?