### **"Good [morning/afternoon], everyone. We are Group 9, and today, we are excited to present our project on Diabetes Health Indicators. This project focuses on analyzing health-related data to predict diabetes risk, identify key contributing factors, and explore actionable insights for public health interventions.**

### **Our team—Aishwarya Balmoori, Krishna Priyanka Challa, Peihua Tsai, and I - Tara Canugovi—collaborated to leverage advanced data analytics techniques such as classification trees, logistic regression, and neural networks. We aim to demonstrate how data-driven methodologies can address one of the most pressing public health challenges globally: diabetes."**

### 

### **Description for Slide: Data Resource & Introduction**

"Diabetes remains one of the most pressing public health issues globally. Its prevalence has doubled over the past three decades, now affecting 1 in 10 Americans, with 1 in 5 unaware of their condition. Early detection and strategic interventions are vital to mitigating the rising economic and healthcare burden caused by diabetes, which exceeds $400 billion annually in the U.S.

For our analysis, we utilized the **2015 Behavioral Risk Factor Surveillance System (BRFSS)** dataset, sourced from Kaggle. This dataset is comprehensive, containing over **253,680 records and 21 health indicators**, providing robust data for analysis. From this dataset, we focused on **12 independent variables**, such as BMI, physical health, and age, to predict the target variable, *Diabetes\_binary*. This variable classifies individuals into two categories: non-diabetic or diabetic.

By leveraging this data, our goal was to build predictive models that not only identify key diabetes risk factors but also provide actionable insights for public health strategies. Through techniques like classification trees, neural networks, and logistic regression, this project highlights the power of data-driven methods in addressing critical healthcare challenges."

=====================================

classification tree analysis

"To simplify our model and improve its interpretability, we applied Correspondence **Analysis (CA)** to the categorical variables *GenHlth*, *PhysHlth*, and *DiffWalk*. **CA** is a statistical technique that analyzes and visualizes the relationships between categorical variables in a contingency table.

Through this process, we found that:

* *GenHlth* contributed the most variance at **314.596**.
* *PhysHlth* contributed **189.771**, and *DiffWalk* contributed even less.

Based on these results, we decided to **remove PhysHlth and DiffWalk**, simplifying the model without significantly sacrificing predictive power.

After dimension reduction, we repeated the classification tree analysis:

1. With a **complexity parameter (CP) of 0.001**, the model achieved a **training accuracy of 86.55%** and a **validation accuracy of 86.51%**.
2. We then adjusted the CP to **0.0005**, which slightly increased the **training accuracy to 86.61%** and the **validation accuracy to 86.52%**.

**Key Insight:**Removing low-contributing variables did not significantly impact the model’s performance, with validation accuracy decreasing by less than 0.1%. This validated our decision to simplify the model, prioritizing efficiency and real-world applicability."

### **2-Minute Explanation: Logistic Regression**

Lets come to the 2nd model i.e. Logistic regression

We trained a logistic regression model on a 60/40 train-test split to analyze diabetes risk factors. Key predictors include **high blood pressure (0.80)** and **high cholesterol (0.60)**, which significantly increase the likelihood of diabetes. Other factors, like **BMI (0.061)** and **Stroke (0.195)**, have smaller positive effects. The model's fit metrics—**residual deviance of 97,491.51** and **AIC of 97,517.51**—indicate strong performance. The model shows higher confidence in predicting non-diabetic cases (class 0) than diabetic cases (class 1), providing a reliable foundation for further analysis.

**Slide 2: ROC Curve and Predictive Power** The ROC curve measures the model's ability to distinguish between diabetic and non-diabetic cases. The x-axis represents the false positive rate, and the y-axis shows the true positive rate. The curve lying above the diagonal indicates that the model performs better than random guessing. With an **AUC of 0.818**, the logistic regression model demonstrates strong predictive power and a reliable ability to separate classes effectively.

—--------------------

The confusion matrix provides a breakdown of the model's predictions: it correctly classified **2,093 diabetic cases (TP)** and **85,408 non-diabetic cases (TN)**, while misclassifying **1,925 as diabetic (FP)** and **12,045 as non-diabetic (FN)**. This results in an overall accuracy of **86.53%**, meaning the model correctly predicts diabetes status for about 87% of the test dataset, indicating strong performance.

**Key Takeaway** "The logistic regression model provides interpretable and actionable insights, identifying key risk factors like high blood pressure and cholesterol while maintaining robust predictive accuracy."

### **3-Minute Explanation: Neural Network**

Then We trained a neural network with **10 neurons in the hidden layer**, a **regularization parameter of 0.1** to reduce overfitting, and up to **200 iterations** for convergence. The model's performance was evaluated using a confusion matrix and overall accuracy, showcasing its ability to predict diabetes risk. The structure was visualized with NeuralNetTools, highlighting the connections between the input, hidden, and output layers, providing insights into how the model processes data."

**This is the R output for confusion matrix and stats**

**=====================**

**Confusion matrix and Stats:**

**The neural network achieved an accuracy of 86.53%, correctly classifying 2,199 diabetic cases (TP) and 85,602 non-diabetic cases (TN). It has a recall of 87.75%, indicating strong performance in identifying diabetic cases. However, with 1,731 false positives (FP) and 11,939 false negatives (FN), the model exhibits a high false positive rate, suggesting the need to improve precision while maintaining its solid foundation for diabetes risk prediction.**

**Strengths and Challenges**"While the neural network achieved high accuracy and recall, it also exhibited a **high false positive rate**, meaning many non-diabetic cases were incorrectly classified as diabetic. This impacts the model's **precision**, which indicates the need for further fine-tuning.

**Key Takeaway:** The neural network provides a solid foundation for diabetes risk prediction with strong recall and accuracy. However, addressing the false positive rate is crucial to improving its real-world applicability, particularly in clinical or policy decision-making."

This explanation covers the model's setup, performance, and limitations, ensuring the audience understands both the benefits and areas for improvement.

—----------------------

3 comparison:

### **30-Second Summary: Comparative Model Performance**

"This table highlights the performance and trade-offs of the three models. **Logistic regression** achieved an accuracy of 86.53%, offering high interpretability with an AUC of 0.818 but moderate performance with complex data. **Classification tree**, slightly less accurate at 86.51%, provides simplicity and ease of use but lacks flexibility in advanced scenarios. **Neural networks** match logistic regression's accuracy and demonstrate high recall but suffer from a high false positive rate. Each model has unique strengths, suited for different use cases in diabetes risk prediction."

This summary captures the key comparisons and insights effectively.

==================

Managerial implications:

### This analysis highlights key predictors like high blood pressure, cholesterol, and BMI, enabling targeted prevention programs to reduce diabetes prevalence and costs. Simplified models like classification trees ensure practical application, even in resource-limited settings. High recall from neural networks supports early detection and effective screening policies, while logistic regression provides actionable insights for resource allocation. These findings also enhance public health campaigns by promoting lifestyle changes and supporting policy advocacy for increased funding in diabetes prevention and screening.

### 

### **Conclusion**

### **30-Second Conclusion**

"Our analysis of diabetes risk factors using the 2015 BRFSS dataset demonstrates the power of data-driven models in public health. The classification tree balanced simplicity and accuracy with an 86.52% validation accuracy. Logistic regression provided actionable insights with an AUC of 0.818, while the neural network achieved high accuracy but highlighted challenges with false positives.

Key predictors like high blood pressure, cholesterol, and BMI emphasize the importance of targeted interventions. This study showcases how predictive analytics can drive early detection and effective diabetes management strategies."

Thanks everyone