

# EARLY DETECTION OF DIABETES USING MACHINE LEARNING

## Problem Background

- Diabetes, a rapidly escalating global health issue, affected 422 million people worldwide as of 2018 (WHO).
- The number of people affected is expected to have increased significantly since then.
- Approximately 50% of all people with diabetes are undiagnosed due to its long-term asymptomatic phase.
- In the United States, 8.5 million people (23.0% of adults) are undiagnosed. (June 29, 2022)
- Early detection is vital for clinically meaningful outcomes, requiring careful assessment of both common and less common symptoms.

## The Challenge

- Currently, there is a lack of efficient methods to predict the likelihood of developing diabetes.
- The unique asymptomatic phase of diabetes presents a significant challenge for early detection and intervention.

# EARLY DETECTION OF DIABETES USING MACHINE LEARNING

## Our Approach ✨

- We used data mining classification techniques and machine learning models to predict the likelihood of developing diabetes.
- A dataset of 520 instances from Sylhet Diabetes Hospital in Sylhet, Bangladesh, collected via direct questionnaires, will serve as the basis for our model.
- Tools: Alteryx for data preparation and blending, Tableau for data visualization and exploration, and machine learning for predictive modelling and analysis.

## Objective

- To leverage the power of Alteryx, Tableau, and Machine Learning to create a robust risk prediction model.
- The ultimate goal is to facilitate the early detection of diabetes, thus enabling timely interventions and improved patient outcomes.

# Initial Data Dictionary

**Patient Details**

- Age [20-65]
- Sex [Male, Female]

**Binary Medical Condition Variables**

- Polyuria
- Polydipsia
- sudden weight loss .
- weakness
- Polyphagia
- Genital thrush
- visual blurring
- Itching
- Irritability
- delayed healing
- partial paresis
- muscle stiffness
- Alopecia
- Obesity

**Target Variable**

- Class (Positive/Negative)

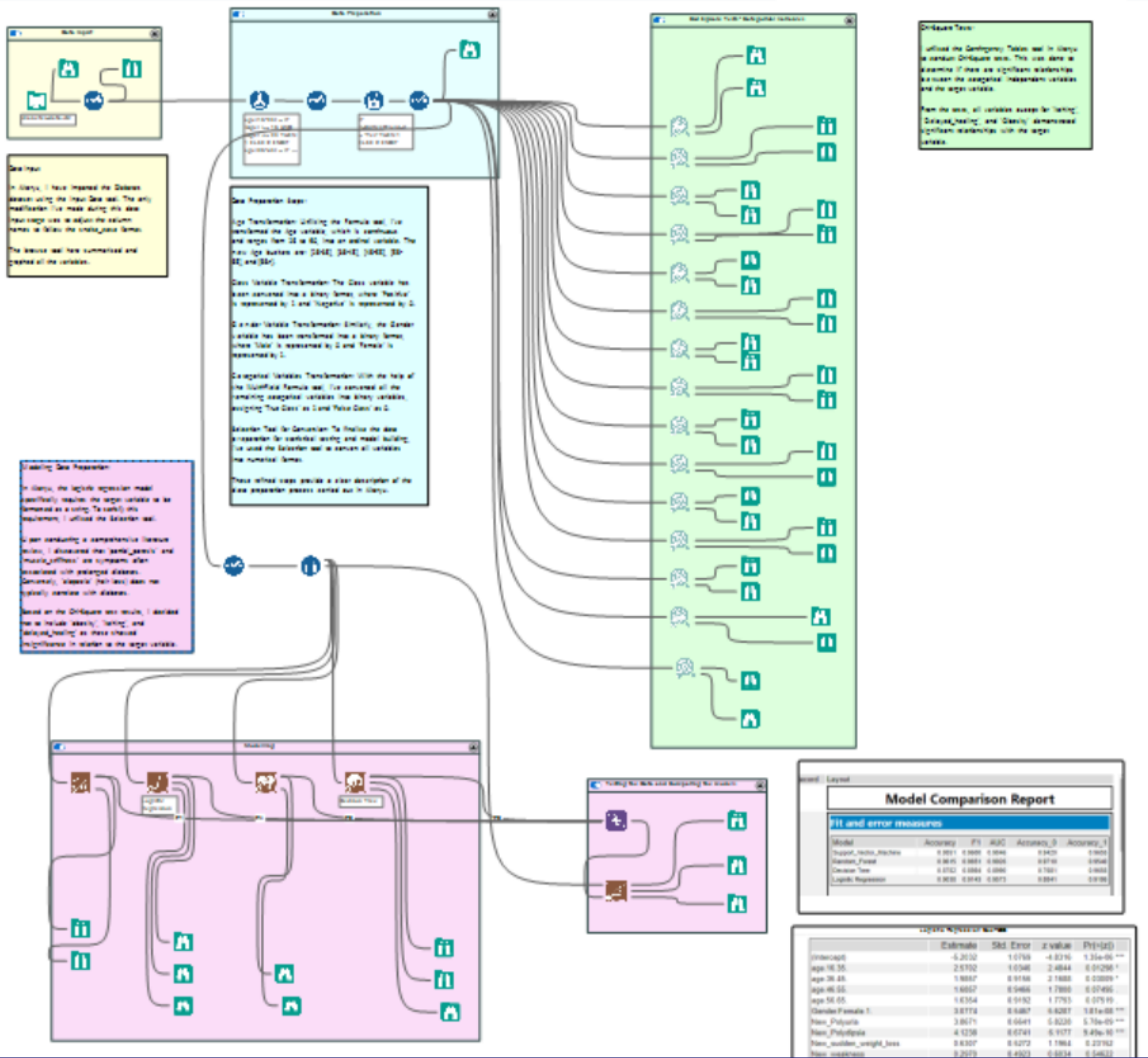
# Final Data used in Modeling

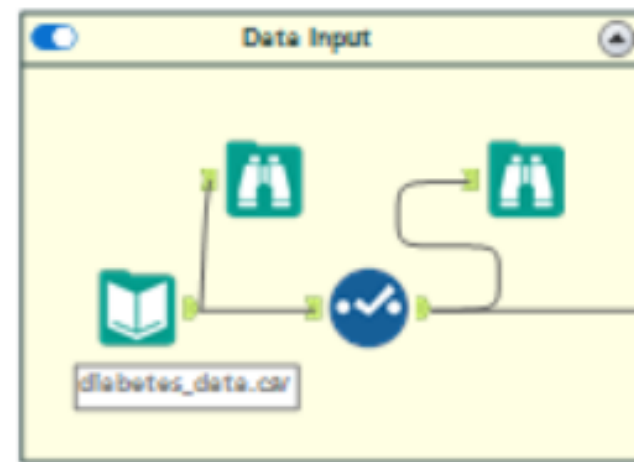
Upon conducting a comprehensive literature review, I discovered that 'partial\_paresis' and 'muscle\_stiffness' are symptoms often associated with prolonged diabetes. Conversely, 'alopecia' (hair loss) does not typically correlate with diabetes

Based on the Chi-Square test results, I decided not to include 'obesity', 'itching', and 'delayed\_healing' as these showed insignificance in relation to the target variable.

Patient Details	Binary Medical Condition Variables	Target Variable
<ul style="list-style-type: none"><li>• Age Ordinal Bucketing</li><li>• Sex [Male, Female]</li></ul>	<ul style="list-style-type: none"><li>• Polyuria</li><li>• Polydipsia</li><li>• sudden weight loss .</li><li>• weakness</li><li>• Polyphagia</li><li>• Genital thrush</li><li>• visual blurring</li><li>• Obesity</li></ul>	<ul style="list-style-type: none"><li>• Class (Positive/Negative)</li></ul>

# ALTERYX WORKFLOW

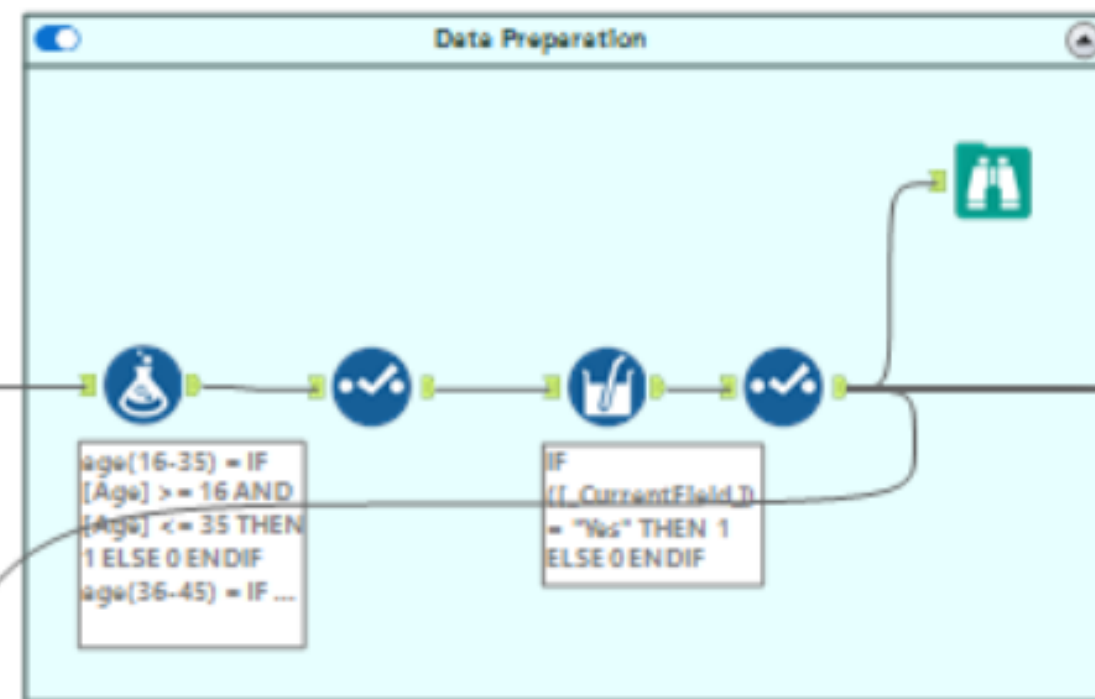




#### Data Input

In Alteryx, I have imported the Diabetes dataset using the Input Data tool. The only modification I've made during this data input stage was to adjust the column names to follow the snake\_case format.

The browse tool here summarized and graphed all the variables.



#### Data Preparation Steps:

**Age Transformation:** Utilizing the Formula tool, I've transformed the Age variable, which is continuous and ranges from 16 to 90, into an ordinal variable. The new Age buckets are: [16-35], [36-45], [46-55], [56-65], and [66+].

**Class Variable Transformation:** The Class variable has been converted into a binary format, where 'Positive' is represented by 1 and 'Negative' is represented by 0.

**Gender Variable Transformation:** Similarly, the Gender variable has been transformed into a binary format, where 'Male' is represented by 0 and 'Female' is represented by 1.

**Categorical Variables Transformation:** With the help of the Multi-Field Formula tool, I've converted all the remaining categorical variables into binary variables, assigning 'True Class' as 1 and 'False Class' as 0.

**Selection Tool for Conversion:** To finalize the data preparation for statistical testing and model building, I've used the Selection tool to convert all variables into numerical format.

These refined steps provide a clear description of the data preparation process carried out in Alteryx.

# DATA PREP ALTREYX WORKFLOW





# Final Models

Record

Layout

## Model Comparison Report

### Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Support_Vector_Machine	0.9551	0.9600	0.9846	0.9420	0.9655
Random_Forest	0.9615	0.9651	0.9926	0.9710	0.9540
Decision Tree	0.8782	0.8984	0.8990	0.7681	0.9655
Logistic Regression	0.9038	0.9143	0.9573	0.8841	0.9195

**MODEL PERFORMANCE ON THE TEST DATA. RANDOM FOREST IS THE ONE WITH HIGHEST AUC AND ACCURACY.**



# Logistic Regression | Coefficient's

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-5.2032	1.0769	-4.8316	1.35e-06	***
age.16.35.	2.5702	1.0346	2.4844	0.01298	*
age.36.45.	1.9857	0.9156	2.1688	0.03009	*
age.46.55.	1.6857	0.9466	1.7808	0.07495	.
age.56.65.	1.6354	0.9192	1.7793	0.07519	.
Gender.Female.1.	3.0774	0.5467	5.6287	1.81e-08	***
New_Polyuria	3.8671	0.6641	5.8228	5.78e-09	***
New_Polydipsia	4.1238	0.6741	6.1177	9.49e-10	***
New_sudden_weight_loss	0.6307	0.5272	1.1964	0.23152	
New_weakness	0.2970	0.4923	0.6034	0.54622	
New_Polyphagia	1.4431	0.5637	2.5601	0.01046	*
New_genital_thrush	0.9075	0.5362	1.6925	0.09055	.
New_visual_blurring	-0.4126	0.5158	-0.7999	0.4238	
New_Obesity	0.2036	0.6067	0.3355	0.73721	
age...66.	NA	NA	NA	NA	

Significance codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1 )