DIABETES ANALYSIS FOR PREGNANT WOMEN BACKED BY SUPERVISED MACHINE LEARNING MODEL USING SUPPORT VECTOR MACHINE ALGORITHM

Overview:

The National Institute of Health estimates that gestational diabetes makes an impact about 14% of pregnancies worldwide, and its incidence is rising in tandem with the rise of type 2 diabetes and obesity. Pregnancy – related diabetes, especially high blood glucose, puts the developing fetus and the mother at serious risk for health problems. In additions to influencing fetal growth and the child's long term metabolic health, poor glycemic management during pregnancy can result in issues such as pre-eclampsia, premature birth, and an increased risk of cesarean delivery. Early detection and balancing of glucose abnormalities during pregnancy are crucial due to rising incidence of diabetes caused by genetic and lifestyle factors. Effective clinical data analysis and early risk predictions are made possible by machine learning approaches, which improve maternal and newborn outcomes through prompt intervention and awareness.

Problem Statement:

Many pregnant women are unaware of the early warning signals and dangers of elevated glucose levels until difficulties occur, even with routine prenatal screening. Conventional diagnostic methods frequently depend on manual interpretation and predefined threshold which may miss intricate relationships between several clinical factors affecting diabetes risk. Preventive care and lifestyle changes may be delayed as a result of this lack of early predictive insight. In order to enable proactive management and better clinical decision making, a data-driven machine learning system that can precisely identify and predict diabetes risk in pregnant women using accessible health factors is needful.

Main Objective(s):

- Restrict the study group only to females (due to pregnancy) with their age from 21 to 60 and find their physiological and clinical data (Rare cases may be available above age 60).
- To train the supervised model with support vector machine classifier loading the medical data that predicts diabetes exposure for new data.

DATASET DESCRIPTION:

A. **Source of data:** Kaggle: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

B. **Structure of data:**

   a. **Rows:** 768 records.

   b. **Columns:** 9 features.

C. **Data columns:**

COLUMN NAME	DATATYPE	COLUMN DESCRIPTION
Pregnancies	Integer	Count of pregnancies by women
Glucose	Integer	Oral glucose tolerance test (OGTT – 2 hour plasma test). Value range: 50 to 200 in mg/dL
BloodPressure	Integer	Lower/Diastolic/bottom blood pressure (in mmHg)
SkinThickness	Integer	Thickness measure of triceps skin fold (in mm)
Insulin	Integer	OGTT 2 hour serum Insulin dosage (in µU/ml)
BMI	Float	Clinically measured Body Mass Index (in Kg/m^2)
DiabetesPedigreeFunction	Float	Value calculate based on the feature weightage. (between 0 and 1)
Age	Integer	Age of the women (in years)
Outcome	Integer	Number variable (0 for no diabetes, 1 for chances for diabetes)

D. **Key Variables:**

   a. **X – Independent variables:** Age, Pregnancies, Glucose, BloodPressure, BMI, DiabetesPedigreeFunction.

   b. **Y – Dependent variables:** Outcome.

Y DATA PROFILE:

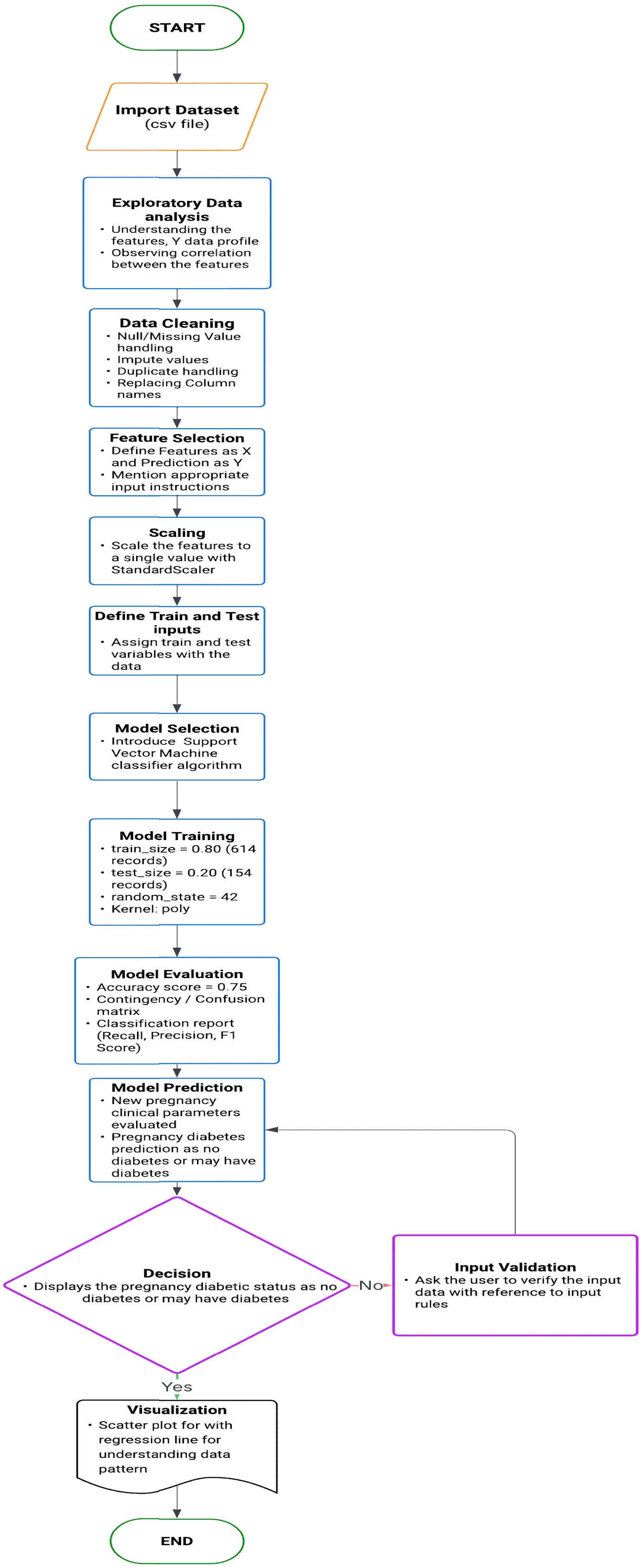
- Profile data:

Pima-Indians-Diabetes-yprofile.html

- Statistical description of the data

FEATURES	COUNT	MEAN	MEDIAN	MODE	STD	MIN	25%	50%	75%	MAX
PREGNANCIES	768	3.8451	3	1	3.3696	0	1	3	6	17
GLUCOSE	768	120.8945	117	99	31.9726	0	99	117	140.25	199
BLOOD PRESSURE	768	69.1055	72	70	19.3558	0	62	72	80	122
BMI	768	31.9926	32	32	7.8842	0	27.3	32	36.6	67.1
DIABETES PEDIGREE FUNCTION	768	0.4719	0.3725	0.254	0.3313	0.078	0.24375	0.3725	0.62625	2.42
AGE	768	33.2409	29	22	11.7602	21	24	29	41	81
OUTCOME	768	0.3490	0	0	0.4770	0	0	0	1	1

METHODOLOGY:


START
Import Dataset (csv file)
Exploratory Data analysis
• Understanding the features, Y data profile
• Observing correlation between the features
Data Cleaning
• Null/Missing Value handling
• Impute values
• Duplicate handling
• Replacing Column names
Feature Selection
• Define Features as X and Prediction as Y
• Mention appropriate input instructions
Scaling
• Scale the features to a single value with StandardScaler
Define Train and Test inputs
• Assign train and test variables with the data
Model Selection
• Introduce Support Vector Machine classifier algorithm
Model Training
• train_size = 0.80 (614 records)
• test_size = 0.20 (154 records)
• random_state = 42
• Kernel: poly
Model Evaluation
• Accuracy score = 0.75
• Contingency / Confusion matrix
• Classification report (Recall, Precision, F1 Score)
Model Prediction
• New pregnancy clinical parameters evaluated
• Pregnancy diabetes prediction as no diabetes or may have diabetes
Decision
• Displays the pregnancy diabetic status as no diabetes or may have diabetes
No
Input Validation
• Ask the user to verify the input data with reference to input rules
Yes
Visualization
• Scatter plot for with regression line for understanding data pattern
END

TRAINING – TESTING SPLIT:

Dataset training and testing size split details are as follows:

SPLIT	SIZE (Total = 1 (768 records)
Training	0.80 (614 records)
Testing	0.20 (154 records)

Split Description:

Training: Different blends of data points learned by the model to trace patterns of the data.

Testing: Left over data points are allocated to examine the model's work following the training before giving a new input.

MODEL SELECTION:

Model: Support Vector Machine Classification (Supervised Model)

Description:

- The outcome column is a class variable affected by multiple independent factors.
- The model learns the pattern for each factor trained to provide precise and anticipated results. Consequently, the system is quick and dependable.
- The result i.e the outcome column is intended to be a single variable correlated with several factors, hence support vector machine classifier model was implemented.

MODEL PARAMETERS:

Parameters Reference:

PARAMETERS	Values
test_size	0.20 (Lesser the better model learns).
random_state	42
Kernel	poly (polynomial curve)

Parameters Description:

- test_size : Collection of data to evaluate the model's effectiveness.

- random_state: By shuffling the data points, the model can be trained in many combinations. This guarantees the model's repeatability by learning data using random combinations.
- kernel: A mathematical function that transformed non linear values into data separable form helping the model to better understand the class differentiation.

MODEL TRAINING:

- Total Dataset used = 768 records
- Test size were listed and used test_size = 0.20. Hence, 614 records were adopted as test case and 154 records were assigned in training the model.
- Scaler used: StandardScaler. Standardizes multiple factors into a single value.
- To assist the model with learning multiple possibilities, random_state 42 was implemented.
- Kernel "poly" utilized as medical parameters interact non-linearly. This would plot the segments with curved boundaries comprising all the related features.
- Metrics such as accuracy score, classification report, confusion matrix were imported to evaluate the performance and accuracy of the model.

MODEL EVALUATION:

Sample data values obtained:

REFER_TEST PARAMETERS						**OBTAIN_PRED**
Age	**Pregnancies**	**Glucose**	**BloodPressure**	**BMI**	**DiabetesPedigreeFunction**	**Outcome**
43	6	98	58	34	0.43	0
21	2	112	75	35.7	0.148	0
21	2	108	64	30.8	0.158	0
34	8	107	80	24.6	0.856	0
50	7	136	90	29.9	0.21	0
55	6	103	72	37.7	0.324	0
22	1	71	48	20.4	0.323	0
44	0	117	0	33.8	0.932	1
37	4	154	72	31.3	0.338	0
65	5	147	78	33.7	0.218	1

EVALUATION METRICS:

Metrics used:

- **Confusion matrix**:
  - Confusion matrix represents a table visual that highlights number of predictions made correct and incorrect for each class by the model
  - The values are count based representation and used for other evaluation metrics.

		ACTUAL VALUES	
		POSITIVE (1)	NEGATIVE (0)
PREDICTED VALUES	POSITIVE (1)	TP	FP
	NEGATIVE (0)	FN	TN

Where:

$TP = True\ positive\ = correct\ prediction\ of\ positive\ classes$

$TN\ = True\ negative\ = correct\ prediction\ of\ negative\ classes$

$FP\ = False\ positive\ = incorrect\ prediction\ of\ positive\ classes$

$FN\ = False\ negative = \ incorrect\ prediction\ of\ negative\ classes$

  - True positive (TP) and true negatives (TN) provide the perfect right as right and perfect wrong as wrong prediction classes whereas false positive (FP) and false negative (FN) values show where the model predicts the right ones as wrong and vice versa.
- **Accuracy score**:
  - Used to measure the closeness of the measured value to the standard value. A single value that summarizes the whole model’s performance
  - The value is calculated by using the below formula:

$$Accuracy\ score = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

$TP = True\ positive\ = correct\ prediction\ of\ positive\ classes$

$TN\ = True\ negative\ = correct\ prediction\ of\ negative\ classes$

$FP\ = False\ positive\ = incorrect\ prediction\ of\ positive\ classes$

$FN\ = False\ negative = \ incorrect\ prediction\ of\ negative\ classes$

  - The higher value of accuracy score marks the model's performance at its best. Low accuracy score shows the model struggles to differentiate between classes.

- **Classification report**:
  - Provides comprehensive performance analysis such as recall, precision, F1 score and shows the behavior of each class.
  - Recall: Ratio of correct predicted positive among all actual positives
    - Formula used: $Recall = \frac{TP}{TP+FN}$
    - Helps to capture exact positive cases
  - Precision: Ratio of correct predicted positives among all predicted positives
    - Formula used: $Precision = \frac{TP}{TP+FP}$
    - Useful to mitigate false positives
  - F1-Score: Utilizes precision and recall and calculates their mean harmonically to balance the false positives and negatives into a single value.
    - Formula used: $F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$
    - Reduces false positives and false negatives

Evaluated Metric Values from the model:

- Confusion matrix:

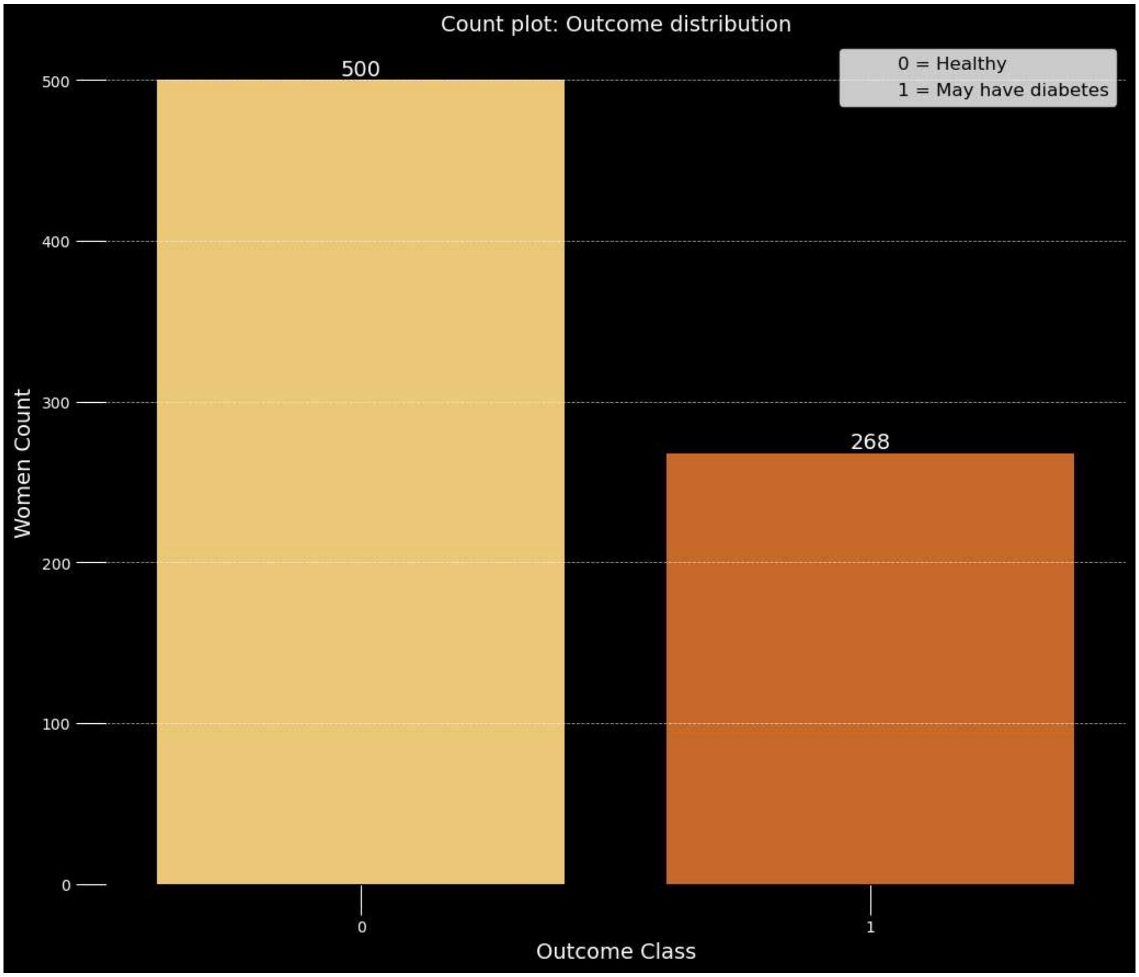
		ACTUAL VALUES	
		POSITIVE (1)	NEGATIVE (0)
PREDICTED VALUES	POSITIVE (1)	25	9
	NEGATIVE (0)	30	90

  - The contingency table utilized 154 test case record
  - 90 records were correctly predicted as negative (no diabetes) case, 9 false positive (have diabetes but no diabetes), 30 false negative (no diabetes but have diabetes), and 25 as correctly predicted as positive case (have diabetes)
  - Hence, the false positive and false negative has to be reduced further.
- Accuracy score = 0.75

- Classification report:
  - Recall: 0.91 (for no diabetes), 0.45 (for having chance of diabetes)
  - Precision: 0.75 (for no diabetes), 0.74 (for having chance of diabetes)
- F1 score: 0.82 (for no diabetes), 0.56 (for having chance of diabetes)

PLOTS:

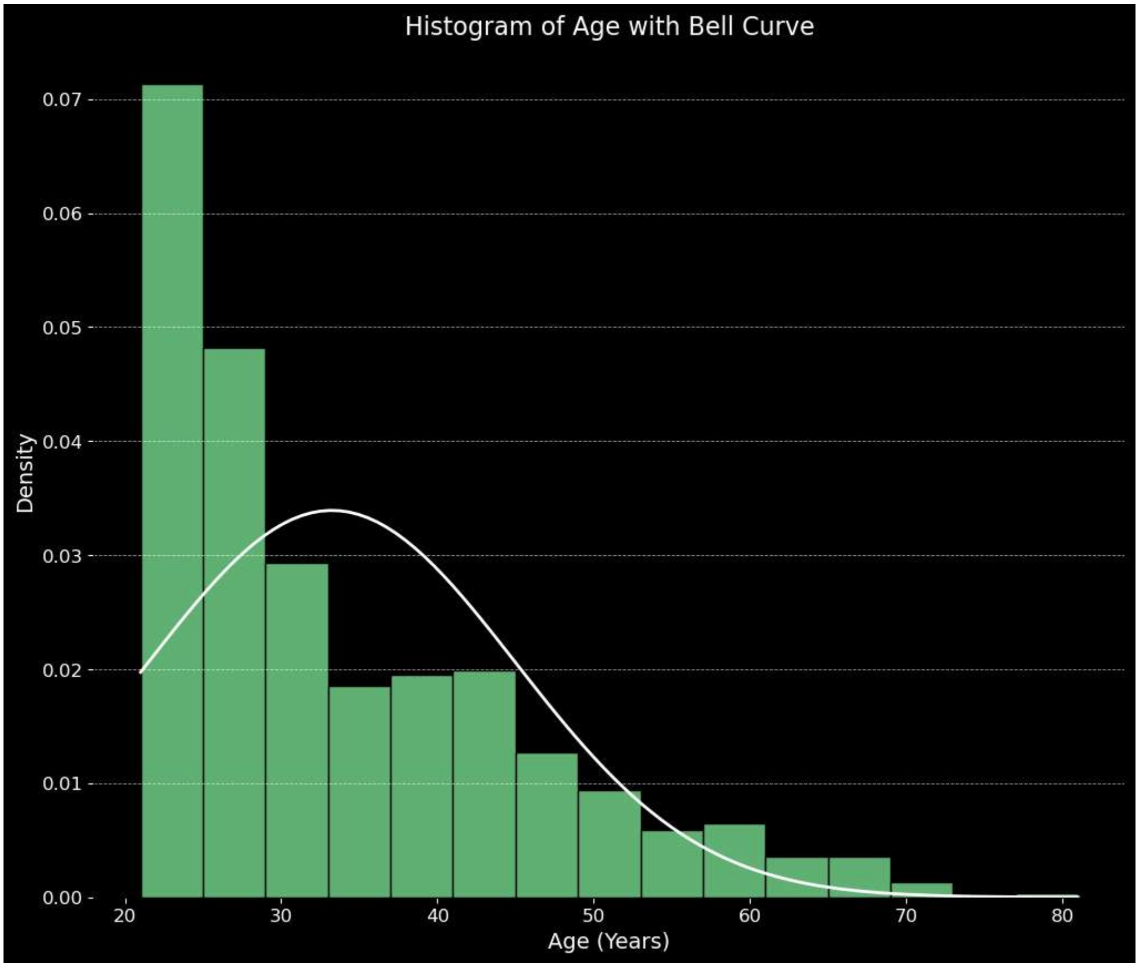
1) Count plot: Outcome distribution of women



PLOT INTERPRETATION:

- An unbalanced dataset is reflected from the count plot, which represents greater score of healthy pregnancies than diabetic risk cases.
- It is brought to the picture, that a significant percentage of women are classified as “May have diabetes“ alarming importance of diabetes screening for public health.

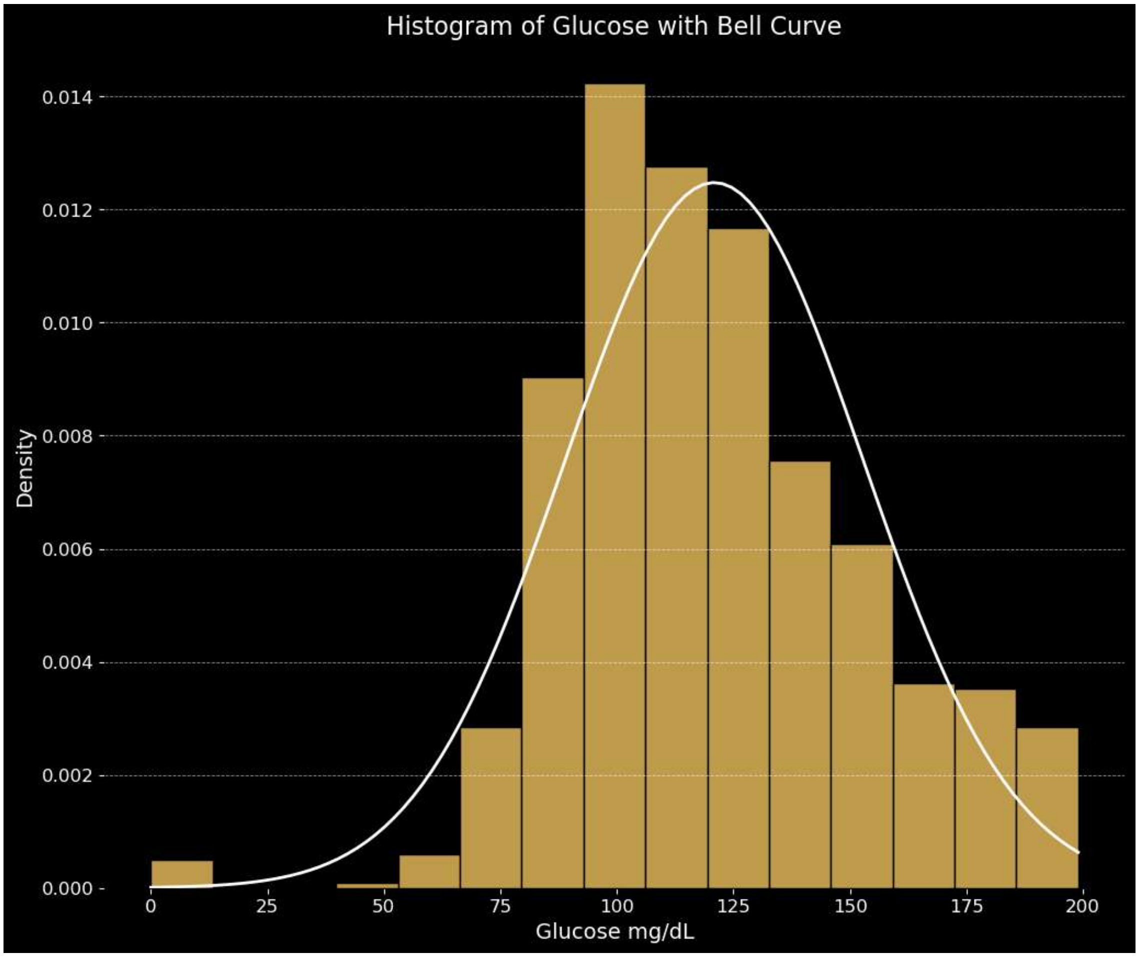
2) Histogram plot: Age distribution with bell curve



PLOT INTERPRETATION:

- From the age distribution, it is observed that the right skewed pattern suggests that older mothers are healthy than the middle aged groups.
- Heavy amount of younger to middle aged groups are prone to having diabetes during their pregnancy that represents the smooth blunt curve at the middle of the plot.
- Here, the distribution of age values illustrate as an important feature working with other clinical parameters.

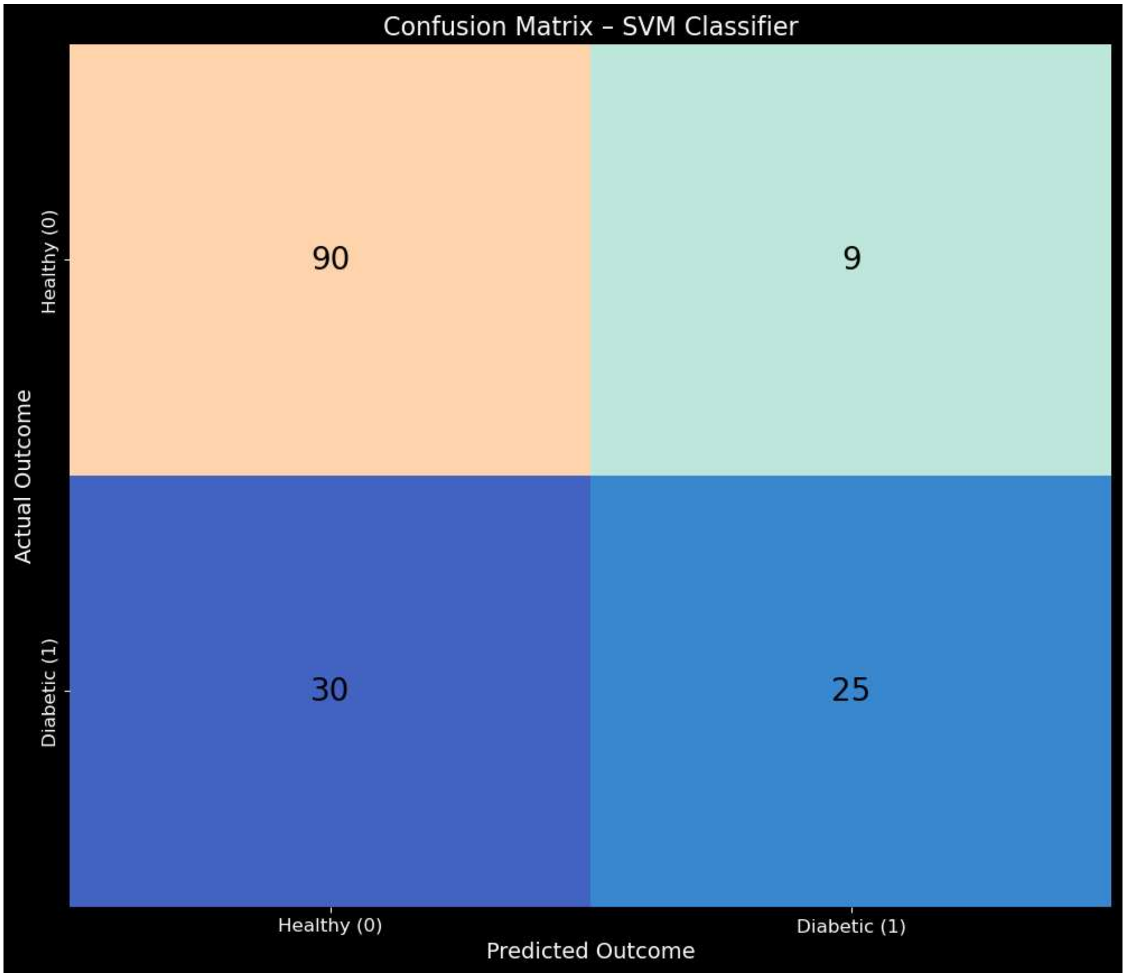
3) Histogram plot: Glucose distribution with bell curve



PLOT INTERPRETATION:

- The middle area of the plot with the noticeable right shift indicates that they have higher glucose level among a portion of total number of pregnant women.
- The plot highlights the peak of the bell curve as a highly depended clinical parameter characteristic that impacts pregnant women
- It proves that glucose is a crucial feature guiding model decision boundaries supported by this glucose distribution.
- Hence, Glucose level during pregnancy period directly affects the mother and the growing foetus to severe health risks.

4) Confusion matrix: SVM Classifier Model performance analysis



PLOT INTERPRETATION:

- As mentioned earlier in the evaluation section, the model demonstrated the capacity to find healthy pregnancies from the value 90 as true negatives
- Clinically, 9 cases might be found healthy, yet the model points them as diabetic pregnancy and on the other hand 30 cases were said healthy which would help mitigate and suggest to reconsider the consultation for better women pregnancy.
- Once confirmed with the specialist, the model could be improved to review the false predictions with added new records.

NEW DATA PREDICTION:

FEATURES	INPUT VALUES			ACCURACY
	SET 1	SET 2	SET 3	
AGE (in years)	26	31	35	0.98
PREGNANCY	1	2	3	
GLUCOSE (in mg/dL)	92	118	165	
BLOOD PRESSURE (in mmHg)	70	78	88	
HEIGHT (in cm)	158	155	152	
WEIGHT (in Kg)	54	66	78	
DIABETIC PEDRIGREE FUNCTION	0.21	0.45	0.78	
	BMI measure			
	No diabetes	May have diabetes	May have diabetes	

CONCLUSION:

In view of the above study, the support vector machine has successfully classified the existing data into the respective class variables and demonstrated the same for new input data. The skewed distributions of glucose and age levels showcased as one of the compulsory features while diagnosing diabetes for pregnant women. The plots represent significant patterns that emerged via exploratory data analysis, supporting its function as the main risk indicator. With an overall accuracy of 75%, the support vector machine classifier model may be used to predict diabetes risk in pregnant women, yet the model has to be monitored, verified by the specialist, and improvises whenever encountering an outlier data each time.