ST JOSEPH COLLEGE OF ENGINEERING

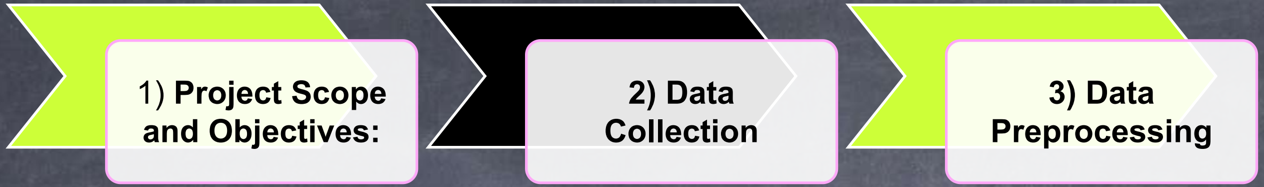
TITLE : AI-BASED DIABETICS PREDICTION MODEL PHASE-3

REPORTED BY SYED UZAIR AHMED
212921104052

PROJ_227128_TEAM_1

03

DEVELOPMENT
PART 1

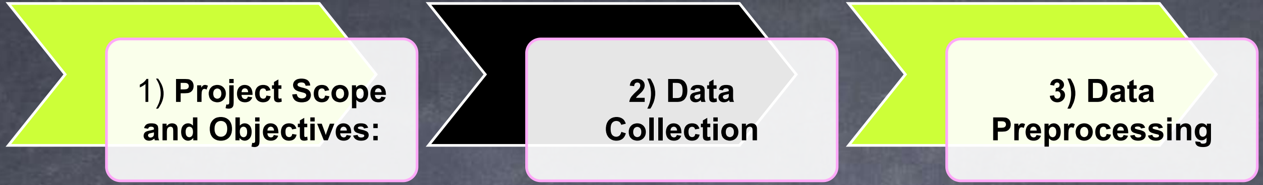


1) **Project Scope and Objectives:**

2) Data Collection

3) Data Preprocessing

1.) **Objective 1: Review existing literature on diabetes diagnosis and prediction.**
Objective 2: Develop a model using machine learning techniques.
Objective 3: Analyze the diabetes dataset and use Support Vector Machine and Random Forest algorithms to develop a prediction engine.
The scope of your project could include:
Developing a user-friendly interface for the prediction system.
Ensuring that the system is accurate and reliable.
Testing the system on a large dataset to ensure that it is effective in predicting diabetes.
Ensuring that the system is secure and that patient data is protected.



1) **Project Scope and Objectives:**

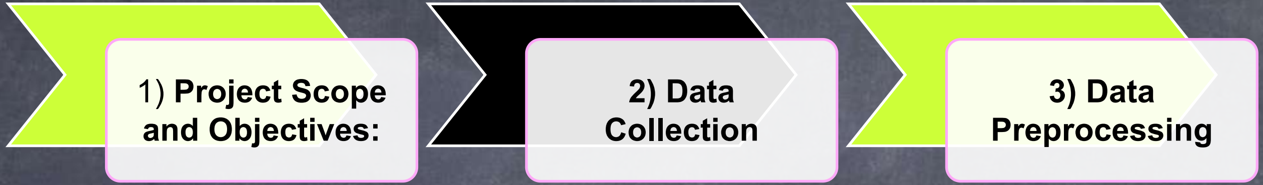
2) Data Collection

3) Data Preprocessing

2.)

There are several AI-based systems that use data collection to predict diabetes. One such system is a decision support system (DSS) that utilizes bidirectional long/short-term memory (BiLSTM)to accurately predict diabetic illness from patient data . Another example is an AI/ML-based medical device that has been approved by the US Food and Drug Administration for automatic retinal screening, clinical diagnosis support, and patient self-management tool.

data set link: https://www.kaggle.com/datasets/mathchi/diabetes-data-set



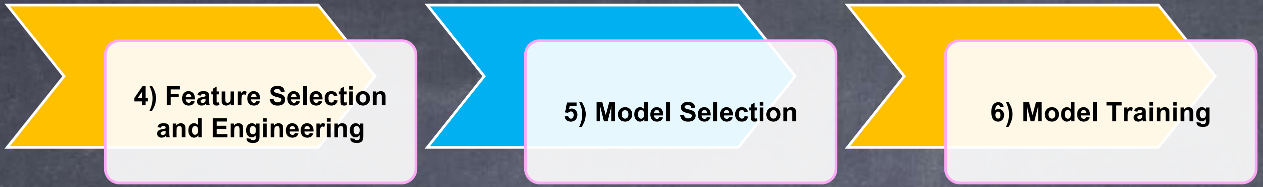
1) **Project Scope and Objectives:**

2) Data Collection

3) Data Preprocessing

Data Preprocessing: Cleaning and preparing the collected data for analysis. This step includes removing missing values, handling outliers, and normalizing the data.

data set link: https://www.kaggle.com/datasets/mathchi/diabetes-data-set

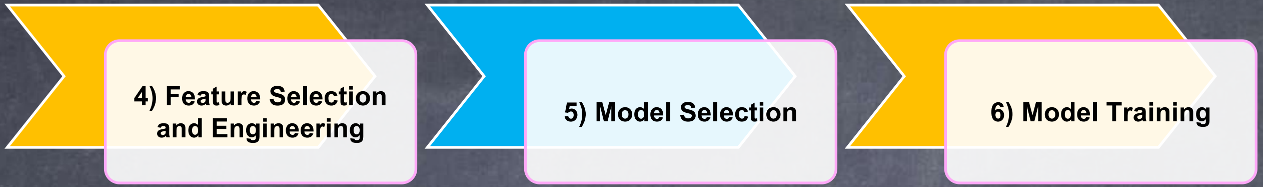


4) Feature Selection and Engineering

5) Model Selection

6) Model Training

The model contains the use of Boruta feature selection, the extraction of salient features from datasets, the use of the K-Means++ algorithm for unsupervised clustering of data, and stacking of an ensemble learning method for classification. The model was evaluated by accuracy, precision, and F1 index on the PIMA Indian diabetes dataset and achieved an accuracy rate of 98% .Another study focuses on data-driven diabetes risk factor prediction using machine learning techniques. The study applies two-fold feature selection techniques (i.e., principal component analysis, PCA, and information gain, IG) to boost the prediction accuracy. Then, the optimal features are fed into five ML algorithms, namely decision tree, random forest, support vector machine, logistic regression, and KNN [2].

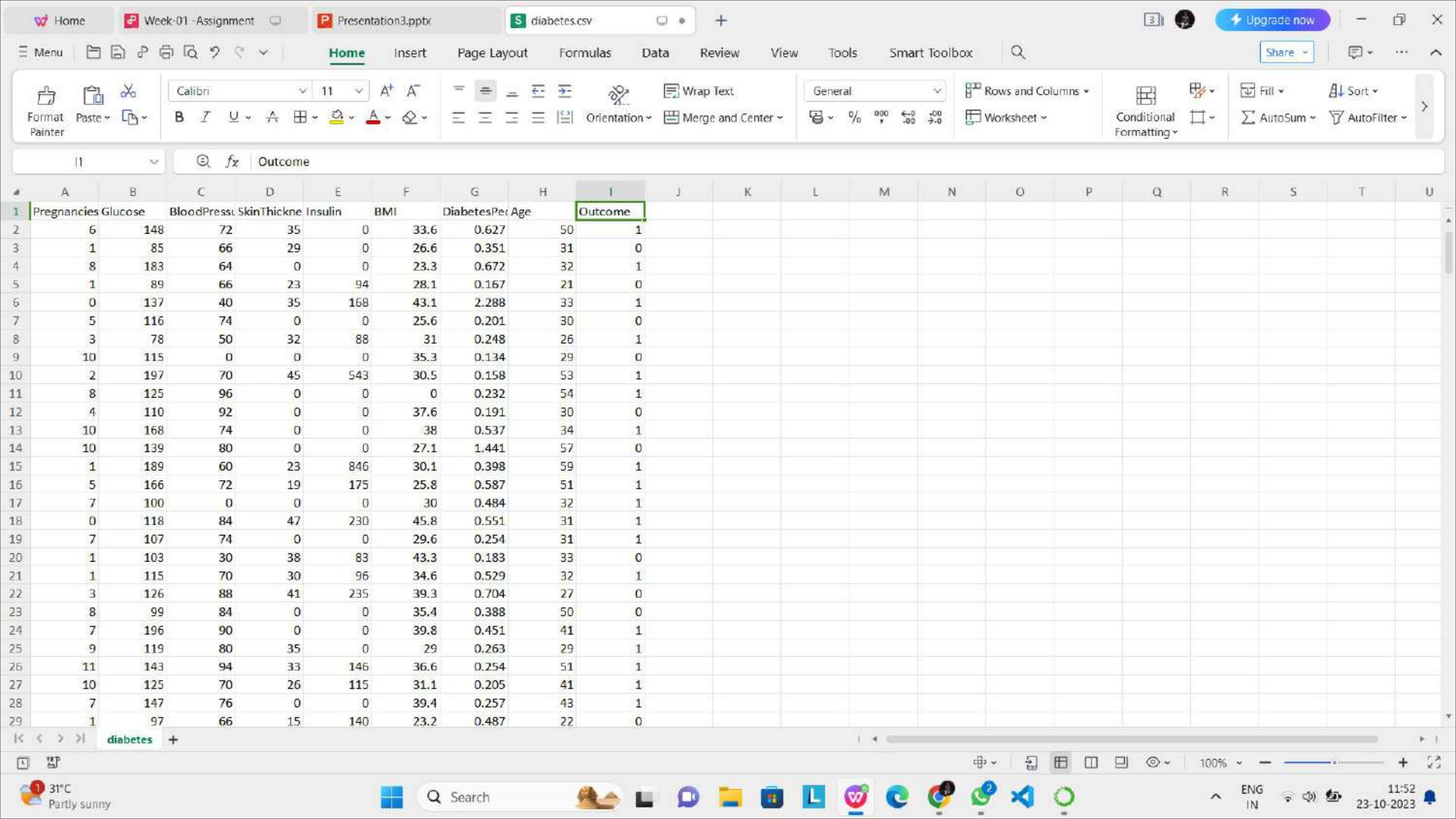


4) Feature Selection and Engineering

5) Model Selection

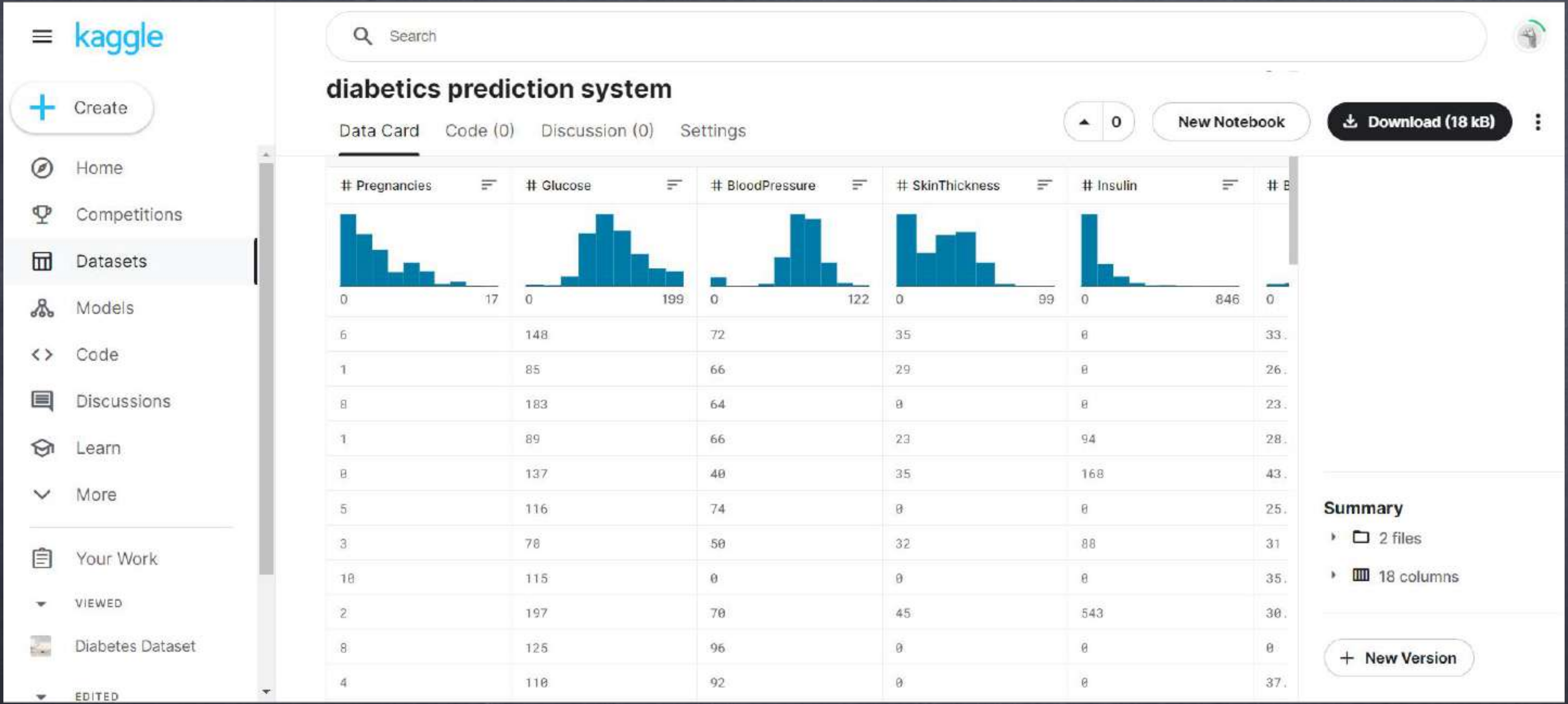
6) Model Training

The model contains the use of Boruta feature selection, the extraction of salient features from datasets, the use of the K-Means++ algorithm for unsupervised clustering of data, and stacking of an ensemble learning method for classification. The model was evaluated by accuracy, precision, and F1 index on the PIMA Indian diabetes dataset and achieved an accuracy rate of 98% The study applies two-fold feature selection techniques (i.e., principal component analysis, PCA, and information gain, IG) to boost the prediction accuracy. Then, the optimal features are fed into five ML algorithms, namely decision tree, random forest, support vector machine, logistic regression, and KNN 2.



I1 *fx* Outcome

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressu	SkinThickne	Insulin	BMI	DiabetesPe	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	9	119	80	35	0	29	0.263	29	1
26	11	143	94	33	146	36.6	0.254	51	1
27	10	125	70	26	115	31.1	0.205	41	1
28	7	147	76	0	0	39.4	0.257	43	1
29	1	97	66	15	140	23.2	0.487	22	0



diabetics prediction system

# Pregnancies	# Glucose	# BloodPressure	# SkinThickness	# Insulin	# B
6	148	72	35	0	33.
1	85	66	29	0	26.
8	183	64	0	0	23.
1	89	66	23	94	28.
0	137	40	35	168	43.
5	116	74	0	0	25.
3	78	50	32	88	31
10	115	0	0	0	35.
2	197	70	45	543	30.
8	125	96	0	0	0
4	110	92	0	0	37.

Summary

- 2 files
- 18 columns

Here are the steps to follow to conclude your project:

1. Download the dataset from Kaggle [3].
2. Open Jupyter Notebook and create a new Python 3 notebook.
3. Import the necessary libraries such as pandas, numpy, matplotlib, seaborn, and sklearn.
4. Load the dataset into the notebook.
5. Preprocess the data by handling missing values, scaling, and encoding categorical variables.
6. Split the data into training and testing sets.
7. Train a machine learning model on the training set.
8. Evaluate the model's performance on the testing set.
9. Use the trained model to predict diabetes for new patients.
10. Write a conclusion summarizing your findings.

SUMMARIZED PART 2 FOR DEVELOPMENT IS GIVEN IN THE NEXT PHASE

THANK YOU !

REPORTED BY SYED UZAIR AHMED
212921104052

PROJ_227128_TEAM_1