DIABETES PREDICTION ANALYSIS

Project 4 Team Members:

1. Eva Brown
2. Jeff He
3. Magdalene Singh

CONTENT

- Introduction
- Proposed Approach
- Background
- Overview of Machine Learning
- Exploring the Data
- Modelling the Data
- Comparison
- Page Launch

INTRODUCTION

Welcome to our Diabetes Prediction Analysis presentation.

Applications used:

- ✓ Python Pandas
- ✓ Python Matplotlib
- ✓ Numpy
- ✓ Scikit-learn
- ✓ Streamlit

PROPOSED APPROACH

- Our project aim was to predict the probability of developing Diabetes by applying Machine Learning.
- Logistic Regression & Random Forest were used to evaluate the accuracy of the prediction outcome.
- The dataset of 100,000 entries was obtained from kaggle.com:

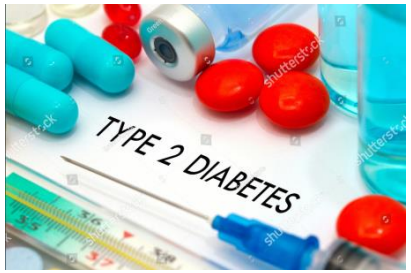
https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

BACKGROUND

DIABETES

- Chronic medical condition
- Elevated levels of blood glucose (or blood sugar).
- The body either does not produce enough insulin or cannot use the insulin it produces.
- Three main types of diabetes:

1. **Type 1 Diabetes:**
    - Immune system destroys insulin-producing beta cells in the pancreas.
    - Insulin injections or an insulin pump required to manage blood sugar levels.



2. Type 2 Diabetes:

- Inability to use insulin properly (insulin resistance) or insufficient production of insulin.
- Lifestyle factors: obesity, lack of physical activity & genetic predisposition.
- Managed through lifestyle modifications, oral medications & insulin therapy.

3. Gestational Diabetes:

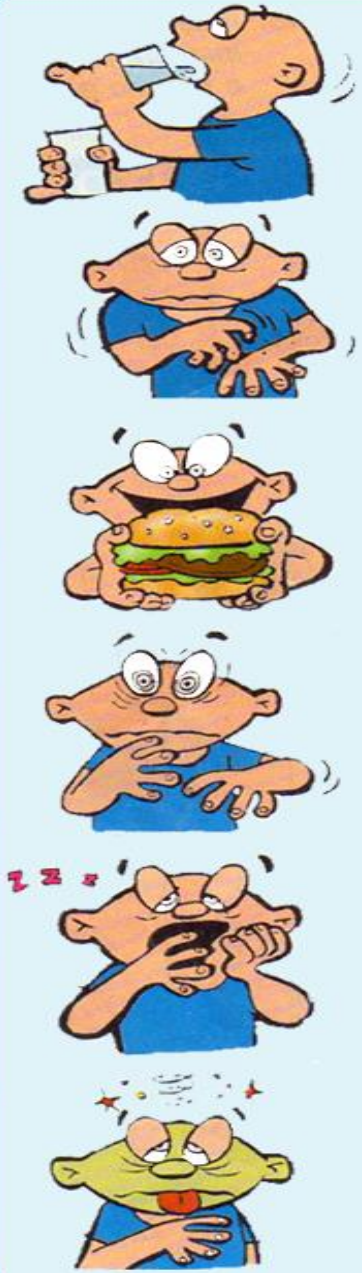
- Occurs during pregnancy.
- Resolves after childbirth.

Common symptoms of diabetes:

increased thirst	unexplained weight loss	fatigue
frequent urination	blurred vision	

If left untreated, diabetes can lead to serious health consequences such as :

heart disease	kidney damage	vision problems
stroke	nerve damage	



Management of diabetes:

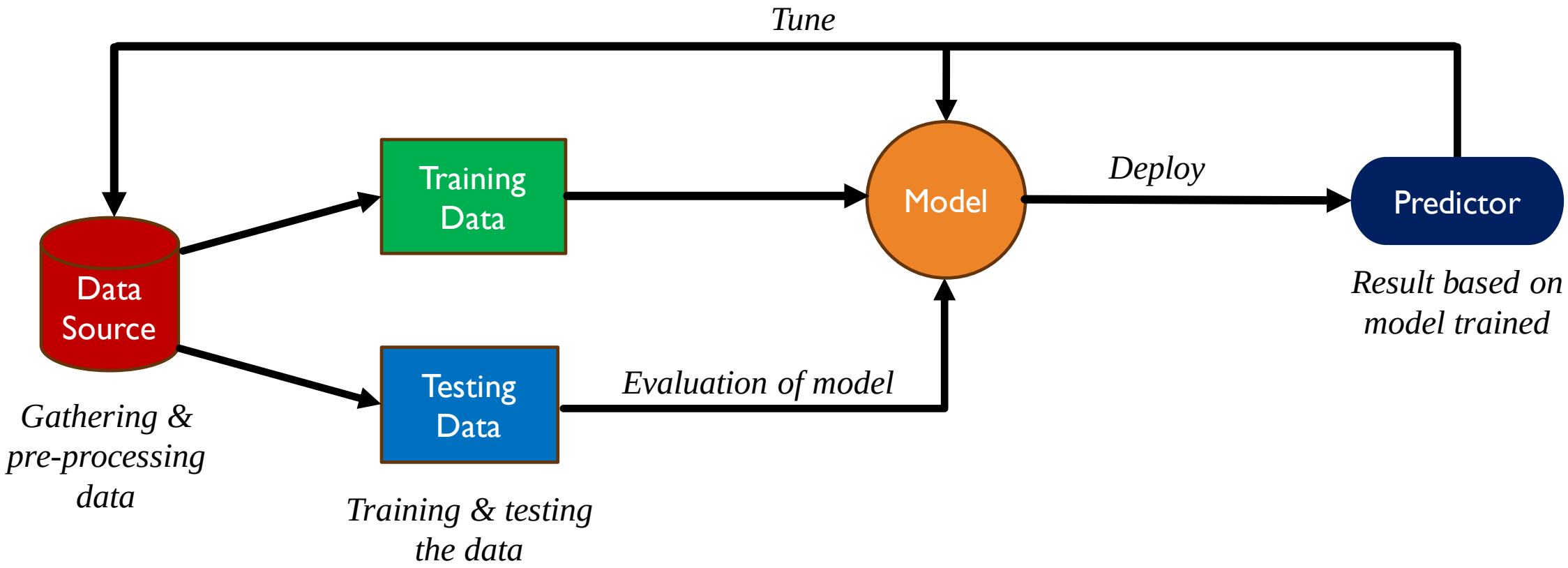
maintaining blood glucose levels within a target range through a combination of medication	lifestyle changes such as healthy eating, regular physical activity, stress management
regular glucose monitoring	regular medical check-ups

Regular medical check-ups are essential for early detection of health consequences and effective diabetes management.

As of January 2022, it is estimated that approximately 10% of the world's adult population has diabetes.

The prevalence of diabetes has been increasing globally and varies by region.

OVERVIEW OF MACHINE LEARNING



EXPLORING THE DATA

1. Read the *diabetes_prediction_dataset.csv* data into a Pandas DataFrame.

2. Fields:

gender	hypertension	heart_disease	HbA1c_level	diabetes
age	smoking_history	bmi	blood_glucose_level	

3. Transformed the data.
   - renamed & regrouped the smoking_history & gender fields

```
# Check the `gender` column's values
diab_pred_df.gender.value_counts()
```

```
Female    56161
Male      39967
Other        18
Name: gender, dtype: int64
```

```
# Check the `smoking_history` column's values
diab_pred_df.smoking_history.value_counts()
```

```
never          34398
No Info        32887
former          9299
current         9197
not current     6367
ever            3998
Name: smoking_history, dtype: int64
```

EXPLORING THE DATA

4. Checked for missing data

```
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   gender               100000 non-null  object
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64
 3   heart_disease        100000 non-null  int64
 4   smoking_history      100000 non-null  object
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64
 8   diabetes             100000 non-null  int64
dtypes: float64(3), int64(4), object(2)
```

EXPLORING THE DATA

5. Dropped duplicated rows

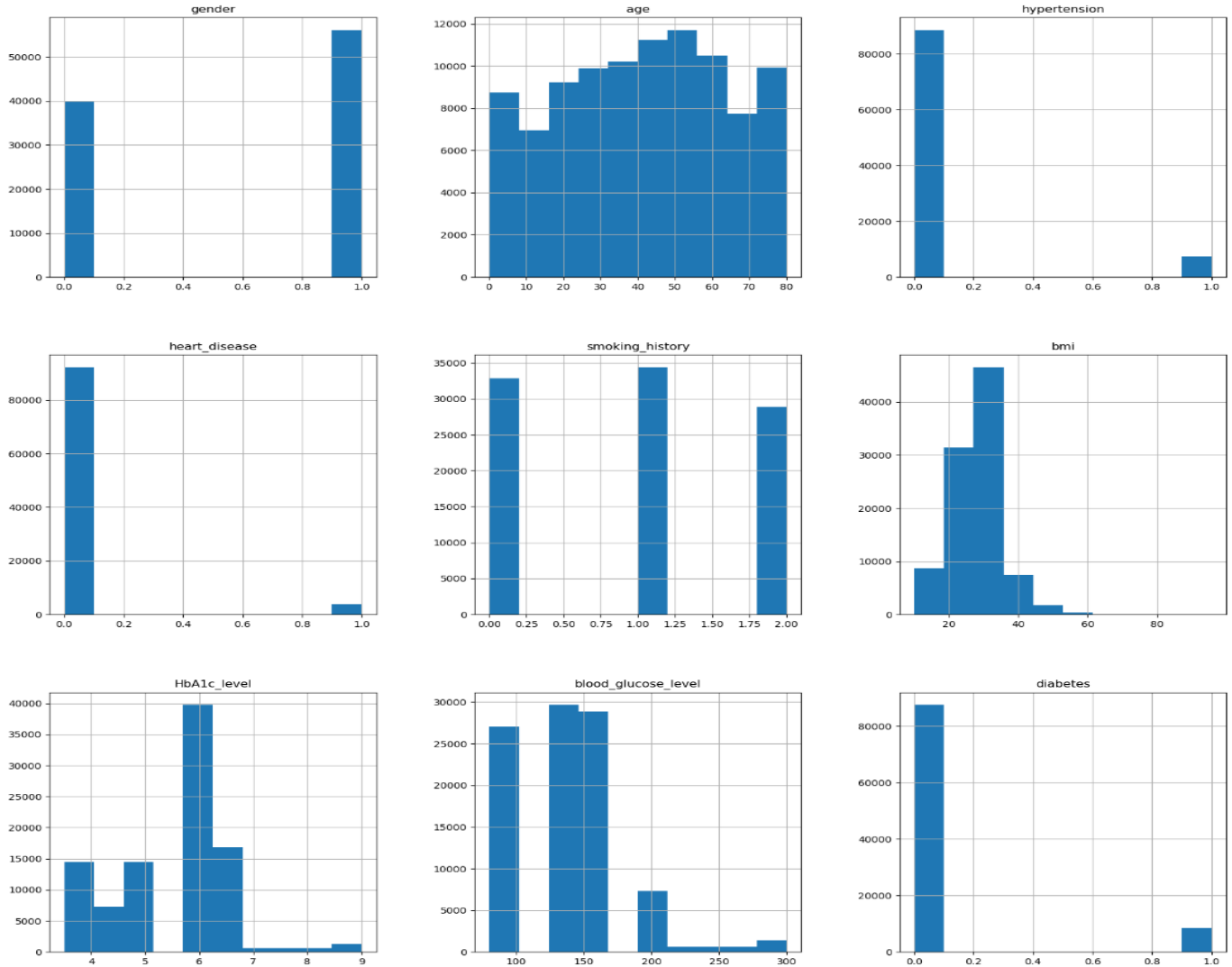
```
diab_pred_df.duplicated().sum()
```

```
3854
```

6. Find the correlation between each feature and diabetes outcome.

```
gender                -0.04
smoking_history        0.12
heart_disease          0.17
hypertension           0.20
bmi                    0.21
age                    0.26
HbA1c_level            0.41
blood_glucose_level    0.42
diabetes               1.00
Name: diabetes, dtype: float64
```

```
# Visualising the data distribution plots
p = df1.hist(figsize = (20,20))
```



MODELLING THE DATA

1. The dataset was split into training and testing datasets.

2. Evaluated the data.

    - Used Logistic Regression as the initial model.

    - According to studies, Random Forest was shown to be a more accurate model for health predictions.

3. We used the Random Forest model for our Diabetes Predictor web page.

COMPARE CLASSIFICATION REPORTS

LOGISTIC REGRESSION

	precision	recall	f1-score	support
0	0.96	0.99	0.98	21912
1	0.83	0.63	0.71	2120
accuracy			0.96	24032
macro avg	0.90	0.81	0.85	24032
weighted avg	0.95	0.96	0.95	24032

RANDOM FOREST

	precision	recall	f1-score	support
0	0.97	1.00	0.98	21912
1	0.94	0.69	0.79	2120
accuracy			0.97	24032
macro avg	0.96	0.84	0.89	24032
weighted avg	0.97	0.97	0.97	24032

- The Random Forest model has a higher precision (94% vs. 83%), indicating fewer false positives.
- The Random Forest model has a higher recall (69% vs. 63%), meaning it captures more true positives among all actual positives.
- The f1-score for class 1 is notably higher for the Random Forest model (79% vs. 71%).

HOLD ON TO YOUR SEATS WHILE WE …

LAUNCH OUR PAGE

The page can be launched directly from GitHub:

- https://github.com/EvaB5050/Diabetes-Prediction-Analysis.git

Thank you for watching our presentation


Questions?

And that, my dear fellow students, marks the end of our Data Analytics BootCamp.

Let's embrace life once again ☺