

**WINTER PEP TRAINING**

**on**

**BIG DATA**

**Submitted by**

**Karanvir Singh**

**Registration No : 12208675**

**Programme Name : Btech. CSE (3rd Year)**

**Under the Guidance of**

***Mr. LOKESH***

**School of Computer Science & Engineering Lovely Professional University, Phagwara**

(January-February,2024)

1 **|** Page

DECLARATION

I hereby declare that I have successfully completed my Winter PEP Training in Big Data from January 14, 2025, to February 12, 2025, under the esteemed guidance of Mr. Lokesh. During this period, I have diligently dedicated myself to the training, actively participating in all learning activities and projects.

Through this training, I have gained in-depth knowledge and practical experience in Big Data technologies, equipping me with the necessary skills to meet the academic and professional requirements for the award of the B.Tech. degree in Computer Science and Engineering at Lovely Professional University, Phagwara.

This declaration is made in acknowledgment of my commitment to excellence and continuous learning in the field of Big Data and Analytics.

Name of Student – Karanvir Singh Registration no: 12208675

## ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to Lovely Professional University for providing me with this golden opportunity to participate in the Winter PEP Training on Big Data. This training has been an enriching experience, allowing me to enhance my skills, complete practical assignments, and expand my knowledge in the field. Through this journey, I have discovered numerous new concepts that have significantly contributed to my learning.

I am also deeply grateful to my friends and peers, whose unwavering support has been invaluable. Their willingness to help whenever I encountered challenges in my coursework has been instrumental in my progress. I truly appreciate their encouragement and the strong sense of collaboration we share.

Furthermore, I extend my sincere appreciation to our esteemed trainer, **Mr. Lokesh** (Senior Trainer), for his guidance, mentorship, and continuous support throughout the training. His constructive feedback, detailed explanations, and dedication to our learning have played a crucial role in refining our understanding of Big Data. His patience in reviewing our assignments, providing insightful corrections, and ensuring we grasp the core concepts is truly commendable.

Lastly, I would like to acknowledge the contributions of everyone who has supported me throughout this training—be it individuals or organizations. Their encouragement and assistance have made this learning journey a memorable and transformative experience.

## **INTRODUCTION**

### **Big Data Analytics: Transforming Industries with Data-Driven Insights**

Big Data Analytics refers to the **process of examining vast volumes of structured, semi-structured, and unstructured data** to uncover hidden patterns, correlations, and valuable insights. With the rapid advancements in digital transformation, organizations across various sectors are increasingly leveraging Big Data to **enhance decision-making, forecast trends, and improve overall operational efficiency**.

### **Key Concepts in Big Data**

The effectiveness of Big Data Analytics is built on five fundamental principles, commonly known as the **5Vs of Big Data**:

* **Volume** – The sheer amount of data generated from multiple sources such as social media, IoT devices, business transactions, sensors, and digital applications. Organizations must employ scalable storage and processing solutions to manage these large datasets efficiently.
* **Velocity** – The speed at which data is generated, processed, and analyzed. In today’s fast-paced digital world, real-time data processing is crucial for applications such as financial trading, online recommendations, and fraud detection.
* **Variety** – The diverse nature of data, which can be structured (databases), semi-structured (XML, JSON), or unstructured (text, images, videos, and social media posts). Handling this variety requires advanced data integration and analytics tools.
* **Veracity** – The reliability, accuracy, and quality of data. Since inaccurate or inconsistent data can lead to flawed insights, organizations must implement data validation, cleansing, and governance strategies to ensure data integrity.
* **Value** – The ultimate goal of Big Data Analytics is to derive **actionable insights** that contribute to **business growth, cost reduction, efficiency improvements, and customer satisfaction**. Extracting value from data requires powerful analytics techniques such as machine learning, artificial intelligence, and statistical modeling.

### **Impact of Big Data Across Industries**

Big Data Analytics is transforming industries by providing deeper insights, **enhancing decision-making processes**, and **optimizing business operations**. Some of its major applications include:

* **Healthcare** – Analyzing patient records, medical imaging, and real-time health data to **improve diagnostics, personalize treatment plans, and enhance disease prevention**. Hospitals use predictive analytics to optimize resource allocation and patient care.
* **Finance** – Financial institutions leverage Big Data for **risk assessment, fraud detection, and algorithmic trading**. AI-driven models analyze historical transaction data to predict market trends and detect suspicious activities in real time.
* **E-Commerce & Retail** – Companies analyze customer purchasing behavior, online interactions, and feedback to **personalize recommendations, optimize pricing strategies, and enhance customer satisfaction**. Demand forecasting helps businesses manage inventory more effectively.
* **Marketing & Advertising** – Businesses use consumer insights to create **targeted ad campaigns, optimize digital marketing strategies, and improve customer engagement**. Social media sentiment analysis helps brands understand customer opinions and tailor their messaging accordingly.
* **Manufacturing & Supply Chain** – Predictive analytics is used to **monitor equipment performance, reduce downtime, and streamline logistics**. Data-driven insights improve efficiency in production planning and inventory management.
* **Smart Cities & IoT** – Governments and urban planners use Big Data to **enhance traffic management, energy efficiency, and public safety**. IoT sensors generate real-time data that can be analyzed for smart infrastructure development.

### **Future of Big Data Analytics**

With continuous advancements in **cloud computing, artificial intelligence, and machine learning**, the future of Big Data Analytics looks promising. Organizations are increasingly adopting **real-time analytics, automated decision-making systems, and AI-driven predictive models** to stay competitive in a data-driven world. Additionally, ethical data usage, privacy concerns, and regulatory compliance will play a significant role in shaping the future of Big Data applications.

As industries continue to embrace **data-driven innovation**, mastering Big Data Analytics will be crucial for businesses and professionals seeking to **gain a competitive edge, drive efficiency, and unlock new opportunities** in the digital era.

## **TECHNOLOGIES USED**

During my training, I worked with various tools and technologies in Big Data Analytics, including:

**Hadoop** – Distributed storage and processing framework.

**Apache Spark** – Real-time data processing engine.

**Hive** – Data warehousing solution for Hadoop.

**Kafka** – Data streaming platform.

**Python & Pandas** – Data analysis and visualization.

**Tableau** – Interactive data visualization tool.

**SQL & NoSQL Databases** – Managing structured and unstructured data efficiently.

**Machine Learning with Big Data** – Applying ML techniques for predictive analytics.

## **TRAINING MODULES & LEARNING OUTCOMES**

### **Week 1: Introduction to Big Data & Hadoop**

* Basics of Big Data, its challenges, and industry applications.
* Hadoop architecture, HDFS (Hadoop Distributed File System), and MapReduce.
* Installation and setup of the Hadoop environment.
* Writing and running MapReduce jobs.

### **Week 2: Data Processing with Apache Spark**

* Introduction to Spark and RDDs (Resilient Distributed Datasets).
* Spark SQL and DataFrames for handling structured data.
* Real-time streaming with Spark Streaming.
* Data transformations using Spark MLlib.

### **Week 3: Data Warehousing & Visualization**

* Using **Hive** for querying large datasets.
* Connecting **Python Pandas** with Big Data frameworks.
* Data visualization using **Tableau**.
* Building interactive dashboards for data insights.

### **Week 4: Hands-on Project & Implementation**

* **Project:** [Diabatic test]
* Collecting, processing, and analyzing large datasets.
* Applying machine learning models for predictive analytics.
* Implementing a recommendation system using collaborative filtering.
* Final assessment and report submission.

# HTML CODE

<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8">

<meta name="viewport" content="width=device-width, initial-scale=1.0">

<title>Diabetes Prediction</title>

<link rel="stylesheet" href="static/styles.css">

<script src="https://ajax.googleapis.com/ajax/libs/jquery/3.6.0/jquery.min.js"></script>

</head>

<body>

<div class="hero-section"></div>

<div class="container">

<div class="form-box">

<h1>Diabetes Predictor</h1>

<form id="predictionForm">

<label for="pregnancies">Pregnancies:</label>

<input type="number" id="pregnancies" name="pregnancies" required>

<label for="glucose">Glucose Level:</label>

<input type="number" id="glucose" name="glucose" required>

<label for="blood\_pressure">Blood Pressure:</label>

<input type="number" id="blood\_pressure" name="blood\_pressure" required>

<label for="skin\_thickness">Skin Thickness:</label>

<input type="number" id="skin\_thickness" name="skin\_thickness" required>

<label for="insulin">Insulin Level:</label>

<input type="number" id="insulin" name="insulin" required>

<label for="bmi">BMI:</label>

<input type="number" step="0.1" id="bmi" name="bmi" required>

<label for="diabetes\_pedigree">Diabetes Pedigree Function:</label>

<input type="number" step="0.01" id="diabetes\_pedigree" name="diabetes\_pedigree" required>

<label for="age">Age:</label>

<input type="number" id="age" name="age" required>

<button type="submit">Predict</button>

</form>

<div id="result"></div>

</div>

</div>

<script>

$(document).ready(function() {

$('#predictionForm').submit(function(e) {

e.preventDefault();

$.ajax({

url: '/predict',

method: 'POST',

data: $(this).serialize(),

success: function(response) {

$('#result').html(`<p class="result">${response.result}</p>`);

} else if (response.error) {

$('#result').html(`<p class="error">${response.error}</p>`);

}

HTML

if (response.result) {

$('#result').html(`<p class="result">${response.result}</p>`);

} else if (response.error) {

$('#result').html(`<p class="error">${response.error}</p>`);

}

}

});

});

});

</script>

</body>

</html>

# CSS CODE

body {

margin: 0;

padding: 0;

color: #c62929;

background-image: url(background.jpg);

background-size: cover;

font-family: Arial, sans-serif;

}

.hero-section {

background: url('static/images/glucose-monitor-7325306\_1280.png') no-repeat center;

background-size: cover;

height: 400px;

width: 100%;

}

.container {

display: flex;

justify-content: center;

align-items: center;

height: 100vh;

}

.form-box {

background-color: rgba(0, 0, 0, 0.9);

padding: 25px;

border-radius: 10px;

box-shadow: 0px 4px 12px rgba(0, 0, 0, 0.2);

width: 100%;

max-width: 400px;

text-align: center;

}

h1 {

color: #c62929;

}

form label {

display: block;

margin: 10px 0 5px;

color: #fefefe;

font-weight: bold;

}

form input {

width: 100%;

padding: 10px;

margin-bottom: 15px;

border: 1px solid #ccc;

border-radius: 5px;

font-size: 16px;

}

form button {

width: 100%;

padding: 10px;

background-color: #007bff;

color: white;

border: none;

border-radius: 5px;

font-size: 18px;

cursor: pointer;

transition: 0.3s;

}

form button:hover {

background-color: #0056b3;

}

#result {

margin-top: 20px;

font-size: 18px;

}

.result {

color: green;

font-weight: bold;

}

.error {

color: red;

font-weight: bold;

}

# PYTHON CODE

from flask import Flask, request, render\_template, jsonify

import joblib

import numpy as np

import pandas as pd

# Load the trained diabetes model and scaler

diabetes\_model = joblib.load('diabetes\_model.pkl')

scaler = joblib.load('scaler.pkl')

app = Flask(\_\_name\_\_)

@app.route('/')

def index():

return render\_template('index.html')

@app.route('/predict', methods=['POST'])

def predict():

try:

# Get user input

features = np.array([[

int(request.form['pregnancies']),

float(request.form['glucose']),

float(request.form['blood\_pressure']),

float(request.form['skin\_thickness']),

float(request.form['insulin']),

float(request.form['bmi']),

float(request.form['diabetes\_pedigree']),

int(request.form['age'])

]])

# Scale input data

scaled\_features = scaler.transform(features)

# Make prediction

prediction = diabetes\_model.predict(scaled\_features)

result = "Diabetic" if prediction[0] == 1 else "Not Diabetic"

return jsonify({'result': result})

except Exception as e:

return jsonify({'error': str(e)})

if \_\_name\_\_ == '\_\_main\_\_':

app.run(debug=True)

# CODE TO TRAIN MODEL

import pandas as pd

import numpy as np

import joblib

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score

# Load dataset

url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"

columns = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']

data = pd.read\_csv(url, names=columns)

# Split data into features and target variable

X = data.drop(columns=['Outcome'])

y = data['Outcome']

# Split into training and test sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Standardize features

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

# Train a Random Forest model

model = RandomForestClassifier(n\_estimators=100, random\_state=42)

model.fit(X\_train, y\_train)

# Test model accuracy

y\_pred = model.predict(X\_test)

accuracy = accuracy\_score(y\_test, y\_pred)

print(f"Model Accuracy: {accuracy:.2f}")

# Save the model

joblib.dump(model, 'diabetes\_model.pkl')

joblib.dump(scaler, 'scaler.pkl')

print("Model and scaler saved as 'diabetes\_model.pkl' and 'scaler.pkl'.")

# IMPORTANT THING TO REMEMBER

# Pregnancies: 0 (for males) or 1-2 (for females)

# Glucose Level: 90 (70-99 mg/dL is normal fasting glucose)

# Blood Pressure: 80 (Normal range: 80-120)

# Skin Thickness: 20 (Typical range: 10-30 mm)

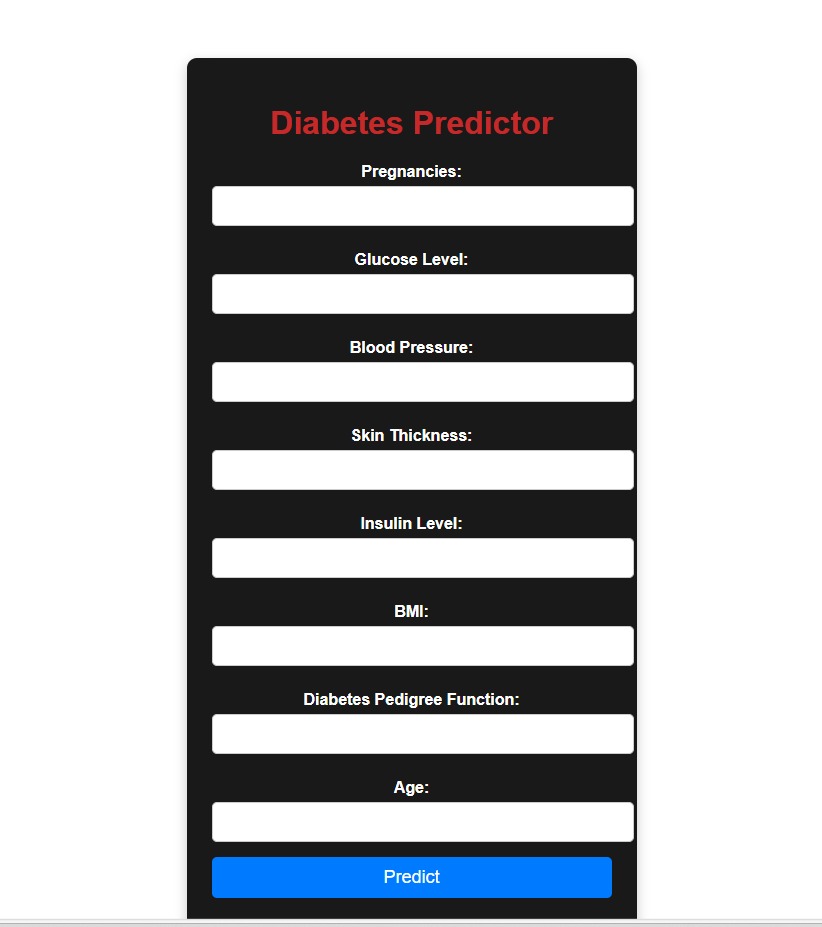
# Insulin Level: 85 (Normal fasting insulin: 16-166 µU/mL)

# BMI: 22.0 (Normal BMI range: 18.5-24.9)

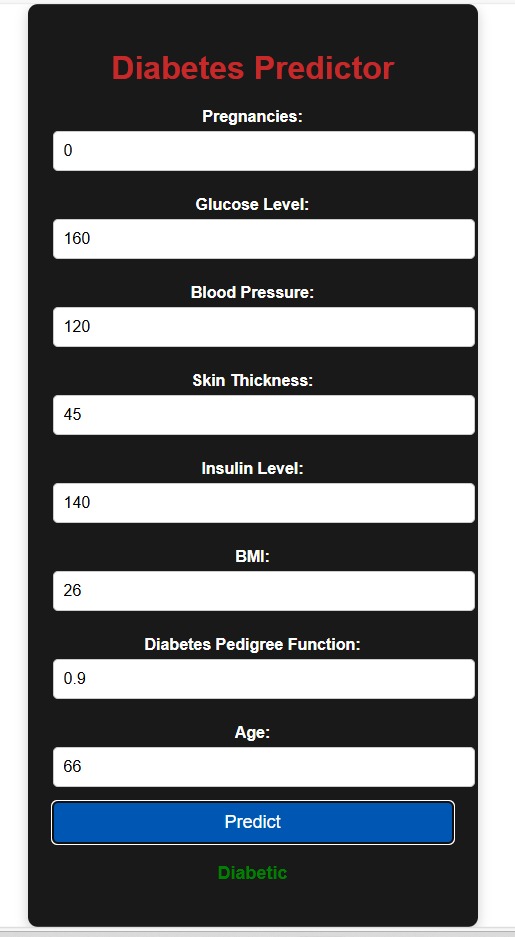
# Diabetes Pedigree Function: 0.3 (Lower values mean lower genetic risk)

# Age: 25 (Younger people have a lower risk)

# OUTPUT OF THE CODE

****

# Person is diabatic



# Person is non diabetic

# non

**CONCUSION**

Practical knowledge means the visualization of the knowledge, which we read in our books. For this, we perform experiments and get observations. Practical knowledge is very important in every field. One must be familiar with the problems related to that field so that he may solve them and become a successful person. After achieving the proper goal in life, an engineer has to enter in professional life. According to this life, he has to serve an industry, may be public or private sector or self own. For the efficient work in the field, he must be well aware of the practical knowledge as well as theoretical knowledge. Due to all above reasons and to bridge the gap between theory and practical, our Engineering curriculum provides a winter pep class. During this period a student work in the industry and get well all type of experience and knowledge about the working of companies and hardware and software tools. I have undergone my winter pep training in 6 th sem. This report is based on the knowledge, which I acquired during my winter pep training.