AI BASED DIABETES PREDICTION SYSTEM

PHASE 3: DEVELOPMENT PART-1

DATASET PROGRAM:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from joblib import dump, load
data = pd.read_csv('diabetes_dataset.csv')
X = data.drop('diabetes_status', axis=1)
y = data['diabetes_status']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
print(f'Accuracy: {accuracy}')
print('Classification Report:\n', report)
dump(model, 'diabetic_prediction_model.joblib')
```

Description:

This dataset is designed to aid in the development of an AI-based diabetes prediction system. It includes both independent variables (features) and the target variable (diabetes status) for a sample of individuals. The dataset contains a mix of demographic, clinical, and lifestyle information that may be relevant to predicting diabetes.

Features (Independent Variables):

1. Age: Age of the individual (in years).

2. Gender: Gender of the individual (categorical: Male, Female).

3. BMI (Body Mass Index): A measure of body fat based on height and weight.

4. Family History: Family history of diabetes (categorical: Yes, No).

5. Blood Pressure: Systolic and diastolic blood pressure (mm Hg).

6. Glucose Level: Fasting blood glucose level (mg/dL).

7. Insulin Level: Fasting insulin level (μU/mL).

8. Cholesterol Level: Total cholesterol level (mg/dL).

9. Physical Activity: Self-reported level of physical activity (categorical: Sedentary, Light, Moderate, Active).

10. Diet: Self-reported dietary pattern (categorical: Healthy, Unhealthy).

11. Smoking Status: Smoking status (categorical: Current Smoker, Former Smoker, Non-Smoker).

12. Diabetes Status: Binary variable indicating diabetes status (categorical: Yes, No).

Dataset Size:

Ideally, the dataset should include data for a diverse and representative sample of individuals, with at least a few thousand records.

Data Collection:

Data should be collected through surveys, medical records, and clinical tests. Ensure that all data collection complies with ethical guidelines and data privacy regulations. Consent should be obtained from participants, and personal identifiers should be removed or anonymized.

Data Preparation:

1. Handle missing values (if any) using appropriate techniques.

2. Normalize or scale numerical features.

3. Encode categorical variables (e.g., one-hot encoding).

4. Split the dataset into training, validation, and test sets.