**Project Report**

**Title:** Drug Side Effects Classification Using Patient Demographics

**Objective:** The objective of this project is to build a machine learning classification model that can predict the side effects of a particular drug based on patient demographic details, specifically age, gender, and race.

**Introduction:** Understanding the potential side effects of a drug is crucial for safe prescription and usage. Different patient groups often respond differently to the same drug, depending on their demographic profile. In this project, a model is built to predict the type of side effect a patient might experience using synthetic patient data.

**Dataset Description:** A synthetic dataset of 400,000 patient records was generated. Each entry includes the following attributes:

* Patient\_ID (anonymized identifier)
* Age (0-100)
* Gender (Male, Female, Other)
* Race (Asian, African American, White, Hispanic, Other)
* Side\_Effect (None, Mild, Severe, Allergic Reaction)

**Data Cleaning and Preprocessing:**

1. **Duplicate Removal:** Any duplicate patient records were removed to ensure unique entries.
2. **Sanitization:** The "Name" column was replaced with anonymized "Patient\_ID" to comply with data protection principles.
3. **Missing Values:** Any entries with missing demographic information were dropped.
4. **Encoding:** Label encoding was applied to convert categorical fields (Gender, Race, Side\_Effect) into numerical values.
5. **Feature Scaling:** The "Age" field was scaled using standardization to improve model convergence and performance.
6. **Data Partitioning:** The dataset was split into training (80%) and testing (20%) sets using scikit-learn's train\_test\_split function.

**Model Building:** A Decision Tree Classifier was chosen as the model for this project because of its interpretability and effectiveness in handling both numerical and categorical variables. The model was trained on the training set and then evaluated on the test set.

**Evaluation Metrics:**

* **Accuracy:** Measures the proportion of correct predictions over total predictions.
* **Confusion Matrix:** Gives a summary of prediction results showing the true vs. predicted labels.
* **Classification Report:** Includes Precision, Recall, and F1-Score for each class.

**Conclusion:** The classification model effectively predicts the potential side effects of a drug based on patient demographic data. Such a system can be used by healthcare professionals to improve drug safety and customize prescriptions for specific patient groups.

**Future Scope:**

* Integration of real-world medical records.
* Expansion to include more patient attributes like medical history, lifestyle, and dosage information.
* Exploring more advanced models like Random Forest, XGBoost, and Neural Networks for improved performance.

**Submitted By:** akshaj masta

--- End of Report ---