Experiment - 06

Aim: To implement SMOTE techniques to generate synthetic data to solve the problem of class imbalance

Theory: SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances.

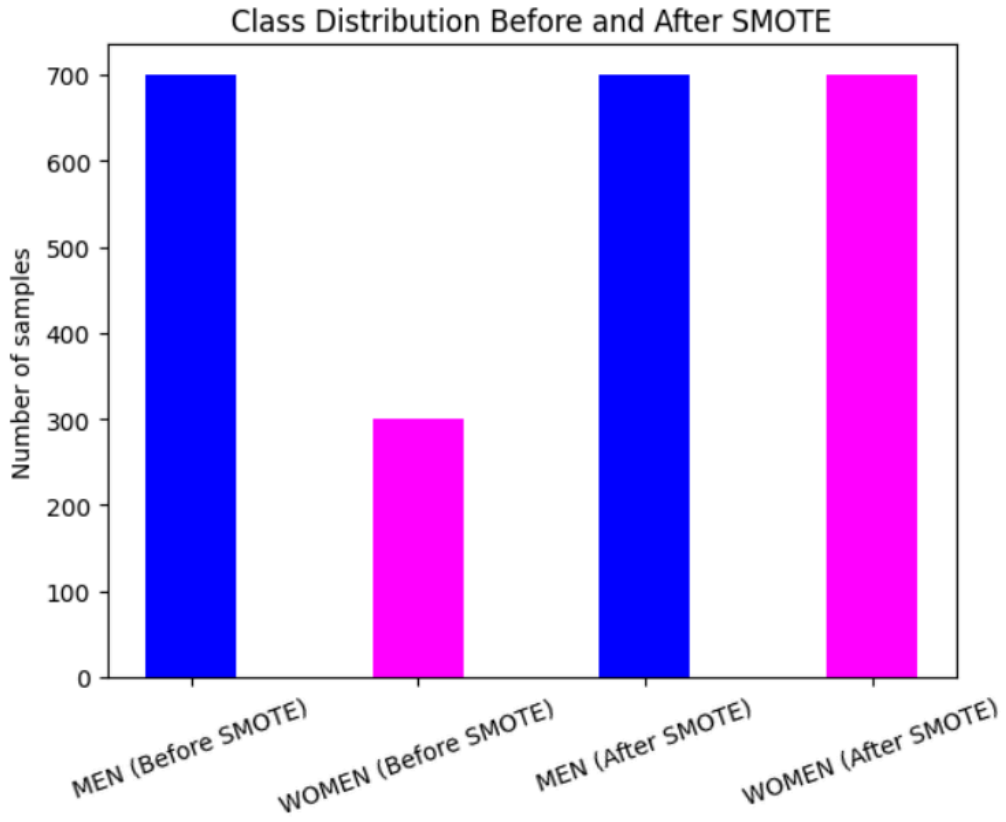
These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

Code:

```
import numpy as np
import matplotlib.pyplot as plt
from collections import Counter
from sklearn.datasets import make_classification
from imblearn.over_sampling import SMOTE

# Step 1: Create an imbalanced dataset
X, y = make_classification(n_classes=2, class_sep=2, weights=[0.7,
0.3],n_informative=3, n_redundant=1, flip_y=0,n_features=5,
n_clusters_per_class=1, n_samples=1000, random_state=42)

# Apply SMOTE
smote = SMOTE(sampling_strategy='auto', random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)

# Class distributions
before_counts = Counter(y)
after_counts = Counter(y_resampled)

# Set positions for bars
x_labels = ["MEN (Before SMOTE)", "WOMEN (Before SMOTE)", "MEN (After
SMOTE)", "WOMEN (After SMOTE)"]
x_positions = np.arange(len(x_labels))  # Generate equally spaced
positions

# Plot the bars
plt.bar(x_positions[:2], before_counts.values(), color=['blue',
'fuchsia'], width=0.4, label="Before SMOTE")
```

```
plt.bar(x_positions[2:], after_counts.values(), color=['blue',
'fuchsia'], width=0.4, label="After SMOTE")

# Adjust x-axis labels
plt.xticks(x_positions, x_labels, rotation=20)  # Rotate labels for
better readability
plt.ylabel("Number of samples")
plt.title("Class Distribution Before and After SMOTE")
plt.show()
```

Output:



Conclusion: Hence, we successfully implemented SMOTE to solve the problem of imbalanced datasets