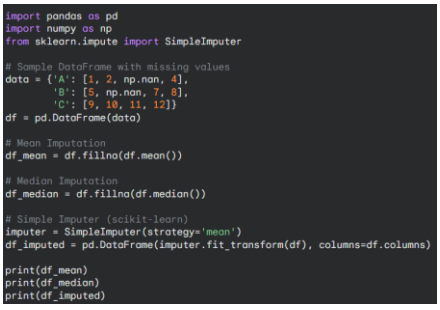
# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 24 SEPTEMBER 2024 |
| Team ID | SWTID1727151090 |
| Project Title | Classification of Arrhythmia by Using Deep Learning with 2-D ECG Spectral Image Representation |
| Maximum Marks | 2 Marks |

**Data Quality Report Template**

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

| Data Source | Data Quality Issue | Severity | Resolution Plan |
|---|---|---|---|
| Dataset | Missing values | Moderate | Imputation<br><br>Resolution using Pandas:<br><br>  |

```python
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer

# Sample DataFrame with missing values
data = {'A': [1, 2, np.nan, 4],
        'B': [5, np.nan, 7, 8],
        'C': [9, 10, 11, 12]}
df = pd.DataFrame(data)

# Mean Imputation
df_mean = df.fillna(df.mean())

# Median Imputation
df_median = df.fillna(df.median())

# Simple Imputer (scikit-learn)
imputer = SimpleImputer(strategy='mean')
df_imputed = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)

print(df_mean)
print(df_median)
print(df_imputed)
```

| | | | |
|---|---|---|---|
| Dataset | Duplicate values | Low | Using Pandas:<br>`import pandas as pd`<br>`# Sample DataFrame with duplicat`<br>`data = {'A': [1, 2, 3, 1, 2],`<br>`'B': [4, 5, 6, 4, 5]}`<br>`df = pd.DataFrame(data)`<br>`# Find duplicate rows`<br>`duplicates = df.duplicated()`<br>`print(duplicates)`<br>`# Remove duplicate rows`<br>`df_unique = df.drop_duplicates()`<br>`print(df_unique)` |
| Dataset | Data inconsistency and accuracy | High | `train_datagen = ImageDataGenerator(`<br>`rescale=1./255,`<br>`shear_range=0.2,`<br>`zoom_range=0.2,`<br>`horizontal_flip=True`<br>`)`<br>`test_datagen = ImageDataGenerator(rescale=1./255,`<br>`shear_range=0.2,`<br>`zoom_range=0.2,`<br>`horizontal_flip=True)` |