

```
[5]: import pandas as pd
import numpy as np
from sklearn.model_selection import KFold
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, roc_auc_score
from sklearn.inspection import permutation_importance

# Load the data
df = pd.read_csv('SE2.csv')
X = df.drop(columns=['CID', 'se_C0000737'])
y = df['se_C0000737']

# Initialize variables to store results
accuracies = []
auc_scores = []
feature_importances = np.zeros(X.shape[1])

# Perform 10-fold cross-validation
kf = KFold(n_splits=5, shuffle=True, random_state=42)
for fold, (train_index, val_index) in enumerate(kf.split(X), 1):
    X_train, X_val = X.iloc[train_index], X.iloc[val_index]
    y_train, y_val = y.iloc[train_index], y.iloc[val_index]

    # Train the model
    rf = RandomForestClassifier(n_estimators=100, random_state=42)
    rf.fit(X_train, y_train)

    # Make predictions
    y_pred = rf.predict(X_val)
    y_pred_prob = rf.predict_proba(X_val)[:, 1]

    # Calculate metrics
    accuracy = accuracy_score(y_val, y_pred)
    auc = roc_auc_score(y_val, y_pred_prob)

    accuracies.append(accuracy)
    auc_scores.append(auc)

    # Calculate feature importances
    perm_importance = permutation_importance(rf, X_val, y_val, n_repeats=5, random_state=42)
    feature_importances += perm_importance.importances_mean

    print(f"Fold {fold} - Accuracy: {accuracy:.4f}, AUC-ROC: {auc:.4f}")

# Calculate average metrics
avg_accuracy = np.mean(accuracies)
avg_auc = np.mean(auc_scores)
print(f"\nAverage Accuracy: {avg_accuracy:.4f}")
print(f"Average AUC-ROC: {avg_auc:.4f}")

# Get top 10 features
feature_importances /= 10 # Average over 10 folds
feature_importance_df = pd.DataFrame({'Feature': X.columns, 'Importance': feature_importances})
feature_importance_df = feature_importance_df.sort_values('Importance', ascending=False)
top_10_features = feature_importance_df.head(10)
print("\nTop 10 Features:")
print(top_10_features)
```

```
Fold 1 - Accuracy: 0.6272, AUC-ROC: 0.6826
Fold 2 - Accuracy: 0.6631, AUC-ROC: 0.6850
Fold 3 - Accuracy: 0.6308, AUC-ROC: 0.6542
Fold 4 - Accuracy: 0.6272, AUC-ROC: 0.7634
Fold 5 - Accuracy: 0.6223, AUC-ROC: 0.6460
```

```
Average Accuracy: 0.6341
Average AUC-ROC: 0.6862
```

```
Top 10 Features:
```

	Feature	Importance
1282	cpi_9606.ENSEP00000255030	0.002795
2876	cpi_9606.ENSEP00000290866	0.002509
9042	cpi_9606.ENSEP00000464149	0.002151
6609	cpi_9606.ENSEP00000360973	0.001792
1732	cpi_9606.ENSEP00000262623	0.001792
4610	cpi_9606.ENSEP00000329684	0.001578
7211	cpi_9606.ENSEP00000367959	0.001506
5889	cpi_9606.ENSEP00000353915	0.001433
5881	cpi_9606.ENSEP00000353820	0.001364
2371	cpi_9606.ENSEP00000272190	0.001362