Cumulative Progress Report #2

Thyroid Disease Detection

Group-F

Loyalist College in Victoria Park, Toronto Campus

2024S-T1 AISC1006 - Step Presentation (Step 1) 01 (M07 Group 1)

Prof. Usman Ahmad

June 15, 2024

Group Members

Student Name	Student Id
Moksh Jaiswal	500240046
Alen Charuvila Saji	500237019
Adarsh Shriram Pednekar	500233484
Utsav Harshadbhai Khamar	500238367
Pranay Sai Jangeti	500240045
Taranjot Singh Bindra	500239542
Smit Rajendraprasad Patel	500238279
Om Kiranbhai Patel	500228172
Tanzima Mohammadyasin Shaikh	500238212
Aravind Seenivasan	500236355

Abstract

The Thyroid Disease Detection project aims to enhance the accuracy and efficiency of thyroid disorder diagnosis using machine learning techniques. Leveraging advanced algorithms and data analytics, the project focuses on developing a robust and reliable system capable of analyzing diverse medical data, including thyroid function tests and imaging results. This system is intended to assist healthcare professionals in the early and accurate identification of thyroid disorders, leading to timely interventions and improved patient outcomes.

Project Goals Recap

- Enhancing diagnostic accuracy.
- Improving diagnostic efficiency.
- Supporting healthcare professionals in early diagnosis.
- Facilitating timely interventions and better patient care.

Tasks Performed So Far

1. **Data Collection and Preprocessing:**
   - **Collected dataset** from the UCI Machine Learning Repository and Kaggle.
   - **Cleaned and preprocessed** the dataset to handle missing values, normalize data, and encode categorical variables.
2. **Exploratory Data Analysis:**
   - **Analyzed the dataset** to understand distributions and relationships.
   - **Visualized data** to identify patterns and relationships.

Week 11 Tasks

In the last week we have performed feature scaling and In this week, we have created data splitting and performed it onto all the base models to compare and get the best model. We also define the model's parameters for hyperparameter tuning using GridSearchCv with cross - validation.

Implementation in Our Project

All team members agreed to address data leakage in our Thyroid Disease Detection project. We decided to follow best practices to ensure the integrity of our model evaluation.

Summary of Last Week

- **Feature Scaling:** Applied to all relevant features.

This Week's Activities

- **Data Splitting:** Performed to ensure robust model evaluation.
- **Model Evaluation:** Developed and compared multiple base models.
- **Hyperparameter Tuning:** Defined model parameters and optimized using `GridSearchCV` with cross-validation.

Implementation

Data Splitting

- **Method:** `train_test_split()` from `sklearn`
  - **Training Set:** 80%
  - **Test Set:** 20%

Model Building and Evaluation

Developed and evaluated the following machine learning models:

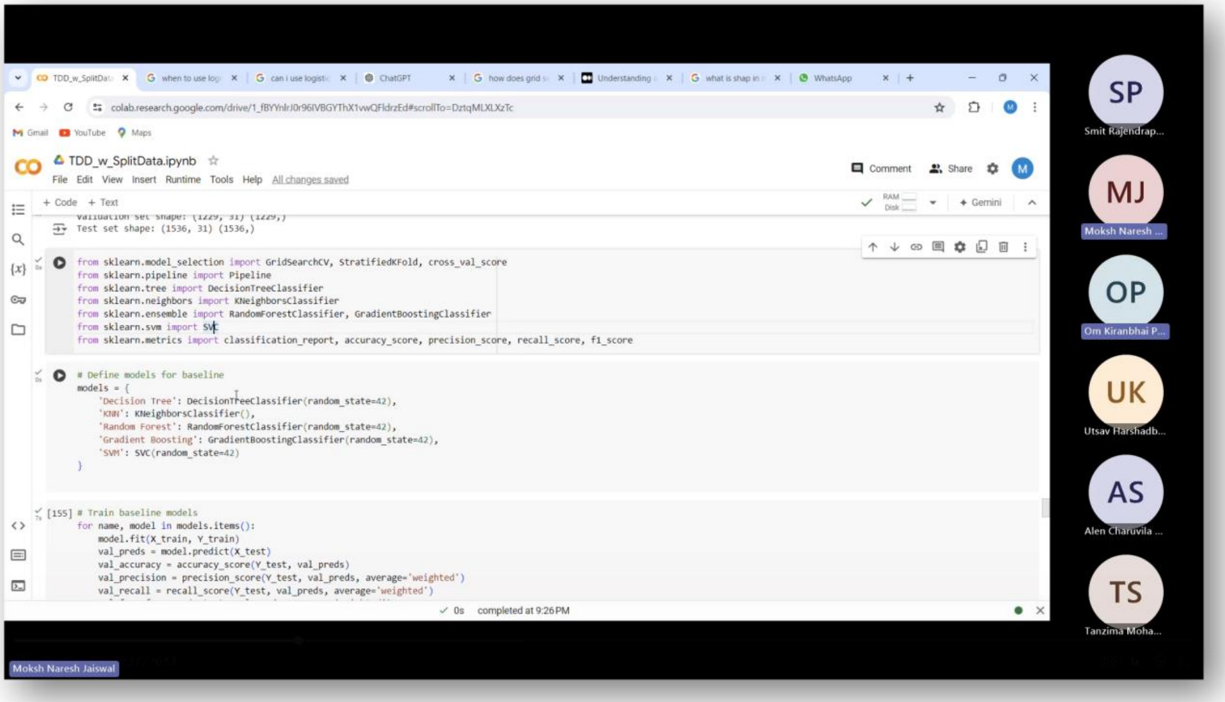
1. **Decision Tree:** Accuracy = 0.9785
2. **K-Nearest Neighbors (KNN):** Accuracy = 0.9147
3. **Random Forest:** Accuracy = 0.9870
4. **Gradient Boosting:** Accuracy = 0.9889
5. **Support Vector Machine (SVM):** Accuracy = 0.8958

Best Models

- **Gradient Boosting:** Accuracy = 0.9889, Cross-Validation Std Dev = 0.0033
- **Random Forest:** Accuracy = 0.9870, Cross-Validation Std Dev = 0.0036

These models showed the highest accuracy and consistency, making them the top performers in our evaluation.

Proof Of Discussion



```
Validation set shape: (1229, 31) (1229,)
Test set shape: (1536, 31) (1536,)
```

```
from sklearn.model_selection import GridSearchCV, StratifiedKFold, cross_val_score
from sklearn.pipeline import Pipeline
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.metrics import classification_report, accuracy_score, precision_score, recall_score, f1_score
```

```
# Define models for baseline
models = {
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'KNN': KNeighborsClassifier(),
    'Random Forest': RandomForestClassifier(random_state=42),
    'Gradient Boosting': GradientBoostingClassifier(random_state=42),
    'SVM': SVC(random_state=42)
}
```

```
[155] # Train baseline models
for name, model in models.items():
    model.fit(X_train, Y_train)
    val_preds = model.predict(X_test)
    val_accuracy = accuracy_score(Y_test, val_preds)
    val_precision = precision_score(Y_test, val_preds, average='weighted')
    val_recall = recall_score(Y_test, val_preds, average='weighted')
```

✓ 0s completed at 9:26 PM

In this meeting, we discussed the work done so far, and Moksh explained further developments regarding the project. We volunteered to take up tasks based on our interests for the project's next steps, i.e., data preprocessing, model building, training the model, and model deployment.

Assigned Roles

Student Name	Roles
Moksh Jaiswal	Data Preprocessing and EDA
Alen Charuvila Saji	Data Preprocessing, Reporting C PPT
Adarsh Shriram Pednekar	Model Deployment
Utsav Harshadbhai Khamar	Data Preprocessing and Model Deployment
Pranay Sai Jangeti	Reporting and PPT
Taranjot Singh Bindra	EDA and Model Building
Smit Rajendraprasad Patel	Model Building
Om Kiranbhai Patel	Model Deployment and Conclusions
Tanzima Mohammadyasin Shaikh	Model Deployment and Conclusions
Aravind Seenivasan	Model Building

Project Management

- **Repository:** https://github.com/pranaysaaij/Thyroid-Disease-Detection-Group-F