Jana Tamer Abdallah 211000963

**Thyroid Disease Classification Project Report**

**1. Introduction**

Thyroid disorders are prevalent endocrine conditions affecting millions worldwide. Early detection and accurate classification of thyroid diseases are crucial for effective treatment. In this project, I obtained the dataset from Kaggle and aimed to develop machine learning models to classify thyroid diseases based on patient data.

**2. Data Description**

The dataset contains the following columns:

age: Age of the patient.

sex: Gender of the patient (0 for male, 1 for female).

TSH: Thyroid-stimulating hormone level.

T3: Triiodothyronine level.

TT4: Total thyroxine level.

T4U: Free thyroxine index.

FTI: Free thyroxine index.

target: Target variable representing different subdiseases of thyroid (0, 1, 2, 3).

**3. Data Acquisition and Preprocessing**

The dataset was sourced from Kaggle, a popular platform for data science competitions and datasets.

Initial exploration revealed missing values and irrelevant features, which I addressed through imputation and feature selection.

Data preprocessing steps included:

Dropping irrelevant columns (e.g., patient\_id, TBG).

Removing rows with missing target values or invalid subdiseases.

Handling missing values using imputation techniques.

Encoding categorical variables (e.g., sex) into numerical format.

Addressing outliers through the IQR method.

**4. Exploratory Data Analysis (EDA)**

I conducted EDA to understand the distribution of features and target classes.

Visualizations such as histograms, boxplots, and pairplots were used to explore feature distributions and correlations.

Relevant features for model training were identified through correlation analysis.

**5. Model Building and Evaluation**

After preprocessing the data, I split it into training and testing sets.

I implemented multiple classification algorithms, including Logistic Regression, Support Vector Machine (SVM), and Random Forest.

To evaluate the performance of each model, I used metrics such as accuracy, precision, recall, and F1-score.

Hyperparameter tuning was conducted to optimize the model performance and prevent overfitting.

**6. Results**

Logistic Regression achieved an accuracy of approximately 97.9%, precision of 98.07%, recall of 97.9%, and F1-score of 97.94%.

SVM yielded an accuracy of approximately 95.8%, precision of 91.90%, recall of 95.8%, and F1-score of 93.81%.

Random Forest exhibited a perfect accuracy of 100%, precision of 100%, recall of 100%, and F1-score of 100%. However, this result may indicate potential overfitting to the training data.

**7. Model Comparison and Selection**

Model performance was compared using accuracy metrics.

While Random Forest demonstrated the highest accuracy, the perfect score raises concerns about overfitting. Therefore, Logistic Regression was selected for its strong performance and slightly higher accuracy compared to SVM.

**8. Deployment**

The trained Logistic Regression model was saved using serialization techniques (pickle) for future use.

Deployment into production for real-time inference on new data was achieved.

**9. Conclusion**

Through this project, I successfully developed machine learning models for thyroid disease classification using a Kaggle dataset.

The Logistic Regression model demonstrated strong performance and was selected for its balanced accuracy and prevention of potential overfitting.

The deployed model can assist healthcare professionals in early diagnosis and treatment planning for thyroid disorders.

**10. Future Enhancements**

Future enhancements could include incorporating additional features or data sources to improve model performance.

Exploration of advanced machine learning techniques such as ensemble methods or neural networks could further enhance the classification accuracy.

Further research could be conducted to enhance model interpretability and explainability, ensuring better understanding and acceptance in clinical practice.

Explanation video: <https://nileuniversity-my.sharepoint.com/:v:/g/personal/j_tamer2163_nu_edu_eg/EflKI5d4DOpMqyAhZbQfAIwBhQPZTuKRkfzJKRBcgx_hWA?nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJTdHJlYW1XZWJBcHAiLCJyZWZlcnJhbFZpZXciOiJTaGFyZURpYWxvZy1MaW5rIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXcifX0%3D&e=54i9w0>