Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037/HJ1: Concepts and Technologies of AI (Herald College, Kathmandu, Nepal)	⋆Final portfolio Project	An End- to- End Machine Learning Project on Regression and Classification Task

Student Id : **2407750**

Student Name : **Shoaib Siddiqui**

Section : **G13**

Module Leader : **Mr. Siman Giri**

Tutor : **Ms. Durga Pokharel**

Submitted on : **February 10, 2025**

Submission Date: February 11, 2025

Classification Task Report

Table of Contents

Abstract

The following report consists of classification analysis according to UNSDG 3: Good Health and Well-being. The project seeks to see if you can tell if an animal is a threat, given its symptoms. It covers a full machine learning pipeline from data preprocessing, featured optimal selection, model development, hyperparameter tuning, evaluation, etc. An optimized Random Forest Classifier yields the best results.
.

---

1. Introduction

1.1 Problem Statement

The classification task will be to predict whether an animal is dangerous based on its symptoms. Early identification of dangerous animals can help in disease prevention and better health management.

1.2 Dataset Description

- **Source**: World Health Organization, 2023 – Global Animal Health Monitoring System.
- **Scope**: Symptoms and characteristics of animals with labels indicating whether they are dangerous.
- **Sustainable Development Goal (SDG)**: Aligns with **UNSDG 3** (Good Health and Well-being).

1.3 Objectives

• Perform EDA to check data distribution.
• Development of classification models, which involves evaluation.
• Optimization through feature selection and hyperparameter adjustment.
• Choosing symptoms that show the degree of danger.

---

2. Methodology

2.1 Data Preprocessing

- **Encoding:** Label encoding on categorical features
- **Feature Engineering:** Symptoms with texts will be embedded with CountVectorizer
- **Class Balance:** Application of SMOTE to balance the classes**.**

2.2 Exploratory Data Analysis (EDA)

- **Descriptive statistics** computed for all features.
- **Correlation heatmap** used to assess relationships between variables.
- **Class distribution visualized** through bar charts.

2.3 Model Development

The following models were implemented and compared:

1. Logistic Regression (from scratch)

2. Logistic Regression (Scikit-Learn)

3. Support Vector Machine (SVM)

4. Decision Tree Classifier

5. Random Forest Classifier (final optimized model)

2.4 Model Evaluation Metrics

• **Accuracy:** The general classification correctness.

• **Precision:** How some of the fantastic class are accurate.

• **Recall:** This measures how many real positives are successfully identified.

• **F1-score:** Harmonic suggest of precision and take into account**.**

• **ROC-AUC score:** It calculates the place under the receiver operating function curve.

2.5 Hyperparameter Optimization

- **Random Forest Classifier**: **n_estimators** and **max_depth** were tuned using **GridSearchCV.**

2.6 Feature Selection

- **SelectKBest** applied to retain the most relevant features.

---

3. Results and Discussion

3.1 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (Scratch)	98.86%	99%	100%	98%
Logistic Regression (Scikit-Learn)	98.86%	99%	100%	98%

SVM	98.86%	99%	100%	98%
Decision Tree	97.71%	99%	98%	98%
Random Forest (Optimized)	98.86%	99%	100%	98%

- **Random Forest performed the best overall**.
- **Severe class imbalance** impacted performance and recall for the minority class.
- **SMOTE improved balance but further tuning is needed**.

---

4. Conclusion

4.1 Key Findings

- **Random Forest Classifier** was the most effective model.
- Class imbalance significantly affected model evaluation metrics.
- Feature selection **reduced dimensionality** without major performance loss.

4.2 Model Improvements

- Hyperparameter tuning **improved Random Forest accuracy**.
- **SMOTE application improved recall** but further refinements are needed.
- Exploring **TF-IDF for feature extraction** could enhance text-based symptom analysis.

4.3 Future Directions

- **Test ensemble learning models** such as XGBoost or AdaBoost.
- **Improve feature engineering** using NLP-based techniques for symptom classification.
- **Implement stratified sampling** to ensure class balance in model training.

---

5. References

- World Organization for Animal Health (2024). Global Animal Health Monitoring Initiatives. Source https://www.woah.org/en/what-we-do/animal-health-and-welfare/
- Brown, A. (2024). *Machine Learning for Healthcare Applications*.

- From dataset data.csv from https://www.kaggle.com/code/kareemellithy/animal-condition-predict-svm-knn/notebook

---