Group Members: Swastik Dubey, Ashish Parida and Anshuman

Algorithms

- **Linear Regression:** Used for prediction tasks by establishing a linear relationship between the features of the dataset and the target
- **Logistic Regression:** Used for binary classification tasks by mapping the input space to [0,1] using the sigmoid function
- **Decision Tree:** Used for classification and Regression by dividing the data into hierarchical structures
- **Support Vector Machine:** Used for classification and Regression by dividing the feature space into 2 halves using a hyperplane
- **AdaBoost:** Used for classification and Regression by utilizing multiple weak learners which create a strong learner by learning from each other's mistakes
- **XGBoost:** Used for classification and regression by using a weak learner and gradually improving it using gradient optimization
- **CatBoost:** Used for classification and regression, it doesn't require a lot of hyperparameter tuning or variable encoding

Modules

- **Pandas:** to parse and handle datasets
- **NumPy:** to perform mathematical operations
- **Matplotlib:** to create graphs and plots
- **Seaborn:** to create heatmaps for correlation matrix
- **Scikit-learn:** provides multiple machine learning algorithms
- **XGBoost:** provides the regressor and the classifier for the same
- **CatBoost:** provides the classifier and regressor for the same

Breast Cancer Classification Report

1. Problem Statement

Breast cancer is a serious health concern affecting many individuals worldwide. The goal of this project is to develop a machine learning model that can classify tumors as either malignant (M) or benign (B) based on various diagnostic features. This classification will help in early detection and treatment of breast cancer.

2. Dataset Description

The dataset used for this project consists of 569 records with 31 features, including:

- **30 numerical features**: These include radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension, among others. Each feature has values derived from three metrics: mean, standard error, and worst case.
- **1 categorical feature**: The target variable "diagnosis," which indicates whether the tumour is malignant (M) or benign (B).

All features are numerical except for the diagnosis column. The dataset does not have missing values, making it suitable for machine learning modelling without significant preprocessing.

Data Pre-processing

Converted the target column to 0 and 1

Top models

These are the models with the best performance on the test data

- **Logistic Regression**: The accuracy of Logistic Regression is **97.37%**. The runtime is 0.02
- **Support Vector Machine:** The accuracy of Support Vector Classifier is **98.25%** . The runtime is 0.00
- **AdaBoost:** The accuracy of AdaBoost is **97.37%**. The runtime is 0.55

Confusion Matrices for the following models are:

Logistic Regression

70	1
2	41

Support Vector Machine

71	0
2	41

AdaBoost

70	1
2	41

4. Choice

We would either go with logistic regression in this case as it provides great performance while being the simplest of the algorithm. We could also go with the SVM model as it has no false negative predictions

Melbourne Housing Price Prediction Report

1. Problem Statement

The objective of this project is to develop a machine learning model to predict house prices in Melbourne based on various property features. Accurate price predictions can help buyers, sellers, and real estate agents make informed decisions.

2. Dataset Description

The dataset consists of 13,580 records and 21 features related to housing prices in Melbourne. Key features include:

- **Categorical Features:** Suburb, Type, Method, SellerG, CouncilArea, RegionName
- **Numerical Features:** Rooms, Price, Distance, Landsize, BuildingArea, YearBuilt, Lattitude, Longtitude, Property

Data Pre-processing

Dropped the columns that were irrelevant to the price target column. One hot encoded 3 columns (Type, ReigonName & Suburb) and dropped the empty rows

Top models

These are the models with the best performance on the test data

- **CatBoost**: The R-square of CatBoost is 0.81. The runtime is 0.82s
- **XGBoost:** The R-square of XGBoost is 0.79, The runtime is 0.76s

4. Choice

We can go with either Catboost or XGBoost, as Catboost provides a better R-square score and can handle missing data much better while XGBoost provides almost as good performance while being slightly faster

Report on Hepatitis Data Analysis and Machine Learning Models

1. Problem Statement

The primary objective of this project is to analyze a hepatitis dataset and develop machine learning models that can help in predicting patient outcomes. Specifically, the goal is to classify whether a patient with hepatitis is likely to have a positive or negative prognosis based on clinical and laboratory data. By leveraging various machine learning algorithms, the analysis aims to improve diagnostic accuracy and support healthcare decision-making.

2. Data Description

The dataset provided contains clinical and laboratory measurements related to hepatitis. Key points about the dataset include:

- **Data Source & Context:**
  The hepatitis dataset is likely sourced from a well-known repository (e.g., UCI Machine Learning Repository) and is used for research on liver diseases.
- **Features:**
  The dataset includes several patient features such as demographic information (e.g., age, gender) and clinical measurements (e.g., bilirubin levels, albumin, alkaline phosphatase, prothrombin time, etc.). Each feature is chosen for its potential relevance in diagnosing or assessing the severity of hepatitis.
- **Target Variable:**
  The target variable is a categorical indicator that represents the patient outcome. This could typically denote the survival status or the stage/severity of the hepatitis condition.
- **Data Characteristics:**
  The dataset may contain missing values, outliers, or imbalanced classes, which necessitate proper pre-processing techniques such as normalization, imputation, and resampling to ensure robust model performance.

Data Pre-processing

Converted all the from (2,1) to 1 and 0 for better classification, performed imputation on all the columns with the normal value of the same test. Finally performed PCA to extract 10 components to decrease the amount of overfitting

Top models

These are the models with the best performance on the test data

- **Logistic Regression**: The accuracy of Logistic Regression is **83.87%.** The runtime is 0.01s
- **CatBoost**: The accuracy of CatBoost is **80.65%.** The runtime is 0.42s

Confusion Matrices for the following models are:

Logistic Regression

3	4
1	23

AdaBoost

1	6
0	24

4. Choice

We should go with logistic regression in this case as it provides better performance while being the simpler of the two. It also has fewer misclassifications and higher True Positive prediction, thus making it more balanced than CatBoost

Report on Australian Credit Approval Data Analysis and Machine Learning Models

1. Problem Statement

The objective of this project is to analyze the Australian Credit Approval dataset to develop machine learning models capable of predicting the approval status of credit card applications. By leveraging various machine learning algorithms, the goal is to enhance the accuracy of credit approval decisions, thereby aiding financial institutions in their assessment processes.

2. Data Description

The dataset used in this analysis is the Statlog (Australian Credit Approval) dataset, which contains information on credit card applications. Key aspects of the dataset include:

- **Data Source & Context:**
  The dataset originates from the Statlog project and is publicly available through the UCI Machine Learning Repository. It is commonly used for benchmarking machine learning algorithms in credit risk assessment.
- **Features:**
  The dataset comprises 15 attributes: 6 numerical and 9 categorical. These attributes represent various aspects of credit applicants, such as demographic information and financial details. All attribute names and values have been anonymized to maintain confidentiality.
- **Target Variable:**
  The target variable is a binary attribute indicating the approval status of the credit application, where typically '1' denotes approval and '0' denotes rejection.
- **Data Characteristics:**
  The dataset consists of 690 instances. It includes a mix of continuous and categorical attributes, some with small numbers of values and others with larger numbers of values. There are also a few missing values present in the dataset.

Data Pre-processing

Scaled the numerical columns using standard scaler

Top models

These are the models with the best performance on the test data

- **Decision Tree**: The accuracy of Logistic Regression is **88.41%.** The runtime is 0.01s
- **CatBoost**: The accuracy of CatBoost is **87.68%.** The runtime is 0.35s

Confusion Matrices for the following models are:

Decision Tree

80	7
9	42

CatBoost

79	8
9	42

4. Choice

We should go with Decision Tree as it is a simpler model and provides the same accuracy as CatBoost. It is also faster to build so scaling it would be easier

Report: Sonar Mines vs. Rocks Classification

1.Problem Statement

The goal of this project is to classify sonar signals to distinguish between metal cylinders (mines) and rocks based on frequency-modulated chirp signals. This classification problem is essential in underwater object detection applications.

2. Dataset Description

- **Source**: The dataset was developed in collaboration with R. Paul Gorman and Terry Sejnowski.
- **Data Composition:** 111 samples of sonar signals reflected from metal cylinders (mines) and 97 samples of sonar signals reflected from rocks.

Each sample consists of 60 continuous attributes representing the strength of sonar signals at different frequencies, along with a label indicating whether the signal is from a mine or a rock.

Data Pre-processing

Converted all the from (R,M) to 1 and 0 for better classification. Finally performed PCA to extract 10 components to decrease the amount of overfitting and improve the accuracy

Top models

These are the models with the best performance on the test data

- **Support Vector Machine**: The accuracy of Support Vector Classifier is **90.48 %**. The runtime is 0.01s
- **XGBoost**: The accuracy of CatBoost is **83.33%.** The runtime is 0.08s

Confusion Matrices for the following models are:

Support Vector Machine

23	3
1	15

XGBoost

21	5
2	14

4. Choice

The obvious choice here is SVM based classifier as it provides higher accuracy with lower misclassifications while being faster

Report on Ionosphere Data Classification

1. Problem Statement

The primary goal of this project is to build a classification model that can effectively distinguish between "good" and "bad" radar returns collected from the ionosphere. These radar returns are indicative of the quality of the ionospheric signal. A "good" signal implies that the radar has detected a clear, predictable pattern, whereas a "bad" signal may indicate noise or interference. The project leverages multiple machine learning algorithms using different libraries to compare model performance and robustness.

2. Data Description

The dataset, known as the Ionosphere dataset, is a well-known collection for binary classification tasks. Key aspects of the data include:

- **Instances:** The dataset consists of 351 examples.
- **Features:** There are 34 continuous features derived from radar signal returns. Each feature represents a characteristic of the signal's reflection from the ionosphere.
- **Target Variable:** The target is a binary class label, typically marked as:
    - **"g"** for a good radar return (indicative of structured, meaningful signals).
    - **"b"** for a bad radar return (indicative of noise or unstructured signals).

Data Pre-processing

Converted all the from (b,g) to 1 and 0 for better classification, scaled the complete dataset as all the features are continuous

Top models

These are the models with the best performance on the test data

- **Support Vector Machine**: The accuracy of SVM is **94.37%**. The runtime is 0.01s
- **XGBoost:** The accuracy of XGBoost is **92.96%.** The runtime is 0.11s
- **CatBoost**: The accuracy of CatBoost is **91.55%.** The runtime is 1.24s

Confusion Matrices for the following models are:

Support Vector Machine

24	4
0	43

XGBoost

23	5
0	43

CatBoost

22	6
0	43

4. Choice

As we see that all the top models perform very similarly, and only differ from each other on 1 false negative misclassification. So we can either go with SVM for simplicity and faster execution or XGBoost for better handling the overfitting due to inherent regularization (L1 & L2)

Report on Heart Disease Classification

1. Problem Statement

The objective of this project is to develop a robust classification model to predict the presence or absence of heart disease using the Statlog (Heart) dataset. This involves analyzing various clinical features of patients to determine whether they have heart disease. The challenge lies in effectively handling the nuances of the data and selecting the best-performing algorithm to accurately identify risk factors.

2. Data Description

The Statlog (Heart) dataset is a widely used dataset in medical research for the prediction of heart disease. Key aspects of the dataset include:

- **Instances:** The dataset comprises several hundred instances (typically around 270–300) representing individual patient records.
- **Features:** The dataset includes various clinical attributes such as age, sex, chest pain type, resting blood pressure, cholesterol levels, maximum heart rate achieved, exercise-induced angina, and more.
- **Target Variable:** The target is a binary variable indicating the presence (or absence) of heart disease.

Data Pre-processing

Converted all the from (2,1) to (1,0) and (3,6,7) to (0,1,2) for better model learnability, performed scaling on only the numerical columns

Top models

These are the models with the best performance on the test data

- **Logistic Regression**: The accuracy of Logistic Regression is **92.59%**. The runtime is 0.01s
- **CatBoost**: The accuracy of CatBoost is **88.89%.** The runtime is 0.34s

Confusion Matrices for the following models are:

Logistic Regression

32	1
3	18

CatBoost

32	1
5	16

4. Choice

We should go with logistic regression in this case as it provides better performance while being the simpler of the two. It also has fewer False Positive misclassifications, thus making it more robust than CatBoost in this case