PAPER PRESENTATION

Team No -04

Group Members

Fairuz Tassnim Prapty | 21101027

Sabiha Alam Chowdhury | 20301192

Ashakuzzaman Odree | 20301268

Syed Ashik Mahamud | 20301124

RA : EHSANUR RAHMAN RHYTHM

ST : MEHNAZ ARA FAZAL

TOPIC

Identification of Iris Flower Varieties Through Machine Learning Techniques

OVERVIEW

- Abstract
- Introduction
- Literature review
- Methodology
- Dataset
- Data Acquisition and Preparation
- Choosing the Model
- Implementation of Cross Validation
- Utilizing the Confusion Matrix
- Conclusion

ABSTRACT

- Dataset and Process:
  - Iris dataset used for implementation.
  - Scikit-learn toolkit employed for implementation.
- Paper's Objective:
  - Employing classification and regression algorithms on the IRIS dataset.
  - Identification and examination of patterns based on sepals and petals sizes.
- Key Finding:
  - SVM classifier outperforms KNN and logistic regression models.
  - Higher accuracy achieved with the SVM classifier.

ABSTRACT

- Classification in Machine Learning:
  - Crucial for analyzing data.
  - Involves a diverse range of algorithms.
- Algorithms for Classification:
  - Decision trees, Naive Bayes, backpropagation, neural networks.
  - Artificial neural networks, multi-layer perceptrons.
  - Multi-class classification, Support Vector Machines (SVM), K-nearest neighbors (KNN).
- Research Focus:
  - Elaborated on three specific classification methods.
  - Implementation using the iris dataset and Scikit-learn toolkit.

INTRODUCTION

- **Machine Learning Overview:**
  - **Subset of computer science focused on creating adaptable programs through self-improvement when exposed to new data. Broadly categorized into supervised and unsupervised learning, with an emphasis on supervised learning involving classification and regression.**

- **Iris Species Identification Study:**
  - **Utilizes Fisher's Iris dataset for identifying Iris flower species. Aiming for high accuracy in predicting unseen data, the study employs supervised learning techniques, particularly classification. The Scikit-learn toolkit is used to implement various machine learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression classifiers.**

METHODOLOGY

- Implemented three key machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbor (KNN) classifiers.
- Employed four essential features from the iris dataset to train and test the classification models.
- Utilized the Python-based scikit-learn toolkit for seamless implementation and execution of the chosen algorithms.
- Conducted a comprehensive comparative analysis of the accuracy of SVM, Logistic Regression, and KNN models.

DATASET

- **Flower Species: i. Iris-setosa, ii. Iris-versicolor, iii.Iris-virginica.**
- Sample Size: i.150 individual samples, ii.50 samples per species
- Each sample includes measurements of four distinct features, capturing the morphological differences.
- Graphical illustration of iris flower samples and their measured features.
- A comprehensive dataset providing valuable insights into the morphological variations among Iris-setosa, Iris-versicolor, and Iris-virginica.

LITERATURE REVIEW

Machine Learning Approaches to IRIS Dataset

1

- Deeptam Dutta et al :
  - Applied Artificial Neural Networks for classification.
  - Focused on pattern recognition and predictive modeling.

2

- Poojitha A et al :
  - Utilized MATLAB for unsupervised clustering with k-means.
  - Emphasized neural network tools for large dataset categorization.

LITERATURE REVIEW

Advanced Methodologies and Statistical Analysis

3

- Vaishali Arya et al.:
  - Introduced a neural fuzzy system for feature selection and rule derivation.
  - Enhanced efficiency in the classification process.

4

- Shashidhar T et al. and Patrick S. et al.:
  - Developed models to predict and forecast IRIS species characteristics.
  - Combined statistical patterns and Java application for data analysis.

DATA ACQUISITION AND PREPARATION

- Obtained from the UCI Machine Learning Repository.
- Comprises 150 samples across three Iris species: setosa, versicolor, and virginica.
- Dataset stored in the variable "Iris" for ease of reference.
- Imported using the scikit-learn toolkit.
- Split into training (60%) and testing (40%) subsets.
- Ensures stable and consistent accuracy assessments across multiple model runs.

CHOOSING THE MODEL

Support Vector Machine (SVM)

- SVM serves as a powerful technique for classifying datasets, whether linear or nonlinear.
- It employs non-linear mapping to project training data into a higher-dimensional space.
- Each data item plotted in an n-dimensional space (n = number of features).
- Conducted classification within this transformed space.
- SVM searches for an optimal separating hyperplane in the transformed space.

Logistic Regression

- Logistic Regression, a statistical technique, analyzes datasets with one or more independent variables influencing the outcome.
- Primarily used for accurate data categorization based on existing information.
- Utilized for segmenting Iris flower data based on length and width attributes.
- Functions effectively, particularly with larger datasets.
- Logistic regression applied to the Iris dataset resulted in a model accuracy of 91%.

K-Nearest Neighbor Classifier (KNN)

- KNN is versatile, handling both classification and regression problems in supervised learning.
- Straightforward and entirely reliant on the training dataset.
- Classifies incoming data based on similarity measured by the distance between instances.
- KNN classifier implemented using the KNeighborsClassifier(n_neighbors=3) function from the sklearn.neighbors package.
- Chosen and evaluated for its effectiveness in predicting Iris flower species.

IMPLEMENTATION OF CROSS-VALIDATION

- Cross-validation involves reserving a portion of the dataset for validation, not used during model training.
- Utilize the reserved sample from the test set to assess the model's performance.
- A model exhibiting favorable results is deemed effective.
- Precise estimate from sample accuracy.
- Enhances model efficiency and effectiveness.
- Comparison of SVM, KNN, and Logistic Regression accuracy with and without cross-validation.
- Analysis reveals improved accuracy when cross-validation is implemented, enhancing overall model performance.

UTILIZING THE CONFUSION MATRIX

- A tabular representation used to evaluate the efficacy of a classification model.
- Utilizes test data with established expected output labels.
- Provides a straightforward indication of prediction accuracy.
- Aids in pinpointing errors made by the model.
- Denotes the target, indicating the classification label for the given sample data.
- The confusion matrix contributes to the determination of the model's accuracy score.
- The confusion matrix acts as a crucial tool in assessing the accuracy and reliability of the classification model, specifically in the context of the Iris dataset.

CONCLUSION

- Algorithmic Powerhouse:
  - SVM, KNN, and Logistic Regression play pivotal roles in extracting patterns.
- Validation Excellence:
  - Cross-validation ensures robust model performance and reliable insights.
- SVM Triumph:
  - Support Vector Machine emerges as the most effective in classification tasks.
- Library Strength:
  - Explore essential functions of the scikit-learn library for seamless implementation.

THANK YOU