

**BỘ GIÁO DỤC VÀ ĐÀO TẠO TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN**

-----🕮-----

TIỂU LUẬN

**Học phần: Ứng dụng Trí tuệ nhân tạo trong kinh doanh và quản lý**

**Topic: Research and implement a software program using the ID3 decision tree technique to classify flower species in the test dataset. From the classification results and the data in the test dataset, evaluate the effectiveness of the classification model**

**Nhóm sinh viên : Nhóm 2**

**Lớp học phần :** **TIHT1123E(224)CLC\_01**

**Giảng viên hướng dẫn : TS. Lưu Minh Tuấn**

***Hà Nội, Tháng 4 Năm 2025***

**DANH SÁCH THÀNH VIÊN NHÓM 2**

|  |  |  |
| --- | --- | --- |
| STT | Họ và Tên | Mã Sinh Viên |
| 1 | Vũ Trọng Đại | 11221186 |
| 2 | Nguyễn Quang Huy | 11222804 |
| 3 | Đặng Bá Lộc | 11223872 |
| 4 | Nguyễn Nam Thái | 11225729 |
| 5 | Nguyễn Anh Tú | 11226661 |

# MỤC LỤC

[MỤC LỤC 3](#_Toc196659319)

[Introduction 6](#_Toc196659320)

[Chapter 1: Topic overview 7](#_Toc196659321)

[**1.1** **Research Subject and Scope** 7](#_Toc196659322)

[**1.2 Research Methodology** 7](#_Toc196659323)

[**1.3 Research Subject and Scope** 7](#_Toc196659324)

[**1.4 Research Methodology** 8](#_Toc196659325)

[**1.5 Structure of the Thesis** 8](#_Toc196659326)

[Chapter 2. Theoretical Background 9](#_Toc196659327)

[**2.1. Supervised Learning** 9](#_Toc196659328)

[**2.2.ID3 decision tree algorithm** 9](#_Toc196659329)

[**Chapter 3. Experimentation and Results** 12](#_Toc196659330)

[**3.1. Supervised Learning** 12](#_Toc196659331)

[**3.2. Checking the input data** 12](#_Toc196659332)

[**3.3. Model design** 21](#_Toc196659333)

[**3.4. Design of the Iris Flower Classification Application Based on the ID3 Model** 26](#_Toc196659334)

[**Chapter 4. Conclusion and Recommendations** 31](#_Toc196659335)

[**4.1. Conclusion** 31](#_Toc196659336)

[**4.2.Achievements** 32](#_Toc196659337)

[**4.3. Limitations** 32](#_Toc196659338)

[**4.4. Recomedations** 32](#_Toc196659339)

[TÀI LIỆU THAM KHẢO 36](#_Toc196659340)

**DANH MỤC TỪ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
| STT | Từ viết tắt | Ý nghĩa |
| 1 | ID3 | Iterative Dichotomiser 3 |

## Introduction

*Machine learning, a subset of artificial intelligence (AI), encompasses algorithms trained to identify patterns and correlations within large datasets, enabling optimal decision-making and predictions based on that analysis. In real-world applications, machine learning is widely utilized in areas such as data retrieval systems, medical diagnostics, credit card fraud detection, stock market analysis, DNA sequence classification, speech and handwriting recognition, and automated translation, among others. To ensure accuracy and prevent biased or misleading data, machine learning still requires human intervention in the process of understanding and selecting appropriate data analysis techniques.*

*Given its broad applications, our team decided to explore the problem of predicting wine quality using machine learning methods. Within the scope of this essay, we aim to address the research topic: “Research and implement a software program using the ID3 decision tree technique to classify flower species in the test dataset. From the classification results and the data in the test dataset, evaluate the effectiveness of the classification model.” Specifically, this essay focuses on investigating the application of the ID3 decision tree technique.*

*Due to our limited knowledge, this essay may contain errors or lack comprehensiveness. We sincerely welcome feedback and suggestions from our instructor to help us refine and improve our research in the future.*

*We express our heartfelt gratitude!*

## Chapter 1: Topic overview

### **Research Subject and Scope**

In the context of the Fourth Industrial Revolution, artificial intelligence (AI) has become a vital tool for businesses to optimize management processes and decision-making. The application of AI not only enhances operational efficiency but also provides a competitive edge through rapid and accurate data analysis. This topic, focusing on AI application in data classification (using the Iris dataset) and testing the ID3 decision rule, highlights AI's potential to address practical needs in modern business management.

### **1.2 Research Methodology**

* Develop a detailed report (in text or visual format) on the application of AI in business and management, using the Iris dataset.
* Build and test a software program applying the ID3 decision rule for data classification, evaluating AI's effectiveness in classification and decision-making.
* Provide research documentation and a directory containing the program code, clearly explaining the purpose, methodology, and experimental results.

### **1.3 Research Subject and Scope**

* Research Subject: The Iris dataset (comprising three types: setosa, virginica, versicolor) with attributes such as sepal length, sepal width, petal length, and petal width; the ID3 decision rule in data classification.
* Research Scope: Application of AI in business and management, focusing on testing the ID3 algorithm and data analysis, conducted from March 17, 2025, to April 29, 2025.

### **1.4 Research Methodology**

* Data Analysis Method: Utilize the Iris dataset to analyze attributes and apply the ID3 algorithm for classification.
* Programming Method: Develop a software program to test the ID3 decision rule, using an appropriate programming language (as per the topic requirements).
* Reporting Method: Compile research findings into a detailed report in English, including documentation files and a code directory.

### **1.5 Structure of the Thesis**

* Chapter 1: Topic Overview (significance, objectives, subject, scope, and research methodology).
* Chapter 2: Theoretical Background (concepts of AI, the ID3 algorithm, and its application in business and management).
* Chapter 3: Experimentation and Results (analysis of the Iris dataset, program testing, and performance evaluation).
* Chapter 4: Conclusion and Recommendations (summary of findings and suggestions for practical application).
* References: List of sources used in the research.

## Chapter 2. Theoretical Background

### **2.1. Supervised Learning**

Supervised machine learning is a technique in the field of Artificial Intelligence (AI) that uses labeled data to train a predictive model. The training data consists of input-output pairs, where the input is typically a feature vector, and the output is the corresponding target label.

Two main types of problems in supervised machine learning:

* Classification: Predicting discrete class labels (e.g., flower species).
* Regression: Predicting continuous values (e.g., house prices).

In this topic, the problem is a classification task—predicting flower species (e.g., Iris setosa, Iris versicolor, Iris virginica) based on features such as petal and sepal length/width.

### **2.2.ID3 decision tree algorithm**

#### **2.2.1.**

A **Decision Tree** is a supervised machine learning model that uses a tree structure to represent the decision-making process. Each node in the tree corresponds to a feature, each branch represents a value of that feature, and each leaf denotes an output label.

The decision tree algorithm is intuitive, interpretable, and widely used in classification tasks.

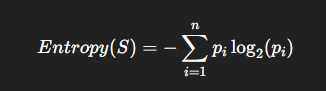
#### **2.2.2. ID3 algorithm**

**ID3 (Iterative Dichotomiser 3)** is one of the most popular decision tree algorithms. It operates on the principle:

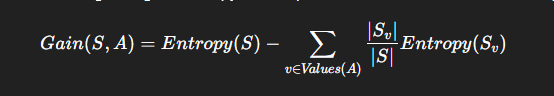
* Select the best feature to split the data based on the Information Gain criterion.
* Continue splitting the data until: all data at a node belongs to the same class, no features remain for further splitting, or the maximum specified depth is reached.

#### **2.2.3. Entropy and Information Gain**

* **Entropy:** Measures the level of uncertainty in a dataset. Lower entropy indicates "cleaner" data, making it easier to classify.
* **Entropy fomula:**



* **Information Gain:** Measures the reduction in entropy after splitting the data according to a feature A
* **Calculation formula:**



#### **2.2.4. Advantage and Disadvantage of ID3:**

**Advantage**:

* Easy to understand, implement, and visualize.
* Can handle both qualitative and quantitative data.
* Enables fast classification after training.

**Disadvantages:**

* Prone to overfitting if tree depth is not limited or pruning is not applied.
* Sensitive to noisy data and irrelevant attributes.
* Only suitable for discrete data; continuous data must be discretized for basic ID3.

### **Chapter 3. Experimentation and Results**

### **3.1. Supervised Learning**

Supervised machine learning is a technique in the field of Artificial Intelligence (AI) that uses labeled data to train a predictive model. The training data consists of input-output pairs, where the input is typically a feature vector, and the output is the corresponding target label.

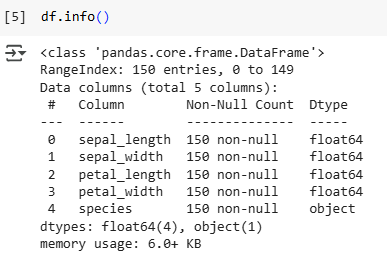
Two main types of problems in supervised machine learning:

* Classification: Predicting discrete class labels (e.g., flower species).
* Regression: Predicting continuous values (e.g., house prices).

In this topic, the problem is a classification task—predicting flower species (e.g., Iris setosa, Iris versicolor, Iris virginica) based on features such as petal and sepal length/width.

### **3.2. Checking the input data**

#### **3.2.1. Overview of the data**



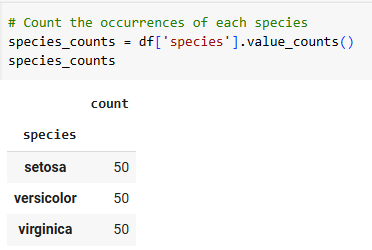
The dataset used in the analysis consists of 150 samples (entries), each corresponding to information about an Iris flower. Each sample in the dataset contains details on the flower’s morphological characteristics, specifically the length and width of the sepal and petal, along with the corresponding species name. The data is stored as a DataFrame object using the pandas library in Python, with 5 columns and 150 rows.

Below is a detailed description table of the columns in the dataset:

|  |  |  |
| --- | --- | --- |
| Tên cột | Kiểu dữ liệu | Mô tả |
| sepal\_length | float64 | Chiều dài của đài hoa (sepal), đơn vị tính là centimet |
| sepal\_width | float64 | Chiều rộng của đài hoa (sepal), đơn vị centimet |
| petal\_length | float64 | Chiều dài của cánh hoa (petal), đơn vị centimet |
| petal\_width | float64 | Chiều rộng của cánh hoa (petal), đơn vị centimet |
| species | object | Loài của hoa, là một giá trị dạng chuỗi (categorical) |

In terms of data quality, all columns contain 150 non-null values, meaning there are no missing or incomplete entries throughout the entire dataset. This ensures that the data is fully prepared for analysis or machine learning model training without the need for additional handling of missing values.

#### **3.2.2. Data distribution by species**



After examining the data by species groups, the results show that the dataset is evenly distributed among the three Iris flower species: Setosa, Versicolor, and Virginica. Specifically, each species has exactly 50 samples, accounting for one-third of the total dataset. This result is presented in the following table:

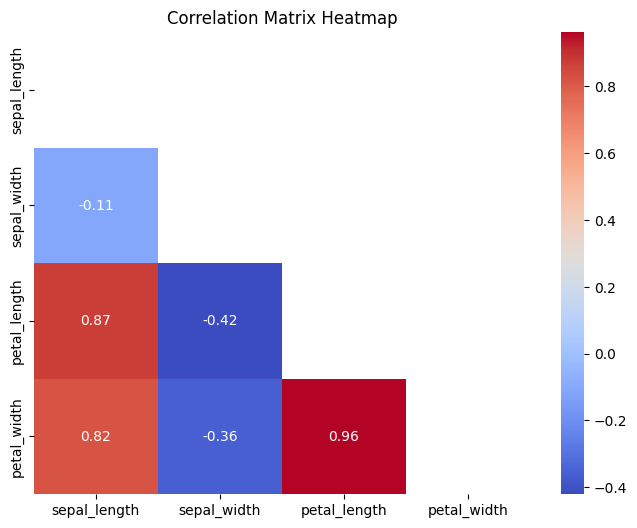
|  |  |
| --- | --- |
| Loài hoa (species) | Số lượng mẫu (count) |
| Setosa | 50 |
| Versicolor | 50 |
| Virginica | 50 |

Such an even distribution across the label classes is an ideal characteristic for classification tasks, as it helps prevent the issue of class imbalance. As a result, the machine learning model, when trained on this dataset, will not be biased toward any particular species and will be able to make fair and balanced predictions for all three classes.

Moreover, this uniform distribution also facilitates the evaluation of the model’s performance using metrics such as accuracy, recall, and precision, as the number of samples in each class is equal.

#### **3.2.3. Correlation matrix analysis**

To gain a deeper understanding of the relationships between the attributes in the dataset, we calculate the Pearson correlation coefficients between the pairs of quantitative variables, including sepal\_length, sepal\_width, petal\_length, and petal\_width. The results are visualized as a heatmap, as shown below:



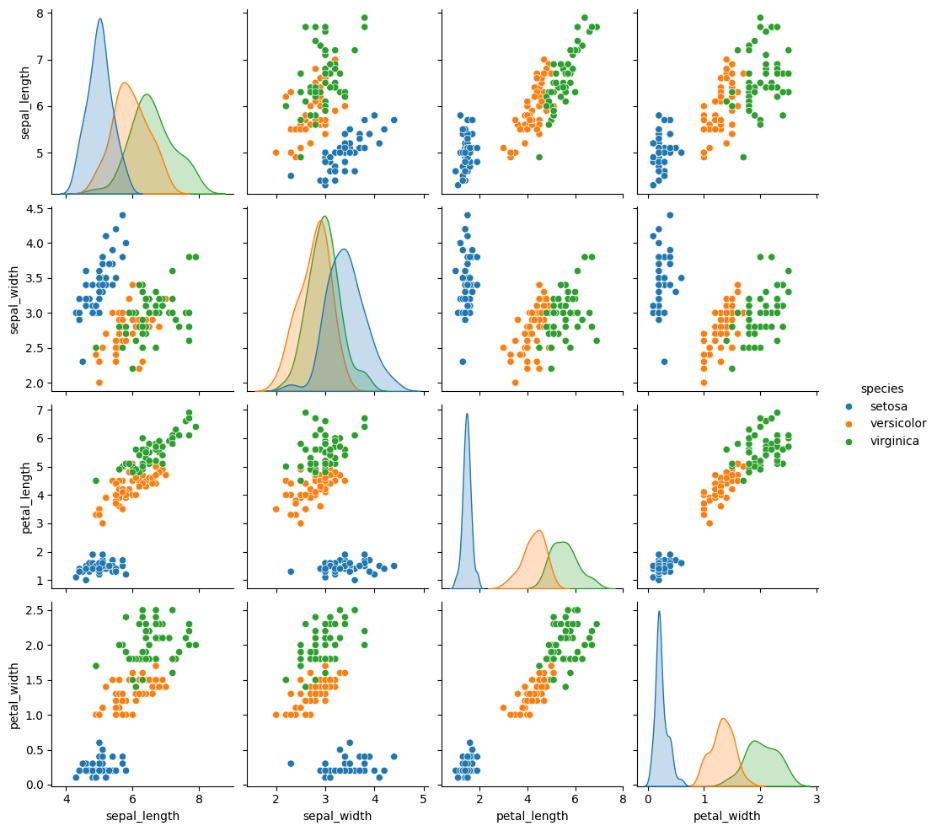
In the heatmap, each square represents the Pearson correlation coefficient, with a color scale ranging from blue (negative correlation) to red (positive correlation). Values closer to 1 or -1 indicate a stronger relationship, while values near 0 suggest a weak or no correlation between the variables.

**Some key observations from the heatmap include:**

* **Petal length and petal width have a very high correlation (0.96):** This indicates that as the petal length increases, the petal width also increases in a nearly linear fashion. This is the strongest correlation in the entire matrix, clearly demonstrating the close relationship between these two features.
* **Sepal length has a fairly high correlation with petal length (0.87) and petal width (0.82):** Although not as strong as the petal pair, sepal length still shows a significant relationship with the petal-related features.
* **Sepal width has a negative correlation with other features, especially with petal length (-0.42) and petal width (-0.36):** This is noteworthy because sepal width tends to decrease as the other attributes increase. However, the strength of this negative correlation is moderate.
* **Sepal length and sepal width have a very weak correlation (-0.11):** This suggests that there is almost no linear relationship between these two features.

Detecting high correlations between features like the ones mentioned above can be useful for feature selection in machine learning models. Additionally, the strong relationship between the petal-related features suggests that these attributes have a good potential for distinguishing between the different Iris species.

#### **3.2.4. Evaluating Feature Distribution and Pairwise Relationships through Pair Plot in ID3 Model Building**



**Chart Structure**:

* **Main Diagonal:** The density plot for each feature on the main diagonal shows the distribution of each feature by species. These plots display the distribution of each feature for each species.

**The Remaining Sections:** The scatter plots show the relationships between pairs of features, with the points color-coded by species (Setosa: blue, Versicolor: orange, Virginica: green). These plots visually depict how different feature combinations relate to each other and how well the species are separated based on these relationships.

**Distribution of Each Feature (Main Diagonal):**

* **Sepal Length**: Setosa has the smallest sepal length, Versicolor is in the middle, and Virginica has the largest sepal length. There is some overlap between Versicolor and Virginica.
* **Sepal Width**: The distributions of sepal width are quite close between the species, with Setosa having a slightly larger sepal width, but there is significant overlap between the species.
* **Petal Length**: There is a clear distinction: Setosa has the smallest petal length (below 2), Versicolor ranges from about 3 to 5, and Virginica has the largest petal length (above 4.5). There is little overlap, making this feature very effective for classification.
* **Petal Width**: Similar to petal length, Setosa has the smallest petal width (below 1), Versicolor ranges from about 1 to 2, and Virginica is above 1.5. This feature also has a high ability to differentiate between the species.

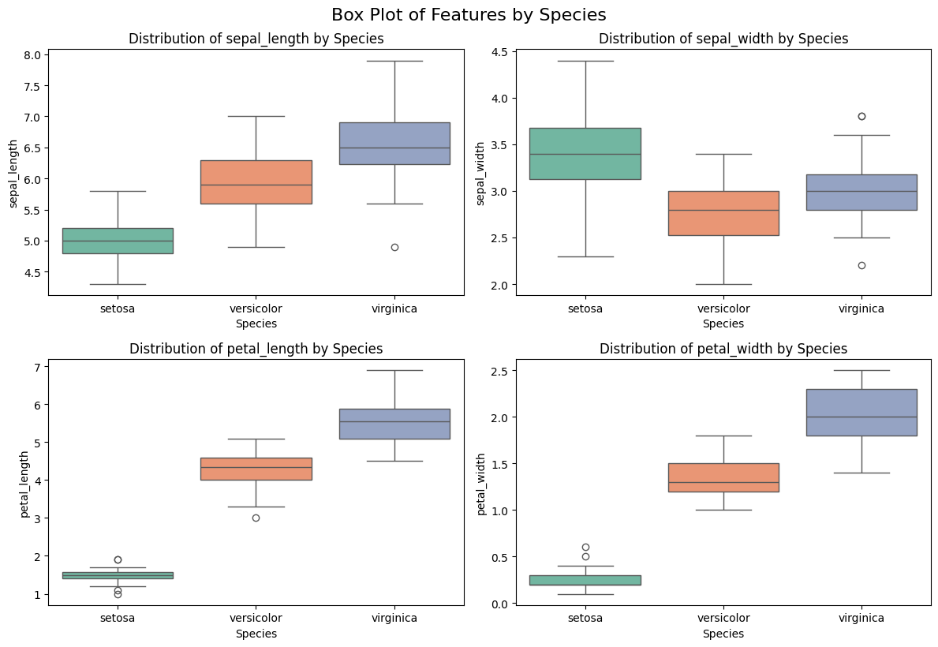
**Relationship Between Feature Pairs (Scatter Plot)**

* **Petal Length vs Petal Width**: The distinction between species is very clear. Setosa is completely separated, while Versicolor and Virginica show a slight overlap but can still be distinguished. This is the best feature pair for classification.
* **Petal Length vs Sepal Length**: Setosa remains separated, but Versicolor and Virginica overlap more compared to petal length vs petal width.
* **Sepal Length vs Sepal Width**: There is considerable overlap between species, especially between Versicolor and Virginica, making this pair less effective for classification.
* **Petal Width vs Sepal Width**: Setosa is separated, but there is significant overlap between Versicolor and Virginica.

**General Observations:**

* **Best Discriminating Features:** Petal length and petal width are the two features with the highest ability to differentiate between species, especially with Setosa being completely separated.
* **Least Discriminating Feature:** Sepal width has the lowest ability to differentiate between species due to the significant overlap between the species.
* **Relationship Between Features:** Based on the plots, features like petal length and petal width have high discriminative power, indicating that they can be selected as splitting features to optimize purity at the root nodes of the ID3 decision tree. This helps in reducing entropy quickly.

#### **3.2.5. Analyzing Feature Distribution and Outliers through Box Plot in ID3 Model Building**



The box plot displays the distribution of the four features (sepal length, sepal width, petal length, petal width) for each Iris species (Setosa, Versicolor, Virginica), focusing on the median, interquartile range (IQR), and outliers. This visualization helps to identify the central tendency (median), the spread of the data (IQR), and any potential outliers for each feature and species. It also provides insights into the variation and potential overlap between species for each feature.

**Additional Information Compared to Pair Plot**:

**Median and Dispersion**:

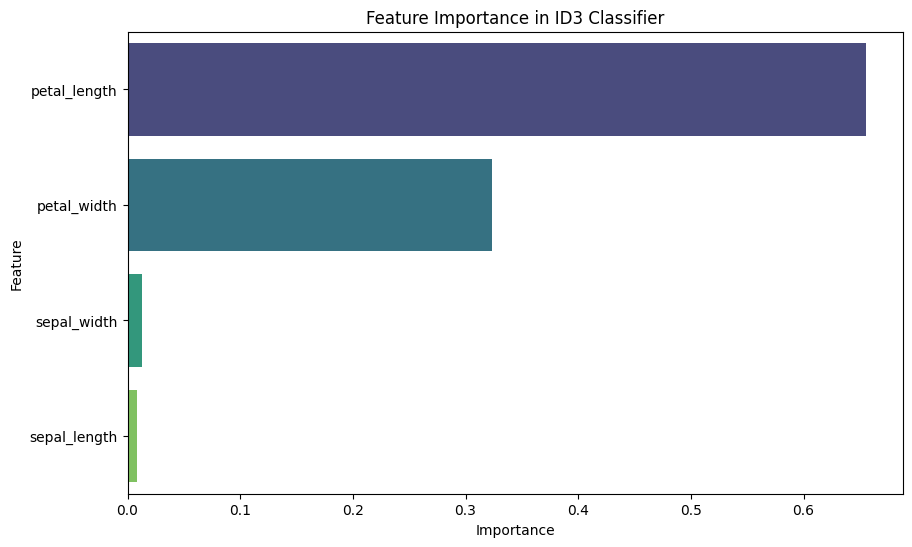
* Petal length: Setosa (~1.5), versicolor (~4.3), virginica (~5.6).
* Petal width: Setosa (~0.2), versicolor (~1.3), virginica (~2.0).
* Sepal length: Setosa (~5.0), versicolor (~5.9), virginica (~6.5).
* Sepal width: Setosa (~3.4), versicolor (~2.8), virginica (~3.0).

**Outliers: Sepal width has the most outliers (Setosa: 2.0 and 4.4; Virginica: 2.2), while other features have fewer outliers (e.g., Sepal length for Versicolor is around 7.7).**

**Application for ID3**:

* **Petal length and petal width** are still good features for splitting due to the clear difference in median, which aligns with the results from the pair plot.
* **Outliers** (especially in sepal width) can affect the splitting in ID3 and need to be addressed (e.g., by removing or replacing with the median).

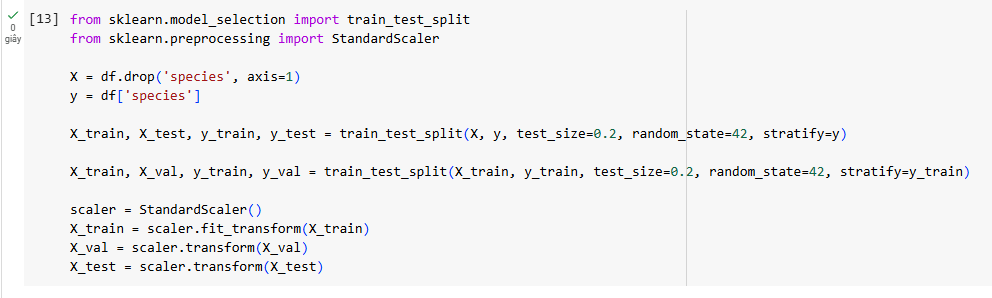
#### **3.2.6. Evaluating the Importance of Features in the ID3 Model**



The chart illustrates the feature importance of the ID3 model when classifying the Iris dataset. The results show that petal length has the highest importance (around 0.6), followed by petal width (about 0.35), while sepal width and sepal length contribute very little (below 0.05). This aligns with the analysis from the pair plot and box plot, where petal length and petal width were identified as strong discriminative features between the flower species. These results confirm that the ID3 model primarily relies on petal length and petal width to construct the decision tree and suggest that sepal length and sepal width could potentially be considered for removal to simplify the model without significantly affecting performance.

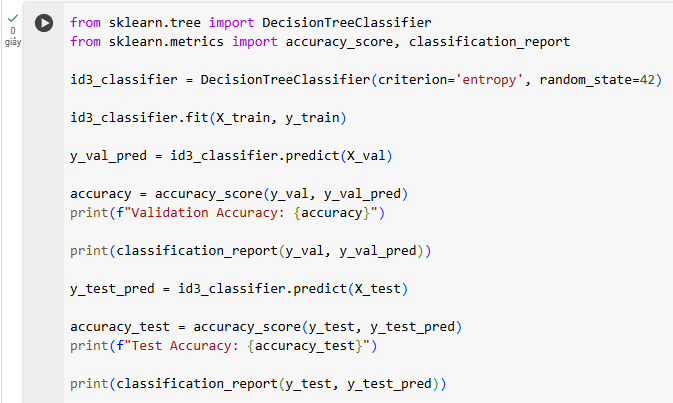
### **3.3. Model design**

#### **3.3.1. Data Normalization and Splitting**

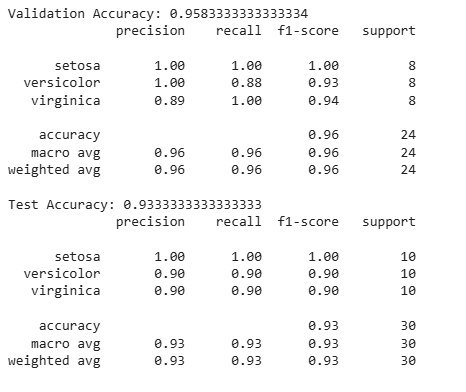


Split the data into training (64%), validation (16%), and test (20%) sets, ensuring that the species ratio is maintained using the stratify parameter. The data is also normalized to ensure consistency. These steps support training, tuning, and evaluating the ID3 model, ensuring generalization and avoiding bias.

#### **3.3.2. Training and Evaluating the Model**

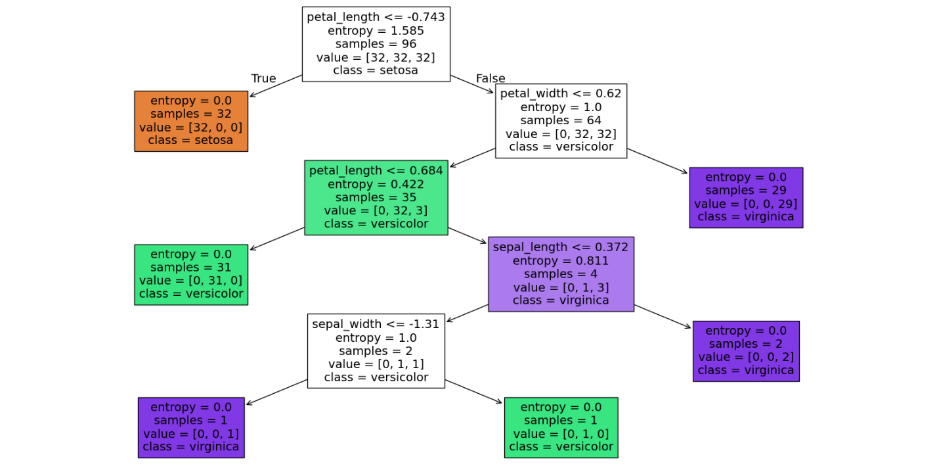


This stage focuses on building, training, and evaluating the ID3 model for classifying flower species in the Iris dataset. The main idea is to implement a decision tree using the entropy criterion to optimize feature selection for splitting, train the model on the training set, and then predict and evaluate on the validation and test sets. Performance is measured through accuracy and detailed metrics such as precision, recall, and F1-score for each species (Setosa, Versicolor, Virginica), in order to assess the model's classification ability and generalization on new data, as well as to detect any potential overfitting. This process helps confirm the effectiveness of ID3 in the classification task and provides a foundation for fine-tuning or comparing with other models.



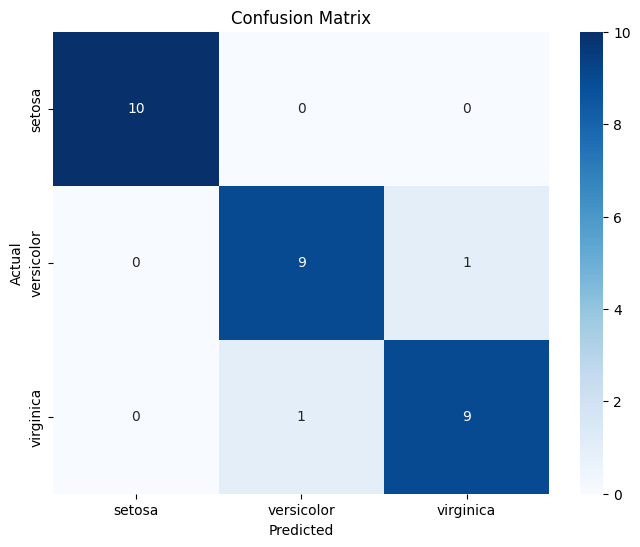
The ID3 model achieves high performance on both the validation and test sets of the Iris dataset. On the validation set, the accuracy reaches 0.96, with Setosa being classified perfectly (precision, recall, F1-score all 1.00), while Versicolor and Virginica perform well, but Versicolor is slightly misclassified (recall 0.88). On the test set, the accuracy is 0.93, with Setosa continuing to perform perfectly, while Versicolor and Virginica both achieve precision, recall, and F1-score of 0.90, demonstrating stable classification performance. The small difference between the validation set (0.96) and the test set (0.93) indicates that the model is not overfitting and has good generalization ability.

#### **3.3.3. Analysis of Decision Tree Structure of ID3 Model**



The chart above illustrates the decision tree structure built by the ID3 model on the Iris dataset, showing how the model classifies the flower species based on features. The tree starts with the root node using the condition **petal\_length <= -0.743**, splitting the data into two branches: the left branch (True) perfectly classifies Setosa (entropy = 0, 32 samples), and the right branch (False) continues to split based on **petal\_width <= 0.62**. The tree further splits with conditions such as **petal\_length <= 0.684** and **sepal\_length <= 0.372**, ultimately classifying samples into Versicolor and Virginica with entropy decreasing progressively (from 1.585 to 0 at the leaf nodes). This result demonstrates that **petal length** and **petal width** play a crucial role in the classification, which aligns with the earlier feature importance analysis.

#### **3.3.4. Confusion Matrix Analysis of ID3 Model on Test Set**



The confusion matrix visualizes the classification performance of the ID3 model on the test set of the Iris dataset. The matrix shows the model correctly classified 10/10 setosa samples, 9/10 versicolor samples, and 9/10 virginica samples. However, there were minor misclassifications: 1 versicolor sample was incorrectly predicted as virginica, and 1 virginica sample was incorrectly predicted as versicolor. This aligns with the previously reported test set accuracy of 0.93 and indicates that the model performs perfectly on setosa but needs improvement in distinguishing versicolor and virginica, possibly due to overlap between these species, as observed in pair plots and box plots. The confusion matrix provides detailed insight into classification errors, aiding in model or data refinement to enhance performance.

### **3.4. Design of the Iris Flower Classification Application Based on the ID3 Model**

#### **3.4.1. Adjusting Feature Parameters Using Sliders**

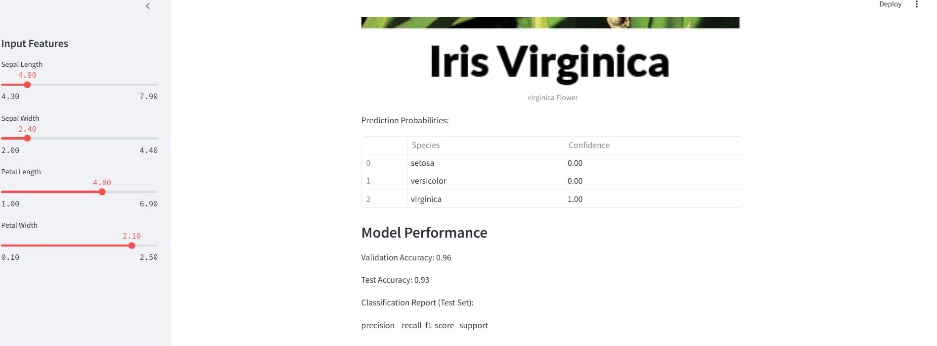
To illustrate the practical application of the ID3 model, a web application has been developed using Streamlit, allowing users to interact directly with the Iris flower classification model. This application provides a user-friendly interface that supports data input, prediction, data visualization, and result comparison. The main features of the application are described in detail below.



The application allows users to adjust the features of the Iris flower (sepal length, sepal width, petal length, petal width) through sliders. Each slider is set with an appropriate range of values based on the distribution of the Iris dataset (e.g., sepal length from 4.30 to 7.90, petal width from 0.10 to 2.50). Users can easily input feature values by sliding the bars, and then the application will use the trained ID3 model to predict the corresponding flower species. This functionality enables users to interact directly with the model, test the classification capability of ID3 across different feature values, and demonstrate the flexibility of the model in real-world scenarios.

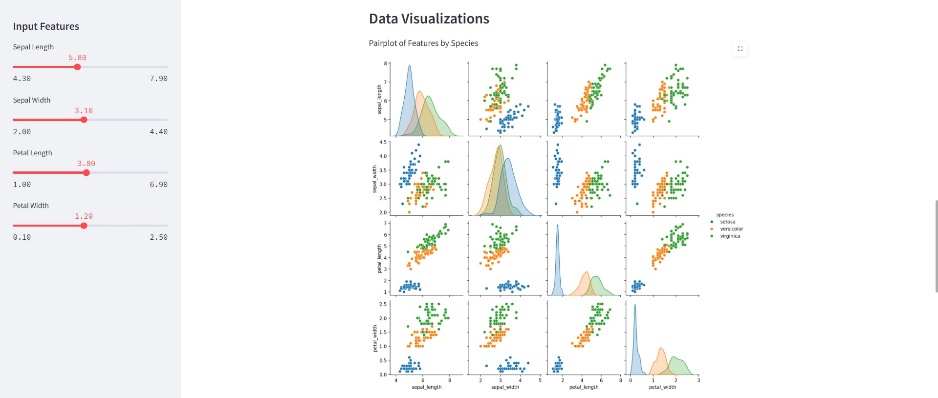
#### **3.4.2. Predicting Flower Species and Displaying Prediction Probabilities**

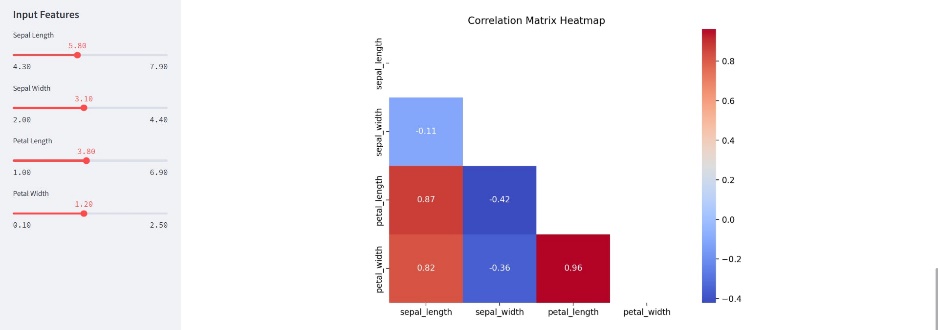
****

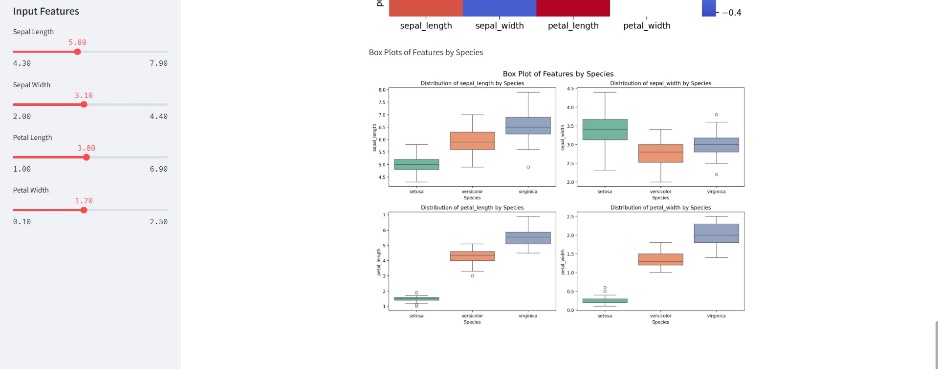
****

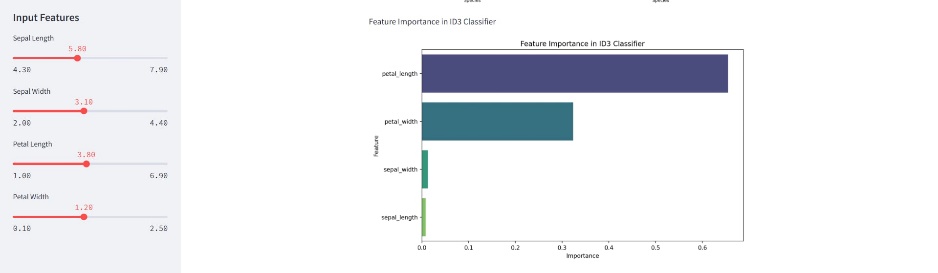
After the user adjusts the features, the application displays the predicted Iris flower species (Setosa, Versicolor, or Virginica) based on the ID3 model. For example, with feature values like **sepal length = 5.80**, **sepal width = 3.10**, **petal length = 3.80**, and **petal width = 1.20**, the application predicts the flower species as **Versicolor**; or with **petal length = 4.80** and **petal width = 2.10**, the predicted result is **Virginica** with a probability of **1.00**. The application also shows the prediction probability for each species, providing a clear insight into the confidence level of the prediction. This functionality not only helps users understand the classification results but also illustrates how the ID3 model makes decisions based on input features, while increasing the transparency of the prediction process.

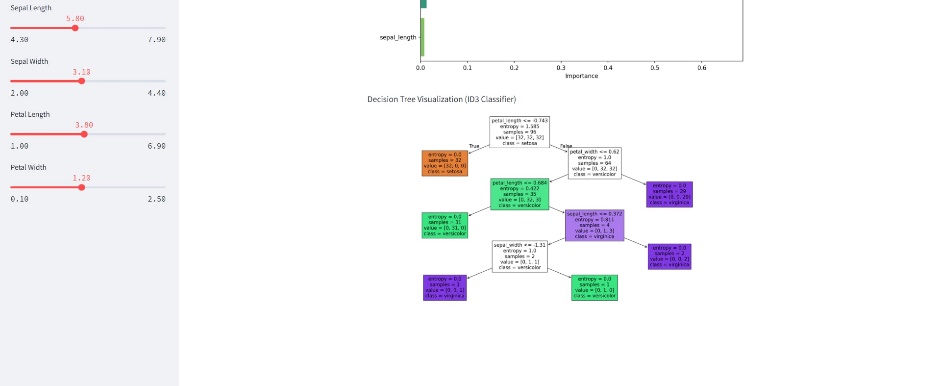
#### **3.4.3. Displaying Visual Charts**

****

****

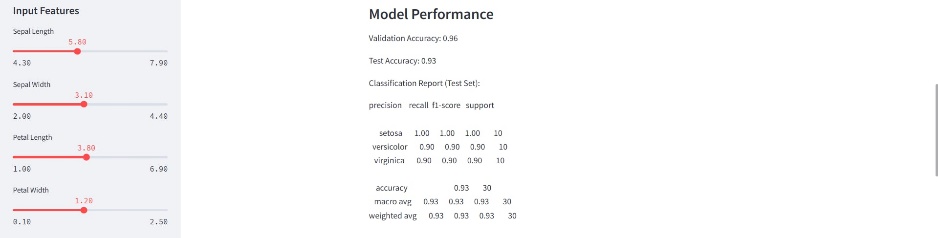
****





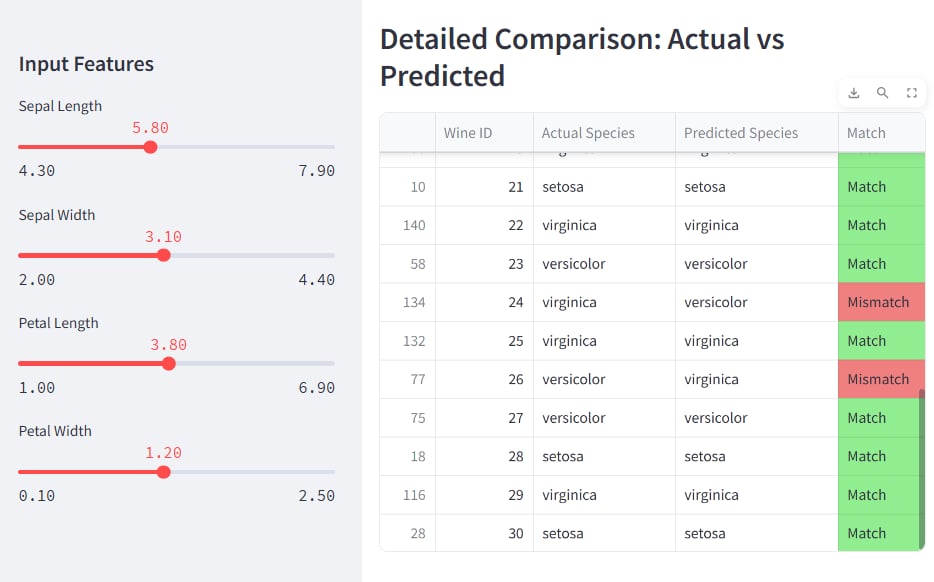
The application integrates the visual charts analyzed earlier to provide an overview of the Iris dataset and the performance of the ID3 model. These charts are displayed directly within the interface, allowing users to easily reference and gain a deeper understanding of the data distribution, feature importance, decision tree structure, and the model's classification errors. This functionality supports users not only in making predictions but also in conducting deeper analyses of the data and the model, enhancing the educational and visual appeal of the application.

#### **3.4.4. Display Model Performance Results**



The application provides performance metrics for the ID3 model, including accuracy on the validation set (0.96) and test set (0.93), along with a detailed classification report (precision, recall, F1-score for each species). For example, the classification report on the test set shows setosa with an F1-score of 1.00, while versicolor and virginica achieve 0.90. This information is presented clearly, enabling users to assess the model's reliability before making predictions. This feature ensures users have a comprehensive understanding of ID3's performance, thereby increasing confidence in the application's prediction results.

#### **3.4.5. Comparison Between Actual and Predicted Results**

****

The application provides a detailed comparison table between actual and predicted labels on the test set, with columns "Actual Species," "Predicted Species," and "Match." For example, sample 24 has an actual label of virginica but was incorrectly predicted as versicolor (Mismatch), while other samples like 21 (setosa) and 22 (virginica) were predicted correctly (Match). This table helps users identify correct and incorrect classifications, thus providing insight into the strengths and limitations of the ID3 model, particularly in distinguishing between versicolor and virginica, as highlighted in the confusion matrix.

### **Chapter 4. Conclusion and Recommendations**

### **4.1. Conclusion**

In this project, we successfully researched and implemented a software program using the ID3 decision tree technique to classify flower species from the Iris dataset. Through careful analysis, model design, training, and evaluation, we obtained the following conclusions:

* The ID3 decision tree model demonstrated **high accuracy** in classifying Iris flower species, achieving **96% accuracy on the validation set** and **93% on the test set**.
* Analysis of feature importance revealed that **petal length** and **petal width** are the most significant attributes contributing to the classification, while **sepal length** and **sepal width** contributed relatively little.
* The decision tree structure showed that the model made clear and logical splits, primarily based on petal-related features, which aligns with the exploratory data analysis results.
* The confusion matrix analysis confirmed that the model **perfectly classified Setosa samples**, while a few misclassifications occurred between Versicolor and Virginica due to slight feature overlap.
* The developed **Streamlit web application** allowed interactive feature input, prediction display, probability visualization, and comparison between actual and predicted results, making the model accessible and interpretable to users.

Overall, the ID3 model proved to be an effective, interpretable, and reliable method for classifying the Iris flower species based on morphological features.

### **4.2.Achievements**

* Gained a deeper understanding of **machine learning concepts**, especially supervised learning and decision trees.
* Successfully applied the **ID3 algorithm** to solve a practical classification problem.
* Designed and implemented a **user-friendly application** to interactively use the trained model.
* Practiced critical steps in the machine learning pipeline, including **data preprocessing, feature analysis, model evaluation, and result interpretation**.

### **4.3. Limitations**

* Despite high overall accuracy, the model showed slight confusion between the Versicolor and Virginica classes, suggesting that **ID3 alone may struggle** when classes have overlapping features.
* The model’s performance could be further affected if the dataset included noisy or more complex samples.
* The current model is based on a relatively simple and balanced dataset (Iris), which may not fully reflect challenges in real-world, imbalanced, or large-scale datasets.

### **4.4. Recomedations**

For future improvement and further research:

* **Experiment with pruning techniques** to avoid overfitting and enhance model generalization.
* **Combine ID3 with ensemble methods** such as Random Forests to improve robustness and accuracy, especially for complex datasets.
* **Test the model on larger, more challenging datasets** to assess scalability and real-world performance.
* **Incorporate cross-validation more systematically** during model training to obtain more stable and reliable performance estimates.
* **Further develop the application** by adding functionalities like feature selection options, model comparison, and explanation of decision paths for each

# TÀI LIỆU THAM KHẢO

1. https://ndquy.github.io/posts/Phan-lop-danh-gia-he-thong-phan-lop
2. [https://www.kaggle.com/code/sohaibhussain/predict-wine-quality-using-knn](http://www.kaggle.com/code/sohaibhussain/predict-wine-quality-using-knn)
3. https://scikit-learn.org/
4. https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning- tips-and-tricks
5. How to Evaluate the Quality of a Wine - From The Vine (wtso.com)
6. KNN Algorithm: When? Why? How?. KNN: K Nearest Neighbour is one of the…

| by Aditya Kumar | Towards Data Science

1. Machine Learning cơ bản (machinelearningcoban.com)
2. Giới thiệu về thuận toán K láng giềng (K nearest neighbor) trong machine learning – FLINTERS Developer's Blog
3. The 4 Factors and 4 Indicators of Wine Quality | JJ Buckley Fine Wines
4. UCI Machine Learning Repository: Wine Quality Data Set