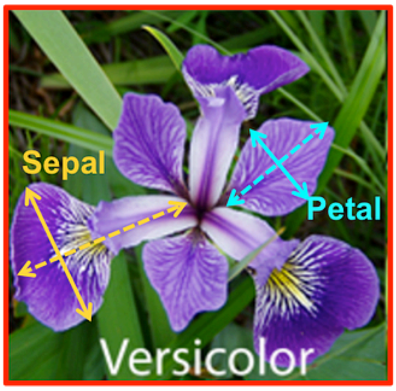
Iris Dataset

The Iris flower data set or Fisher's Iris data set is a multivariate data set used and made famous by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.

The data set consists of 50 samples from each of three species of Iris (Iris Setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.



The dataset is balanced i.e. equal records are present for all three species.We have four numerical columns while just one categorical column which in turn is our target column.A strong correlation is present between petal width and petal length.The setosa species is the most easily distinguishable because of its small feature size.The Versicolor and Virginica species are usually mixed and are sometimes hard to separate, while usually Versicolor has average feature sizes and virginica has larger feature sizes.

Why logistics regression

So while plotting the data using scatter plot where x = petal length and y = sepal width, We can see that most of the data's are Linearly separable .That is , From the scattered plot we can say that there is a simple classification between the data's. Also from the scatter plot we can say that there will be 100 % accuracy on the species setosa ,and there won't be 100% accuracy on versicolor and Virginia. We know that the iris dataset is a small dataset of 150 flowers . Logistic regression was selected because the dataset is small and also the data can be separated linearly .

Methodologies

Exploratory Data Analysis (EDA) and Logistic Regression are common techniques used in data analysis and predictive modelling.

steps

1. **Load the Datase**t : Import necessary libraries (e.g., pandas, numpy, matplotlib, seaborn) and load the dataset
2. **Understand the dataset** : Check the basic information about the dataset (e.g., shape, data types) and Explore the first few rows of the dataset.
3. **Statistical Summary :** Generate summary statistics to understand the central tendency, dispersion, and shape of the distribution of a dataset.
4. **Data Visualization** : Use visualisations like histograms, pair plots, and box plots to understand the distribution of features.
5. **Correlation Analysis :** Explore the correlation between features.

Logistic Regression

1. **Data Preparation**: Split the dataset into features (X) and target variable (y). And Encode categorical target variables if needed.
2. **Train and Test Split :** Split the dataset into training and testing sets.
3. **Logistic Regression Model :** Import and fit a logistic regression model.
4. **Model Evaluation :** Evaluate the model using accuracy, confusion matrix, precision, recall, and F1-score.
5. **Confusion Matrices:** Heap Maps are used for Performance Evaluation.

Google Drive Link - 🗀 NEXUS INTERNSHIP DATASCIENCE