



Research Article

Classification of The Iris Dataset

Abhigyan Gandhi, 2018A7PS0168U, f20180168@dubai.bits-pilani.ac.in

Submitted to: Ma'am A Razia Sulthana, Instructor-In-Charge,
Bits F464 Machine Learning

Abstract - This report is for the project assignment for the machine learning course. This is the first time I am studying machine learning, so I chose the most basic problem for this project as it is very useful for beginners to learn more about machine learning and get started to make even complex projects. My topic for the project is the classification of the Iris dataset, which is a very popular dataset and is widely considered the hello world of machine learning. The main part of this report is the literature review done by me of the 15 papers based on the topic of classification algorithms for pattern recognition in machine learning and the implementation of different classification algorithms for the classification of the Iris dataset.

I have listed all the papers which I reviewed in a list and also provided the objectives and methodology for my project, along with the results I found and the most suitable algorithm with the highest accuracy.

INTRODUCTION

This project is for beginners who want to understand more about machine learning and classification algorithms used for pattern recognition. I have chosen the Iris flower dataset for my project which is a multivariate dataset created by Ronald Fisher in 1936. The dataset has been widely used in statistical classification learning as it implements machine learning techniques such as the support vector machine.

The dataset consists of 50 different samples from the three species of the Iris flower: Iris setosa, Iris virginica, and Iris versicolor. Four features were measured from each sample, the length and width of the sepals and the petals. Fisher developed a linear discriminant model based on this dataset.

The set consists of only two clusters, so it is commonly used in cluster analysis. But this dataset gives a good understanding of the difference between supervised and unsupervised techniques in data science. All the attributes of the dataset are numeric so it helps in understanding how to handle and load data. This is a good project as it is very well understood.

Classification algorithms are a part of machine learning which make use of a training set and use it to get boundary conditions for approximating the target dataset. These algorithms consist of classifiers and classification models. The classifier maps the train set into different classes. The classification model uses this map created by the classifier to determine and distinguish the target dataset into different classes. Classification can be

binary or multi-class. This project is based on multi-class classification. Our objective will be to

determine which algorithm has the highest accuracy out of the five which we will implement.



Figure 1: Different types of the Iris flower

LITERATURE REVIEW

I reviewed around 15 research papers based on classification algorithms used for pattern recognition since 2016. The advancements made in the recent decades have been very massive and thousands of researches have been conducted and applications have been developed. Machine learning has helped the world in every aspect. The papers I reviewed were very useful in understanding the problems faced by the authors and how the newer papers were able to overcome issues in the previous ones.

^[1]In 'Random Forest Algorithm for Land Cover Classification' by Arun D. Kulkarni, he compares the different classification approximations of random forest algorithm with other widely used algorithms like maximum likelihood and support vector machines.

^[2]In 'Intelligent Pattern Recognition of a SLM Machine Process and Sensor Data' by Eckart Uhlmann et al, the authors come up with a different pattern detection technique using a different process and sensors data from a SLM machine. The findings are calculated using a smart tool used for algorithm configuration.

^[3]In 'Performance analysis of machine learning and pattern recognition algorithms for Malware classification' by Barath Narayanan et al, the authors tried to visually see the viruses in an image as they capture the changes. They

implemented principal component analysis for feature extraction.

^[4]In 'Pattern Recognition Approaches for Breast Cancer DCE-MRI Classification' by Roberta Fusco et al, the authors used ANN, SVM, LDA, TC, BC and reviewed several pattern analysis approaches.

^[5]In 'Survey of Machine Learning Algorithms for Disease Diagnostic' by Meherwar Fatima et al, the authors reviewed different machine learning algorithms for the detection of different diseases heart diseases, liver diseases and diabetes diseases.

^[6]In 'A Review Of Point Clouds Segmentation And Classification Algorithms' by E. Grilli et al, the authors reviewed different algorithms for classifying 3D point clouds. Strong and weak points both have been shown and discussed.

^[7]In 'Machine learning in pain research' by Alfred Ultsch et al, the authors discuss ways of machine learning to understand pain. They extract knowledge from complicated pain-related data and find useful data from it.

^[8]In 'Curvature-based pattern recognition for cultivar classification of Anthurium flowers' by Alireza Soleimani Pour et al, the authors did the classification of Anthurium flowers by using image data, b-spline curves and machine learning algorithms. Then they applied a classification model on the data for the distinction.

^[9]In '*Recognition of CAPTCHA Characters by Supervised Machine Learning Algorithms*' by Jan Klecka et al, the authors compared various classification methods like the k-nearest neighbour and support vector machines. They found that the main difference is between the learning times.

^[10]In '*Towards automated statistical partial discharge source classification using pattern recognition techniques*' by Hamed Janani et al, the authors try to classify particle discharge by finding the relationship between the source and PRDP by using a interpretation involving the probability approximation of a classification model.

^[11]In '*Automatic Speaker Recognition System based on Machine Learning Algorithms*' by Tumisho Billson Mokgonyane et al, the authors use the WEKA data mining tool to figure out the best classification algorithm for the automatic detection of the speaker.

^[12]In '*A New Automatic Method for Control Chart Patterns Recognition Based on ConvNet and*

Harris Hawks Meta Heuristic Optimization Algorithm' by Noorbakhsh Amiri Golilarz et al, the authors have proposed a new algorithm called ConvNet based on deep learning techniques and optimization algorithms for nine control chart patterns.

^[13]In '*Functional localization in the brain of a cynomolgus monkey based on spike pattern recognition with machine learning*' by Mixia Wang, the authors describe a functional localization method based on spike patterns recognition through machine learning algorithms.

^[14]In '*Woven Fabric Pattern Recognition and Classification Based on Deep Convolutional Neural Networks*' by Babar Khan et al, the authors propose a model which uses residual network, where the texture is extracted and classified automatically.

^[15]In '*Human activity recognition using machine learning methods in a smart healthcare environment*' by Tayeb Brahimi et al, the authors present a smart healthcare system for delivering pervading human activity recognition by using machine learning techniques.

TABLE OF PAPERS STUDIED

References	Objectives	Problem statement	Methodology	Dataset	Algorithm	Advantage	Disadvantage	Performance-measure value
1	To use random forest algo for land classification	Random Forest Algorithm for Land Cover Classification	Thematic mapper imagery	Land cover dataset	Random forest algorithm	Uses bootstrap aggregating	Required much computational power	7
2	To approach pattern recognition using a different process	Intelligent Pattern Recognition of a SLM Machine Process and Sensor Data	Information was extracted from sensor data to build database for SLM machine	Sensor dataset	K-mean algorithm	Simple to implement	Dependant on initial values	6
3	To classify malware	Performance analysis of machine learning and pattern recognition algorithms for Malware classification	Used PCA for feature extraction, then implement classification algorithms	Malware dataset	KNN, SVM	No training period	Does not work well with large dimensions	7
4	To classify breast lesions using DCE-MRI	Pattern Recognition Approaches for Breast Cancer DCE-MRI Classification	Clinical characterisation of patients	Breast cancer MRI scans	ANN	Has fault tolerance	Unexplained behaviour of the network	6
5	To find the type of disease	Survey of Machine Learning Algorithms for Disease Diagnostic	Applied various techniques to determine the best one	High-dimensional biomedical data	WEKA tool	Feature selection and big range of data preparation	Do not implement newer methods	8
6	To segment 3D point clouds	A Review Of Point Clouds Segmentation And	Segmentation based on different methods like edge-	3D dataset	Canupo segmentation algorithm	Allows to create own classes and use	Cost heavy	8

		Classification Algorithms	based, model fitting, etc.			pre-existing classifiers		
7	To understand the complexity of pain	Machine learning in pain research	Computational science methods using complex clinical data to understand pain	Pain-related data	Random forest algo, ANN and KNN	Uses bootstrap aggregating	Does not work well with large data sets	7
8	Classification of flowers based on four different features	Curvature-based pattern recognition for cultivar classification of Anthurium flowers	Detect the boundary of flowers and reconstruct using b-spline curve	Images of anthurium flowers	SVM	Works well with unstructured data	Long training time	9
9	To compare several classification algorithms for character segmentation	Recognition of CAPTCHA Characters by Supervised Machine Learning Algorithms	Implement different algorithms and figure out the best one	Optical CAPTCHA character dataset	Neural networks	Has ability to work with insufficient data	Hardware dependence	8
10	Online monitoring and recognition of PD patterns	Towards automated statistical partial discharge source classification using pattern recognition techniques	Probabilistic interpretation of classification results to find relationships	Particle discharge dataset	Decision trees	Less effort during data preprocessing	Complex calculation	7
11	To detect the speaker by listening to their voice	Automatic Speaker Recognition System based on Machine Learning Algorithms	To apply classification algorithms and test them using WEKA tool	Voice recordings of different language	WEKA tool	Feature selection and big range of data preparation	Do not implement newer methods	8
12	To recognize CCPs using ConvNet	A New Automatic Method for Control Chart Patterns Recognition Based on ConvNet and Harris Hawks	Used a new method based on deep learning techniques	CCPs	ConvNet	High statistical efficiency	Classification of images with different positions	7

		Meta Heuristic Optimization Algorithm						
13	To treat Parkinson's disease in monkeys	Functional localization in the brain of a cynomolgus monkey based on spike pattern recognition with machine learning	To implement a functional localization method by using spike pattern recognition	Brain patterns of monkeys	K-mean algorithm	Simple to implement	Dependant on initial values	7
14	To detect different fabric patterns	Woven Fabric Pattern Recognition and Classification Based on Deep Convolutional Neural Networks	Used ResNet to extract patterns and classified in end-to-end fashion	Fabric texture patterns	ResNet	Powerful representational ability	Improvement requires increasing number of layers	7
15	To develop a smart healthcare system	Human activity recognition using machine learning methods in a smart healthcare environment	Used machine learning methods to provide pervasive HAR autonomously	Healthcare monitored data	Nearest neighbour algorithm	No training period	Does not work well with large dimensions	6

OBJECTIVES

From the research papers reviewed, it was found one of the papers was very similar to the one I have chosen for this topic which was the paper based on cultivar classification of the Anthurium flower.

Our main objective during this project will be to accurately classify the types of the Iris flowers into the right category out of the three. We have

two different clusters, the first one has one of the categories and the second has the rest two. Our data is also numeric, so it will be fairly easy to separate it and train the classification model.

My contributions to this project will be to identify the best algorithm for the classification of the dataset, then to implement the classification model and to analyse the results and try to reduce the error.

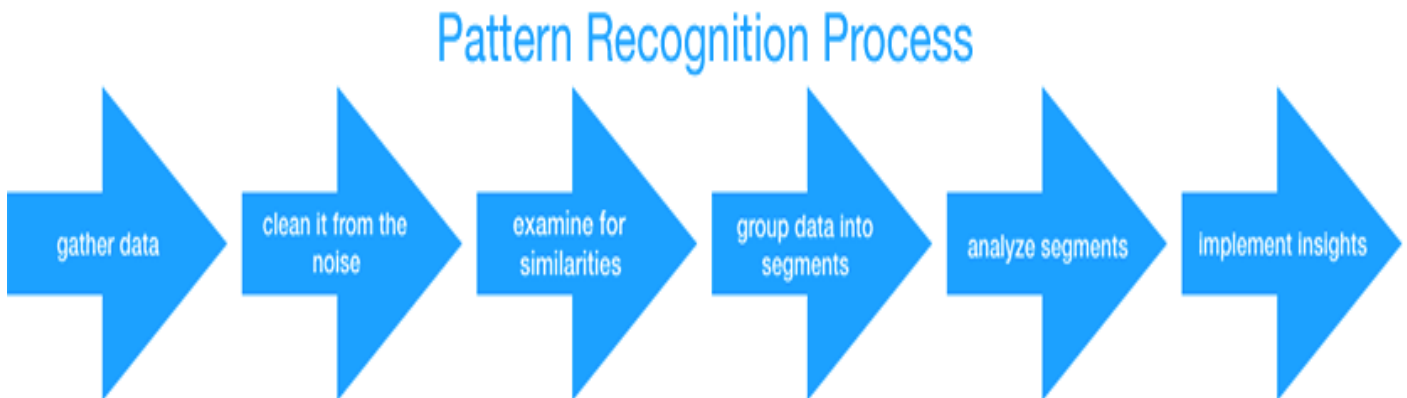


Figure 2: Pattern recognition process

METHODOLOGY

The methodology for this project was a very basic approach as I am a beginner to the field of data science and machine learning. I started out with the SciPy library of Python for this project because this library contains a lot of scientific tools and packages for calculations and computing. The important libraries which I used were the following: SciPy, numpy, matplotlib, pandas and sklearn. Matplotlib was used for plotting of graphs for the visualisation of the dataset and comparing the results. Pandas was used to import the csv dataset file and for generating a scatter matrix. The main library was the sklearn library as it contains all the functions required for the implementation of the different classification algorithms. It uses numpy for performing mathematical operations on large

arrays and matrices and produces accurate and viable results.

The first step after importing all the libraries was to load the dataset. The Iris dataset is easily available on the internet, so I downloaded it as a csv file and imported it as a matrix as the data is linear and numerical. The `read_csv()` function of the pandas library was used for this. Next step was to summarize the data so that we could understand it better. So, I looked at different aspects of the data. It is always good to look at the data from the top, so I displayed the first 20 rows of the dataset and then summarized the data according to the attributes and displayed all necessary properties including the count, mean, minimum and maximum values, standard deviation and percentiles 25%, 50% and 75%. Then I also grouped the data class-wise taking the three types of the Iris flower as individual classes.

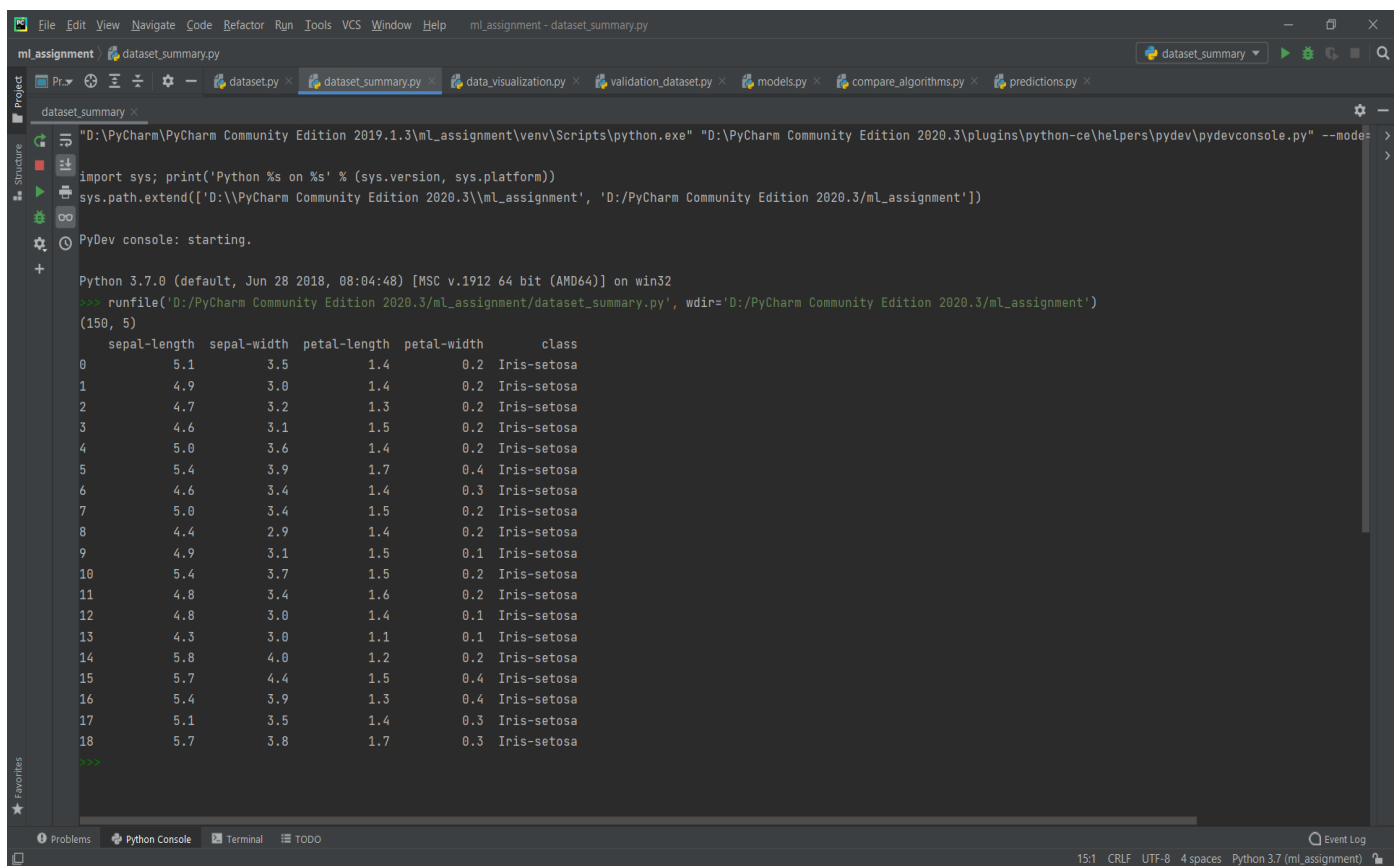


Figure 3: Shape of the dataset and the first 20 rows

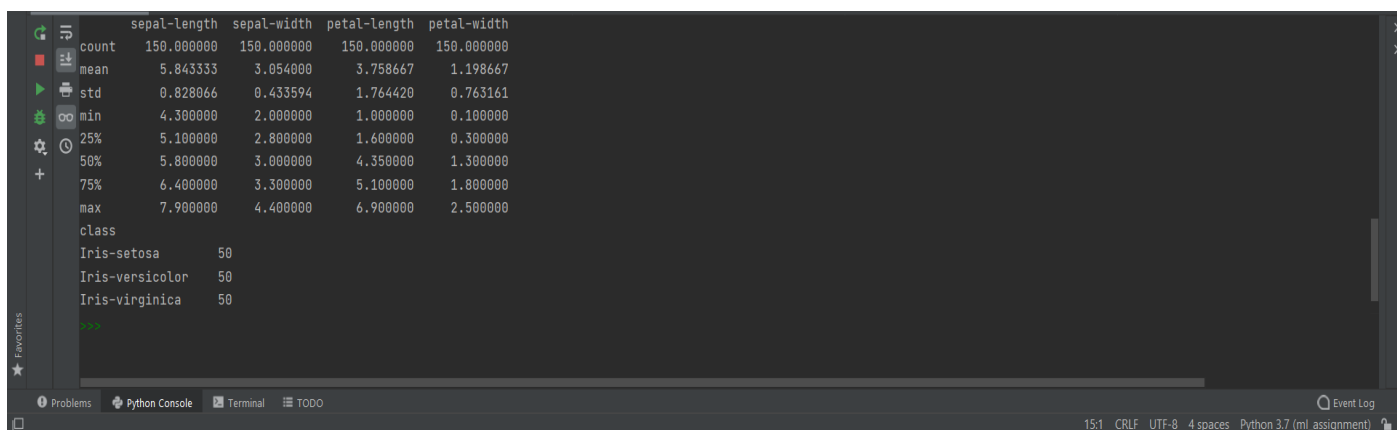


Figure 4: Detailed description of the dataset

After summarizing the data, the next step was to visualize the data by plotting some graphs. For this, I decided to represent the data in two ways, univariate plots and multivariate plots. Univariate plots are a good way of understanding the individual attributes whereas multivariate plots are better for understanding the relation between the attributes. For the univariate plots, I displayed two

types of plots, box and whisker plot and the histogram to understand the distribution of the values. Both of these plots are available in the

matplotlib library and we can see the output using the pyplot function of the same.

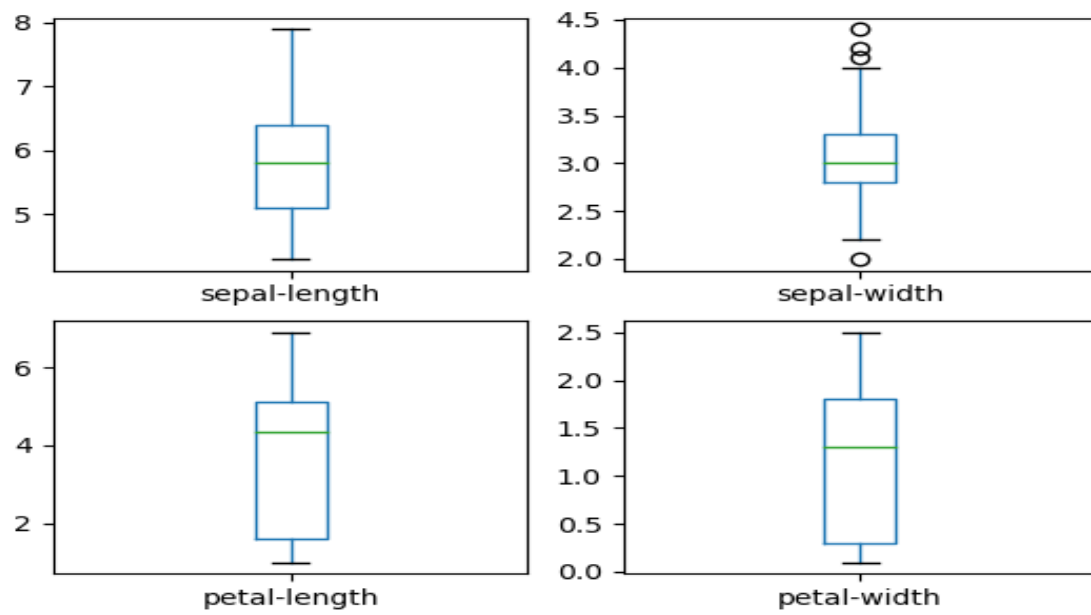


Figure 5: Box and whisker plots of all the attributes

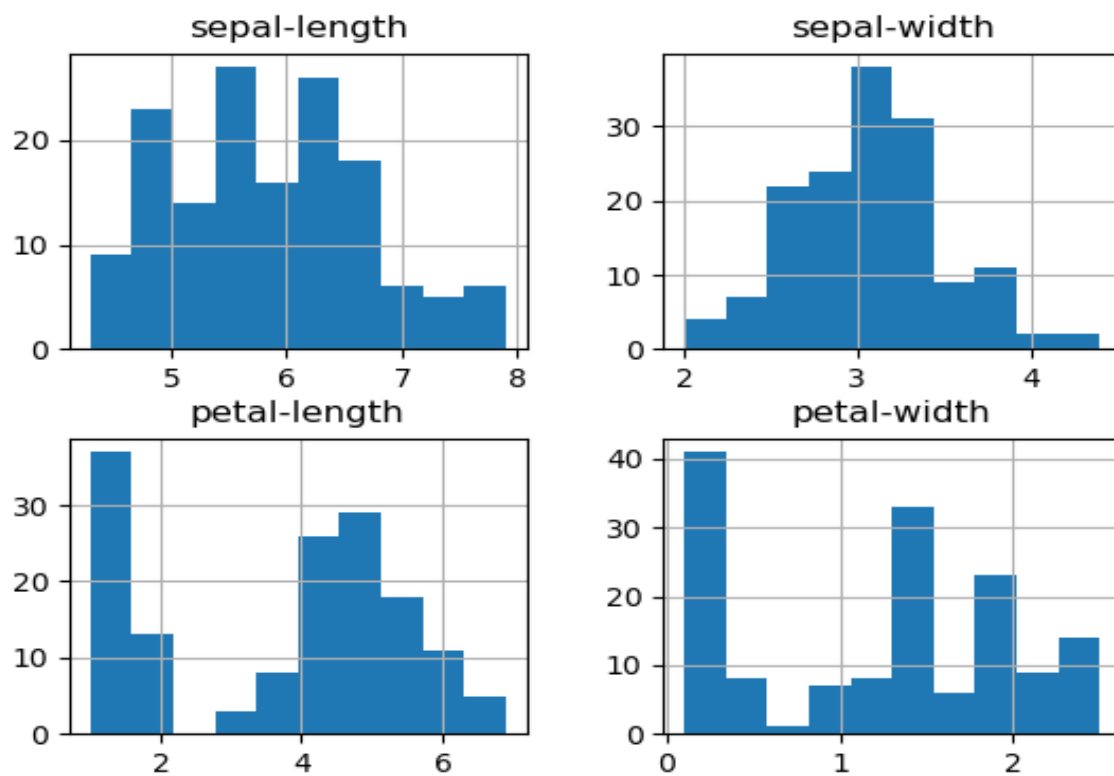


Figure 6: Histograms of all the attributes

For the multivariate plots, I plotted a scatter matrix using the `scatter_matrix` function of the `pandas` library. It was found that a lot of attributes were getting grouped together in a diagonal

manner which means that there was high correlation between them and their relationship can be predicted easily.

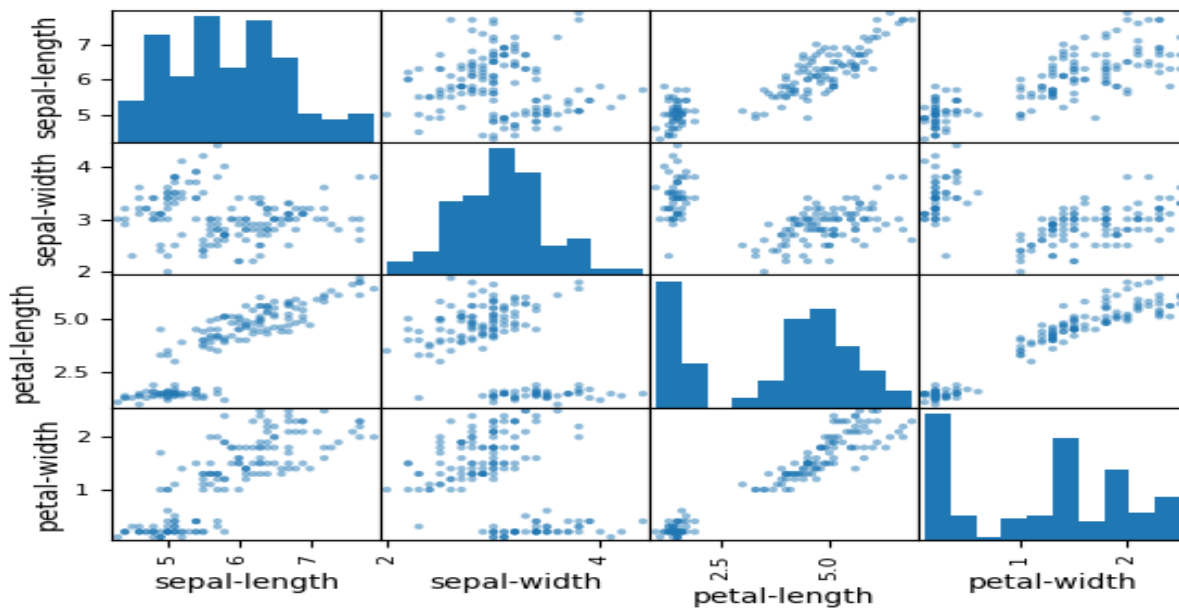


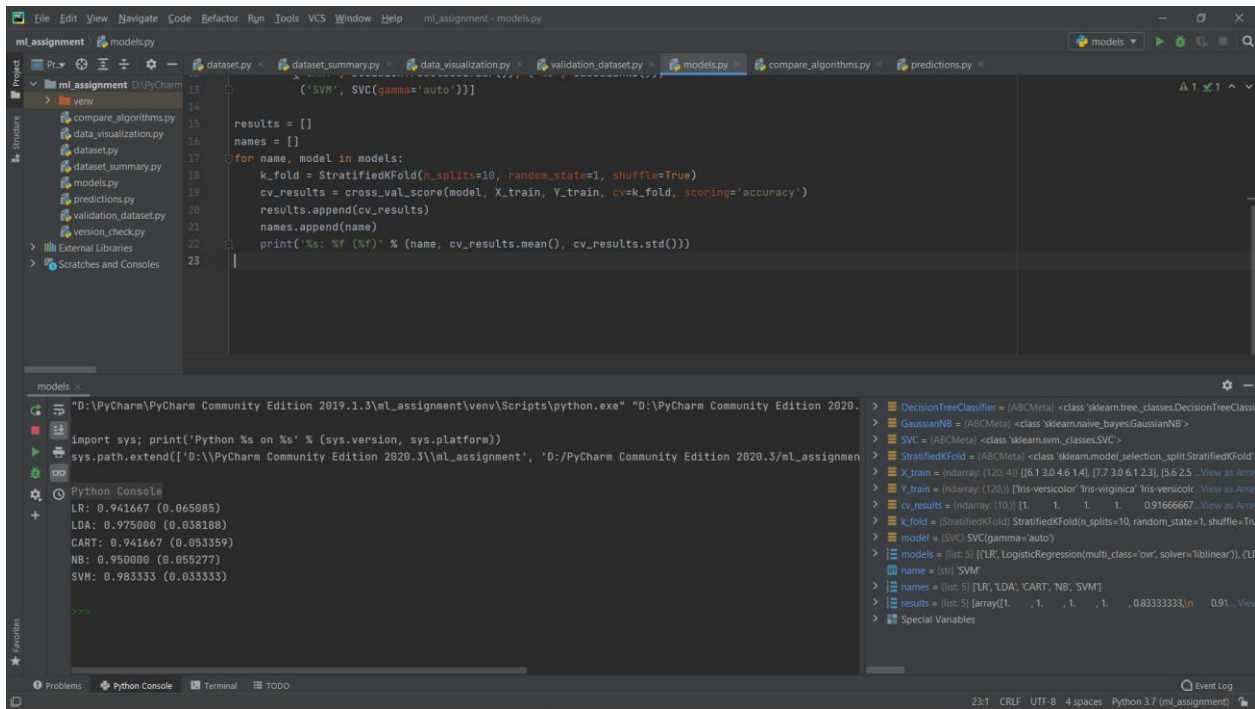
Figure 7: Scatter matrix of the attributes

The next step was to split the dataset into a train set and test set for the algorithms to work on. I used five different models to train the data and find out which was the best. So, 80% of the data was used for training and 20% for testing. The data was split using slicing where arrays are used to first differentiate between input data and output, then by using the `train_test_split` function of the `sklearn` library by specifying the values of the arguments, the data was split into x and y variables for both train set and test set.

Then for the testing and training, I used five different classification algorithms including logistic regression, linear discriminant analysis, regression trees, Gaussian Naïve Bayes and support vector machines. All these models are available in the `sklearn` library, so I imported them and trained them, and found out the accuracy of all the models using cross validation in which the dataset is split into 10 parts and then 9 parts are used for training and 1 part for testing. This completed the building and training of our models.

RESULTS AND PREDICTIONS

After training the data and iterating the models through the 10-fold cross validation, we found the accuracy scores and it was observed that the support vector machine model had the highest accuracy score with 98.3% accuracy.



```
models.py
14
15
16
17
18
19
20
21
22
23

('SVM', SVC(gamma='auto'))

results = []
names = []

for name, model in models:
    k_fold = StratifiedKFold(n_splits=10, random_state=1, shuffle=True)
    cv_results = cross_val_score(model, X_train, Y_train, cv=k_fold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))
```

```
Python Console
LR: 0.941667 (0.065085)
LDA: 0.975000 (0.038188)
CART: 0.941667 (0.053359)
NB: 0.950000 (0.055277)
SVM: 0.983333 (0.033333)
```

```
DecisionTreeClassifier = (ABCMeta) <class 'sklearn.tree._classes.DecisionTreeClassifier'>
GaussianNB = (ABCMeta) <class 'sklearn.naive_bayes.GaussianNB'>
SVC = (ABCMeta) <class 'sklearn.svm._classes.SVC'>
StratifiedKFold = (ABCMeta) <class 'sklearn.model_selection._split.StratifiedKFold'>
X_train = (ndarray) [120, 4] [[6.1 3.0 4.6 1.4], [7.3 3.0 6.1 2.3], [5.6 2.5 ... View as Array]
Y_train = (ndarray) [120, 1] [1. 1. 1. 1. 0.91666667 ... View as Array]
cv_results = (ndarray) [10, 1] [1. 1. 1. 1. 0.91666667 ... View as Array]
k_fold = (StratifiedKFold) StratifiedKFold(n_splits=10, random_state=1, shuffle=True)
model = (SVC) SVC(gamma='auto')
models = (list) [5] [(LR, LogisticRegression(multi_class='ovr', solver='liblinear')), (LDA, LinearDiscriminantAnalysis()), (CART, DecisionTreeClassifier()), (NB, NaiveBayes()), (SVM, SVC(gamma='auto'))]
name = (str) 'SVM'
names = (list) [5] ['LR', 'LDA', 'CART', 'NB', 'SVM']
results = (list) [5] [array([1., 1., 1., 1., 0.98333333]), array([0.94166667, 0.975, 0.94166667, 0.95, 0.98333333]), array([0.065085, 0.038188, 0.053359, 0.055277, 0.033333]), array([0.94166667, 0.975, 0.94166667, 0.95, 0.98333333]), array([0.065085, 0.038188, 0.053359, 0.055277, 0.033333])]
```

Figure 8: Accuracy scores for the algorithms

Then for another comparison between the algorithms, I plotted a box and whisker plot to see the distribution.

All the algorithms performed very well with high accuracy percentages. We have now found the algorithm which we will use to make the predictions which is the SVM model.

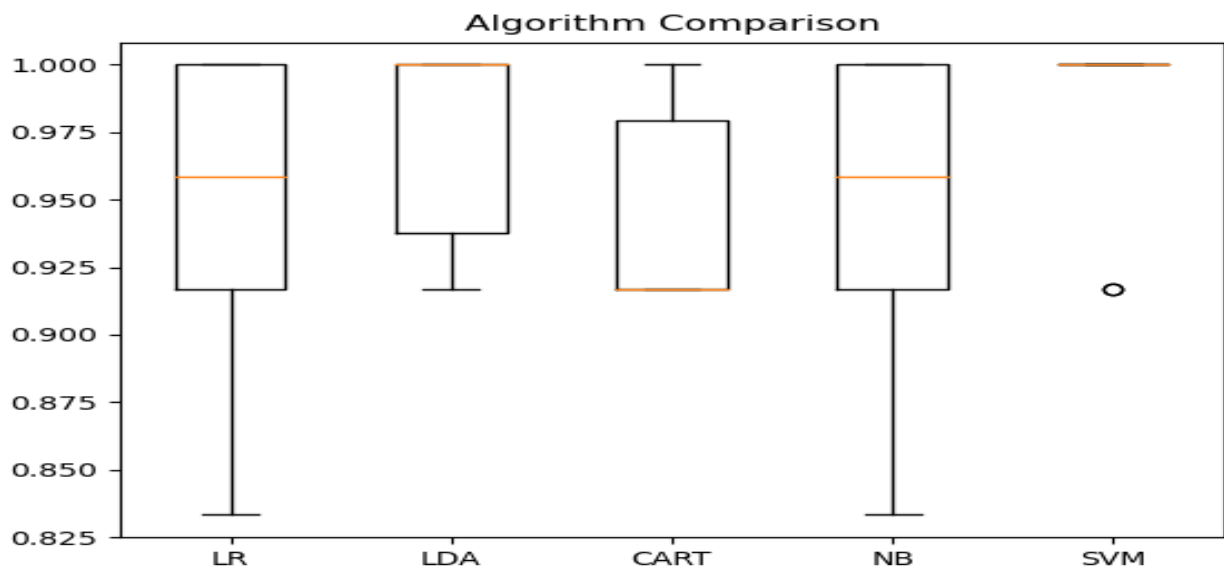


Figure 9: Box and whisker plot for the algorithms

Now, we will use the test dataset which we had created previously on this model. I used functions from the sklearn library to find out the accuracy score, the confusion matrix and the classification report which included the precision, recall, f1-score and support. The accuracy for the test dataset was found to be 96%.

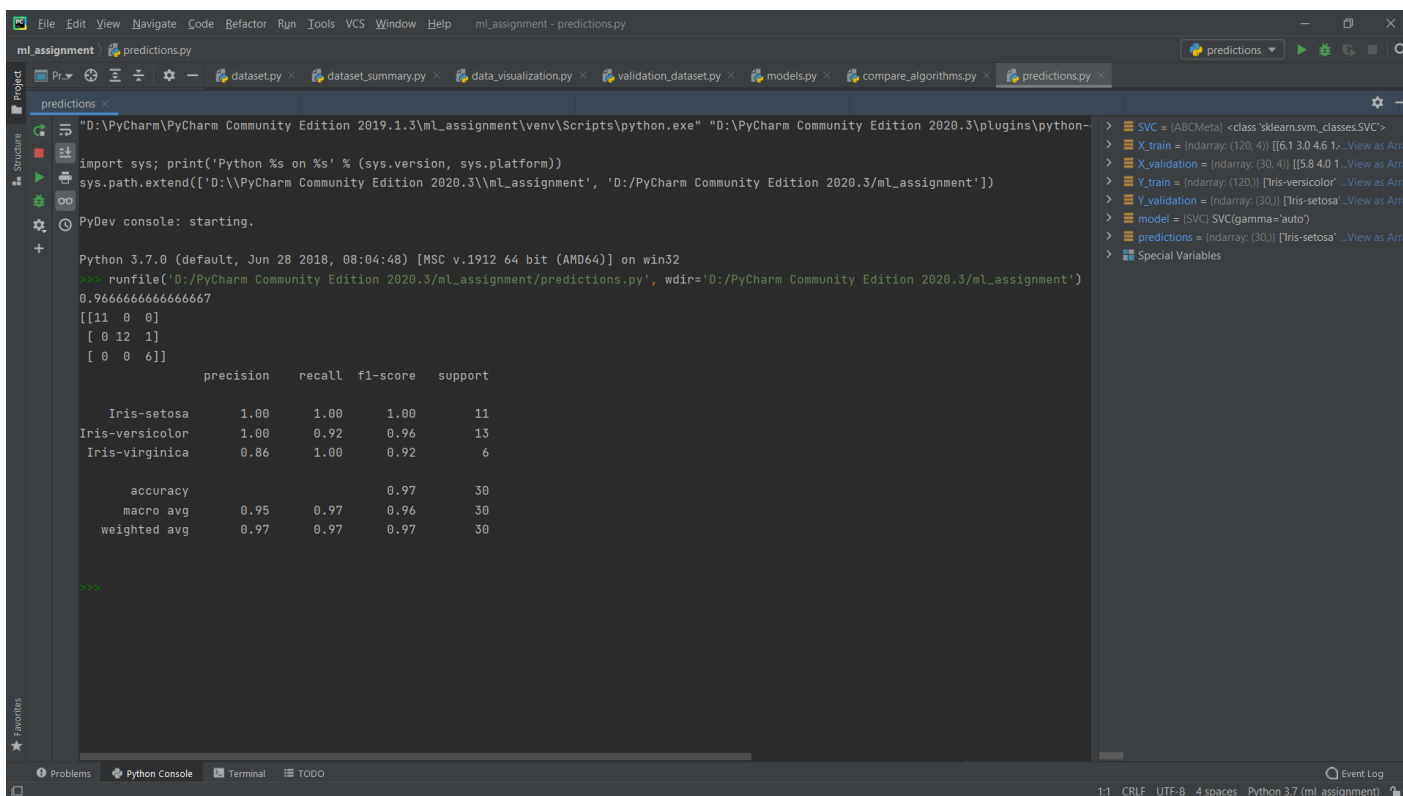


Figure 10: The test results for the SVM model

CONCLUSION

In this paper, five different classification algorithms were implemented and tested on the Iris dataset to find out which algorithm performs the best and by using the sklearn library the support vector machine model was found to be the model with the highest accuracy. This project was a very good process for me to understand how to work in Python for making machine learning projects and to gain more knowledge about classification algorithms

REFERENCES

1. Arun D. Kulkarni, 'Random Forest Algorithm for Land Cover Classification', University of Texas at Tyler, 2016.
https://scholarworks.uttyler.edu/compsci_fac/1/
2. Eckart Uhlmann et al, 'Intelligent Pattern Recognition of a SLM Machine Process and Sensor Data', Elsevier Vol. 62, 2016.
<https://www.sciencedirect.com/science/article/pii/S2212827116306783>
3. Barath Narayanan et al, 'Performance analysis of machine learning and pattern recognition algorithms for Malware classification', IEEE NAECON, 2016.
<https://ieeexplore.ieee.org/abstract/document/7856826>
4. Roberta Fusco et al, 'Pattern Recognition Approaches for Breast Cancer DCE-MRI Classification', Journal of Medical and Biological Engineering, 2016.
<https://link.springer.com/article/10.1007/s40846-016-0163-7>
5. Meherwar Fatima et al, 'Survey of Machine Learning Algorithms for Disease Diagnostic', Journal of Intelligent Learning Systems and Applications Vol.09 No.01, 2017.
https://www.scirp.org/html/1-9601348_73781.htm
6. E. Grilli et al, 'A Review Of Point Clouds Segmentation And Classification Algorithms', 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy, 2017.
https://www.researchgate.net/profile/Fabio_Menna/publication/313959376_A_REVIEW_OF_POINT_CLOUDS_SEGMENTATION_AND_CLASSIFICATION_ALGORITHMS/links/58bd252092851c471d564138/A-REVIEW-OF-POINT-CLOUDS-SEGMENTATION-AND-CLASSIFICATION-ALGORITHMS.pdf
7. Alfred Ultsch et al, 'Machine learning in pain research', PMC journal list, 2017.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5895117/>
8. Alireza Soleimani Pour et al, 'Curvature-based pattern recognition for cultivar classification of Anthurium flowers', Elsevier Vol. 139, 2018.
<https://www.sciencedirect.com/science/article/abs/pii/S0925521417311067>
9. Jan Klecka et al, 'Recognition of CAPTCHA Characters by Supervised Machine Learning Algorithms', Elsevier Vol. 51, 2018.
<https://www.sciencedirect.com/science/article/pii/S2405896318309017>
10. Hamed Janani et al, 'Towards automated statistical partial discharge source classification using pattern recognition techniques', IET journals, 2018.
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8479403>
11. Tumisho Billson Mokgonyane et al, 'Automatic Speaker Recognition System based on Machine Learning Algorithms', PRASA conference, 2019.
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8704837>

12. Noorbakhsh Amiri Golilarz et al, 'A New Automatic Method for Control Chart Patterns Recognition Based on ConvNet and Harris Hawks Meta Heuristic Optimization Algorithm', IEEE Access, 2019.
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8859191>
13. Mixia Wang, 'Functional localization in the brain of a cynomolgus monkey based on spike pattern recognition with machine learning', Springer Link, 2019.
<https://link.springer.com/article/10.1007%2Fs12652-019-01576-9>
14. Babar Khan et al, 'Woven Fabric Pattern Recognition and Classification Based on Deep Convolutional Neural Networks', MDPI journals Vol. 9 issue 6, 2020.
<https://www.mdpi.com/2079-9292/9/6/1048>
15. Tayeb Brahimi et al, 'Human activity recognition using machine learning methods in a smart healthcare environment', ScienceDirect 2020.
<https://www.sciencedirect.com/science/article/pii/B9780128190432000058>