Disease Prediction Using Machine Learning

applied machine learning project
Dr. Enas Abo Fateh

TABLE OF CONTENT

INTRODUCTION

People face various diseases due to the environmental condition and their living habits. And Medical science has large amount of data growth per year. Due to increase amount of data growth in medical and healthcare field . So we proposed Disease Prediction Using Machine Learning

OBJECTIVE

The purpose of constructing this project called “Disease Prediction Using Machine Learning” is to predict the accurate disease of the patient using all their general information’s and also the symptoms

APPROACH

- Data set

The dataset we will use in our project is Disease Prediction consists of 2 CSV files. One of them is for training and the other is for testing your model. Each CSV file has 133 columns. 132 of these columns are symptoms that a person experiences and the last column is the prognosis.

- Machine Learning Model:

We have tried multiple prediction models but using the Random Forest classifier model has shown the best results.

- Data Collection

From Kaggle:
https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning

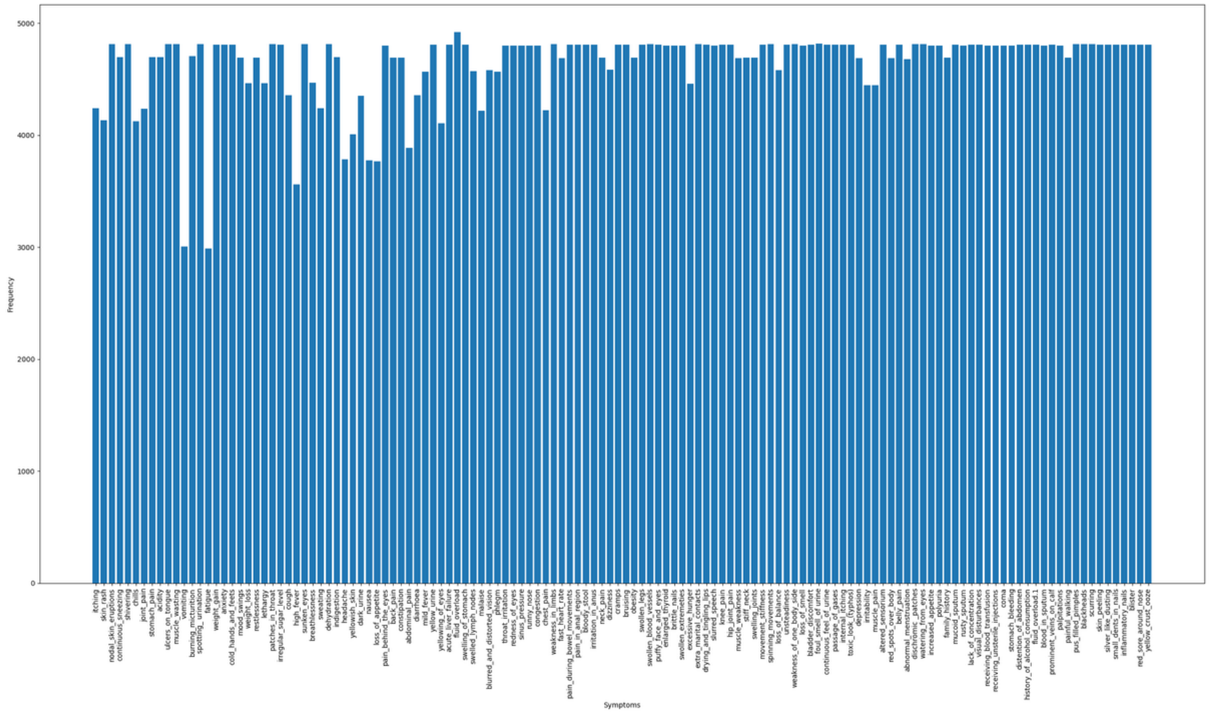
Random Forest classifier

The random forest classifier is a supervised learning algorithm that you can use for regression and classification problems. It is among the most popular machine learning algorithms due to its high flexibility and ease of implementation.

ANALYSIS AND RESULT

Question/hypothesis

This picture shows the frequency of the symptom, we see that many of the symptoms occurred at about the same times, and the conclusion from this analysis inspired us to write these questions.



Predict whether the symptoms indicate a specific disease or not?

Is the increase in the proportion of symptoms a good indicator for the diagnosis of the disease or not?

Import tne important libraries.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.ensemble import RandomForestClassifier
import matplotlib.pyplot as plt
```

Import the dataset.

```
train = pd.read_csv('Dataset/Training.csv')
test = pd.read_csv('Dataset/Testing.csv')
```

Dataset

```
In [4]: train.head()
```

Out[4]:

es	blackheads	scurring	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	blister	red_sore_around_nose	yellow_crust_ooze	prognosis
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection

Preprocessing

```
# check for null values
train.isnull().any()
```

```
# check if balanced
train['prognosis'].value_counts()
```

Splitting the data.

```
A = train[["prognosis"]] # diseases
B = train.drop(["prognosis"],axis=1) # symptoms
C = test.drop(["prognosis"],axis=1) # symptoms - testing
x_train, x_test, y_train, y_test = train_test_split(B,A,test_size=0.2)
```

Training the model

```
# Traning random forest model
mod = RandomForestClassifier(n_estimators = 100,n_jobs = 5, criterion= 'entropy',random_state = 42)
mod = mod.fit(x_train,y_train.values.ravel())
pred = mod.predict(x_test)
```

Performance .

```
metrics.accuracy_score(y_test, pred)
```

```
1.0
```

```
report = classification_report(y_test, pred, output_dict=True)
pd.DataFrame(report).transpose()
```

	precision	recall	f1-score	support
Hepatitis D	1.0	1.0	1.0	20.0
Hepatitis E	1.0	1.0	1.0	24.0
Hypertension	1.0	1.0	1.0	22.0
Hyperthyroidism	1.0	1.0	1.0	26.0
Hypoglycemia	1.0	1.0	1.0	28.0
Hypothyroidism	1.0	1.0	1.0	28.0
Impetigo	1.0	1.0	1.0	30.0
Jaundice	1.0	1.0	1.0	24.0
Malaria	1.0	1.0	1.0	25.0
Migraine	1.0	1.0	1.0	27.0
Osteoarthristis	1.0	1.0	1.0	26.0
Paralysis (brain hemorrhage)	1.0	1.0	1.0	25.0
Peptic ulcer diseae	1.0	1.0	1.0	22.0
Pneumonia	1.0	1.0	1.0	24.0
Psoriasis	1.0	1.0	1.0	27.0
Tuberculosis	1.0	1.0	1.0	22.0
Typhoid	1.0	1.0	1.0	24.0
Urinary tract infection	1.0	1.0	1.0	24.0
Varicose veins	1.0	1.0	1.0	21.0
hepatitis A	1.0	1.0	1.0	28.0
accuracy	1.0	1.0	1.0	1.0
macro avg	1.0	1.0	1.0	984.0
weighted avg	1.0	1.0	1.0	984.0

Performance .

```
cm = confusion_matrix(y_test, pred)
pd.DataFrame(cm)
```

28	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	32	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	19	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	24	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	23	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	21	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	23	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	25	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	27	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	19	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	33	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	18	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	24	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	23	0

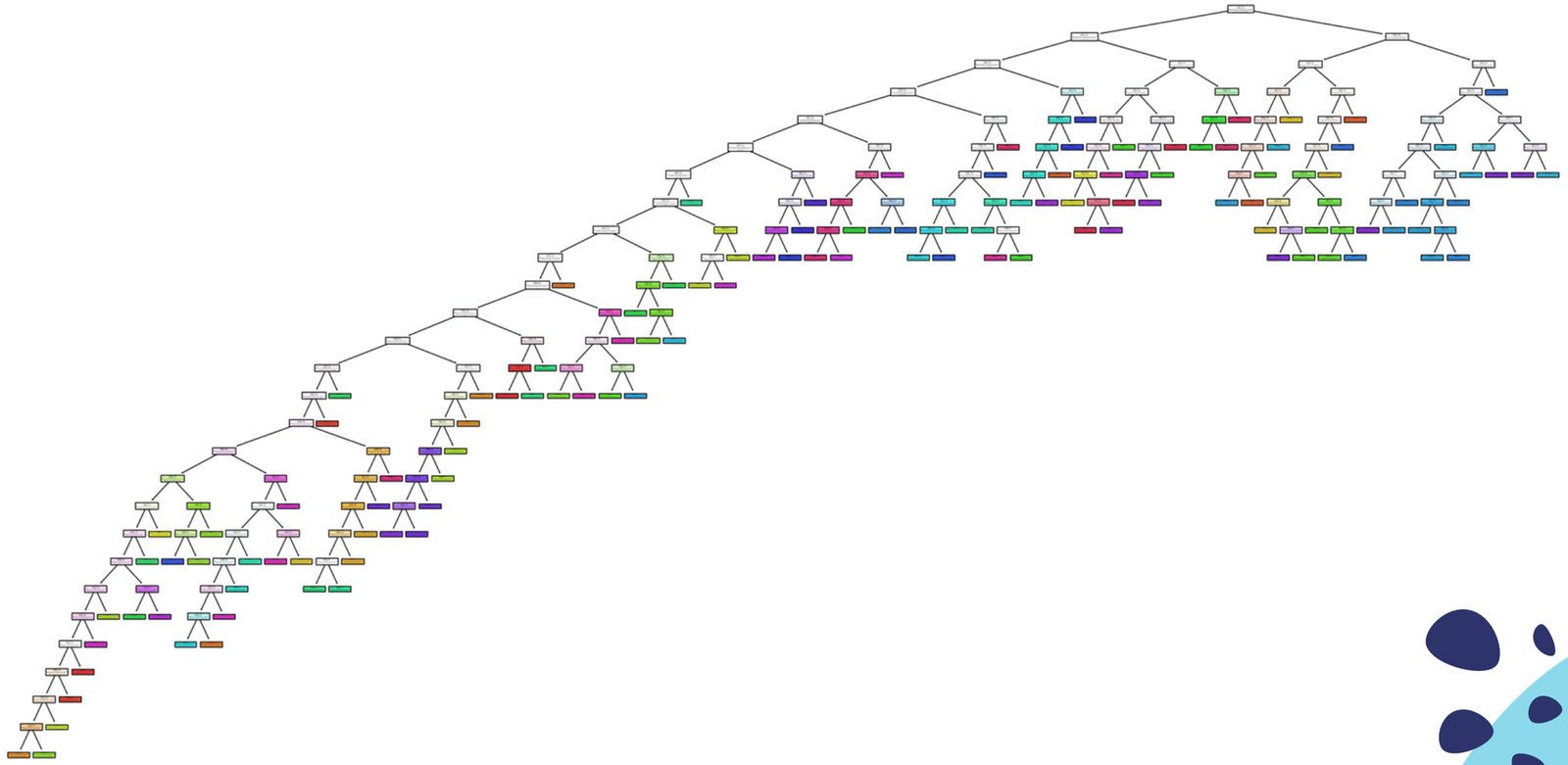
```
test = test.join(pd.DataFrame(mod.predict(C),columns=["predicted"]))[["prognosis","predicted"]]

test['result']= ' '
for i in range(len(test)):
    if test["prognosis"][i] == test["predicted"][i]:
        test['result'].iloc[i] = 'Correct'
    else:
        test['result'].iloc[i] = 'Incorrect'
```

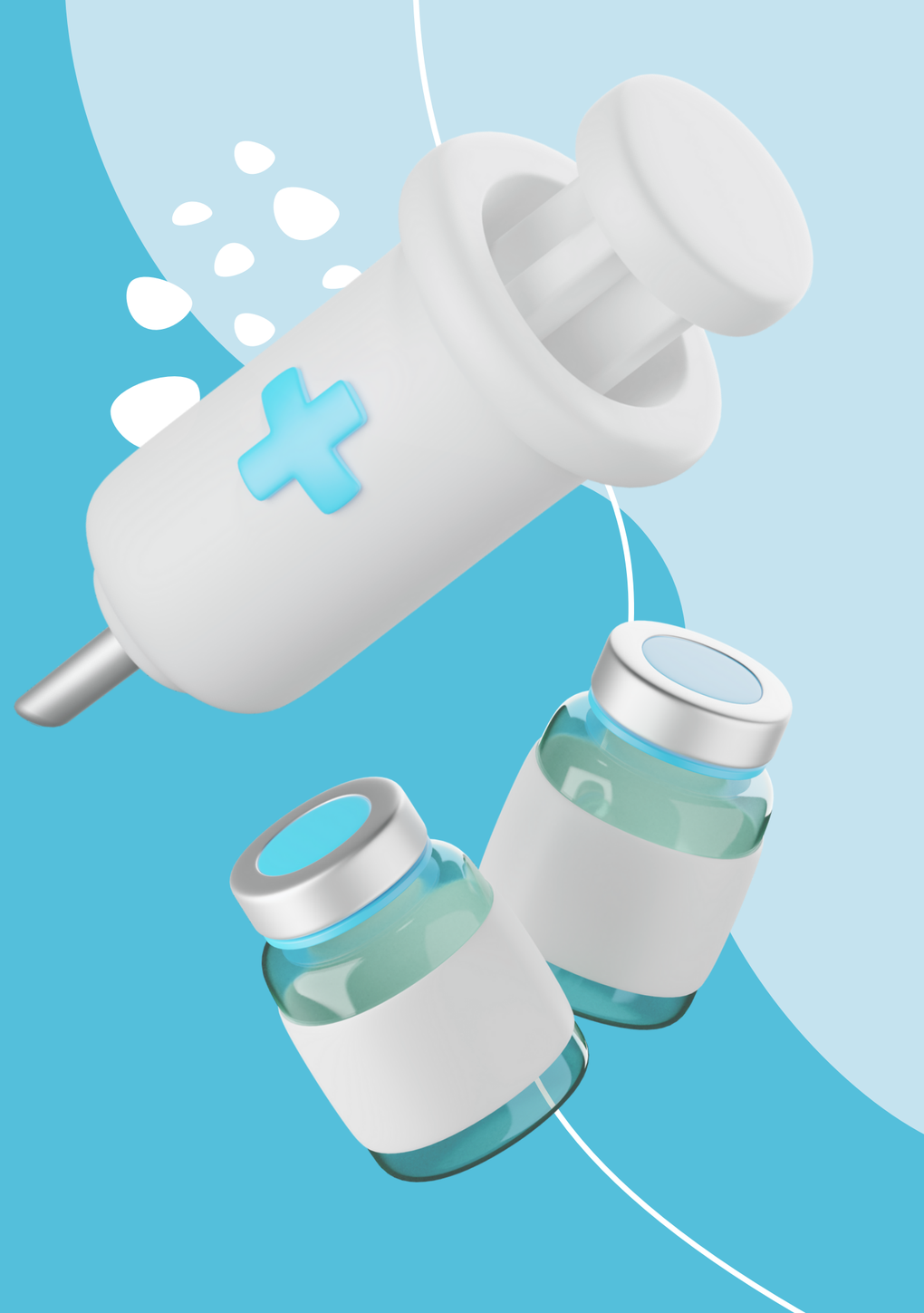
	prognosis	predicted	result
0	Fungal infection	Fungal infection	Correct
1	Allergy	Allergy	Correct
2	GERD	GERD	Correct
3	Chronic cholestasis	Chronic cholestasis	Correct
4	Drug Reaction	Drug Reaction	Correct
5	Peptic ulcer diseae	Peptic ulcer diseae	Correct

Visualising.

```
from sklearn import tree
plt.figure(figsize=(30,15))
tree.plot_tree(mod.estimators_[8],filled = True)
```



Ethical & Professional responsibilities



- **Respect privacy and data security:** Ensure that all data used in the project is collected and stored in a secure manner and that all personal information is handled with respect for the individual's privacy.

- **Adhere to applicable laws and regulations:** Ensure that any code or algorithms used in the project are properly licensed or developed from scratch, and do not infringe on any third-party intellectual property rights.

- **Be transparent:** Make sure that the purpose of the project is clearly stated and understood by all stakeholders, and that any potential risks are identified and discussed.

- **Ensure the accuracy of data and algorithms:** Ensure that machine learning models and algorithms used are free from bias or discrimination. Also, Ensure that the data used to train machine learning models is accurate and up-to-date.

- **Monitor performance:** Monitor the performance of the machine learning system over time to ensure it continues to meet its objectives and take corrective action if necessary.

CONCLUSION

The main aim of this disease prediction model is to predict the disease based on symptoms This model takes the symptoms of the user from which he or she suffers as input and generates the final output as a prediction of disease. In conclusion, for disease risk modeling, the accuracy of risk prediction depends on the diverse feature of the Health centers' data. Findings may help inform future developers of Disease Predictability Software and promote personalized patient care. The model predicts Patient Diseases through the Random Forest algorithm. Model accuracy reaches 100%. Machine learning skills are designed for Disease Prediction successfully

Thank you <3
We hope you enjoyed

جامعة جدة
University of Jeddah

Team Members:

Raneen A. Alshehri	2005560
Ghaid E. Althobaity	2005564
Noor O. Alamoudi	2005841
Taif A. Basheikh	2005890
Raghad Alqithmi	2005999