PROJECT TITLE: BUILDING A SMARTER AI-POWERED SPAM CLASSIFIER.

OVERVIEW OF DATASET:

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according to being ham (legitimate) or spam.

Data processing

- Import the required packages
- Loading the Dataset
- Remove the unwanted data columns
- Preprocessing and Exploring the Dataset
- Build word cloud to see which message is spam and which is not.
- Remove the stop words and punctuations
- Convert the text data into vectors

Building a sms spam classification model

- Split the data into train and test sets
- Use Sklearn built-in classifiers to build the models
- Train the data on the model
- Make predictions on new data

Import the required packages

```
%matplotlib inline

import matplotlib.pyplot as plt

import csv

import sklearn

import pickle

from wordcloud import WordCloud

import pandas as pd

import numpy as np

import nltk

from nltk.corpus import stopwords

from sklearn.feature_extraction.text import CountVectorizer,
TfidfTransformer

from sklearn.tree import DecisionTreeClassifier.
```

Loading the dataset

```
data = pd.read_csv('dataset/spam.csv', encoding='latin-1')

data.head()
```

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

Preprocessing and exploring dataset

```
# Import nltk packages and Punkt Tokenizer Models
import nltk
nltk.download("punkt")
import warnings
warnings.filterwarnings('ignore')
```

	label	text
1990	ham	HI DARLIN IVE JUST GOT BACK AND I HAD A REALLY...
1991	ham	No other Valentines huh? The proof is on your ...
1992	spam	Free tones Hope you enjoyed your new content. ...
1993	ham	Eh den sat u book e kb liao huh...
1994	ham	Have you been practising your curtsey?
1995	ham	Shall i come to get pickle
1996	ham	Lol boo I was hoping for a laugh
1997	ham	\YEH I AM DEF UP4 SOMETHING SAT
1998	ham	Well, I have to leave for my class babe ... Yo...
1999	ham	LMAO where's your fish memory when I need it?

```
ham_words = ''

spam_words = ''

# Creating a corpus of spam messages

for val in data[data['label'] == 'spam'].text:

    text = val.lower()

    tokens = nltk.word_tokenize(text)

    for words in tokens:

        spam_words = spam_words + words + ' '
```

```
# Creating a corpus of ham messages

for val in data[data['label'] == 'ham'].text:

    text = text.lower()

    tokens = nltk.word_tokenize(text)

    for words in tokens:

        ham_words = ham_words + words + ' '



spam_wordcloud = WordCloud(width=500,
height=300).generate(spam_words)

ham_wordcloud = WordCloud(width=500,
height=300).generate(ham_words)

#Spam Word cloud

plt.figure( figsize=(10,8), facecolor='w')

plt.imshow(spam_wordcloud)

plt.axis("off")

plt.tight_layout(pad=0)

plt.show()

#Creating Ham wordcloud

plt.figure( figsize=(10,8), facecolor='g')
```

```
plt.imshow(ham_wordcloud)

plt.axis("off")

plt.tight_layout(pad=0)

plt.show()
```

	label	text
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...
5	1	FreeMsg Hey there darling it's been 3 week's n...
6	0	Even my brother is not like to speak with me. ...
7	0	As per your request 'Melle Melle (Oru Minnamin...
8	1	WINNER!! As a valued network customer you have...
9	1	Had your mobile 11 months or more? U R entitle...

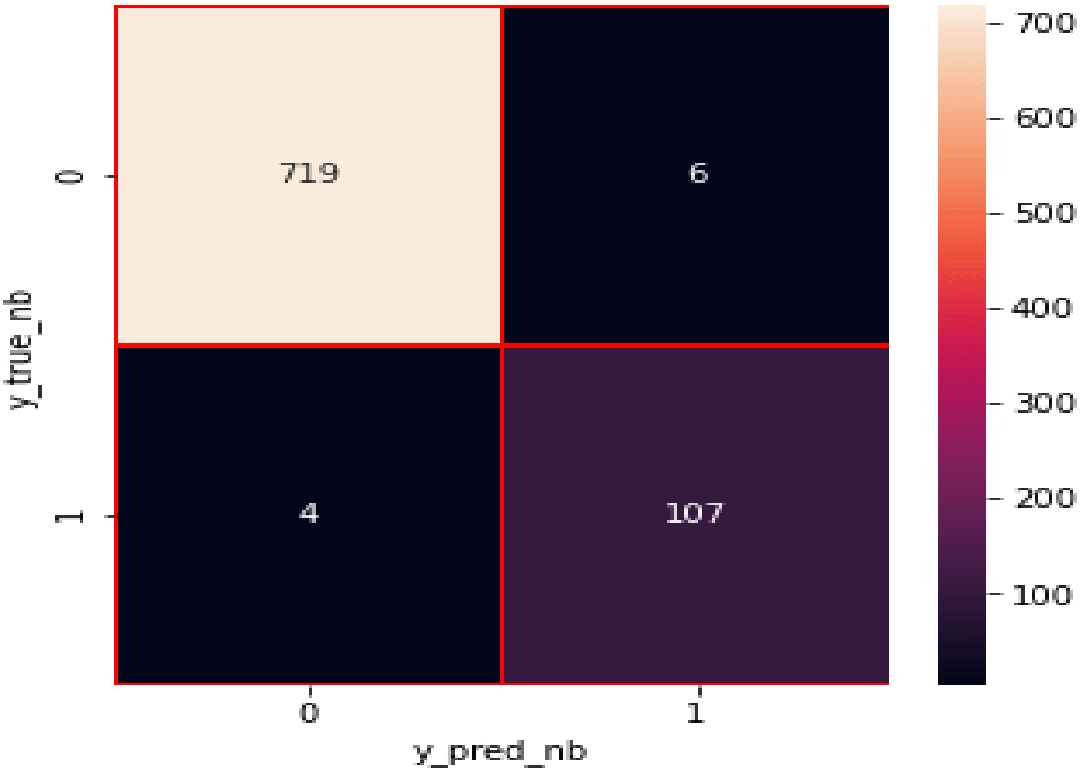
```
import nltk

nltk.download('stopwords')



#remove the punctuations and stopwords
```

```
import string

def text_process(text):



    text = text.translate(str.maketrans('', '', string.punctuation))

    text = [word for word in text.split() if word.lower() not in
stopwords.words('english')]



    return " ".join(text)



data['text'] = data['text'].apply(text_process)

data.head()
```

	label	text
0	0	Go jurong point crazy Available bugis n great ...
1	0	Ok lar Joking wif u oni
2	1	Free entry 2 wkly comp win FA Cup final tkts 2...
3	0	U dun say early hor U c already say
4	0	Nah dont think goes usf lives around though

Checking classification results with confusion matrix

```
from sklearn.metrics import confusion_matrix

import seaborn as sns

# Naive Bayes

y_pred_nb = mnb.predict(X_test)

y_true_nb = y_test

cm = confusion_matrix(y_true_nb, y_pred_nb)

f, ax = plt.subplots(figsize =(5,5))

sns.heatmap(cm,annot = True,linewidths=0.5,linecolor="red",fmt =
".0f",ax=ax)

plt.xlabel("y_pred_nb")

plt.ylabel("y_true_nb")

plt.show()
```



from the confusion matrix, we can see that the Naive Bayes model is balanced. That's it !! we have successfully created a spam classifier.