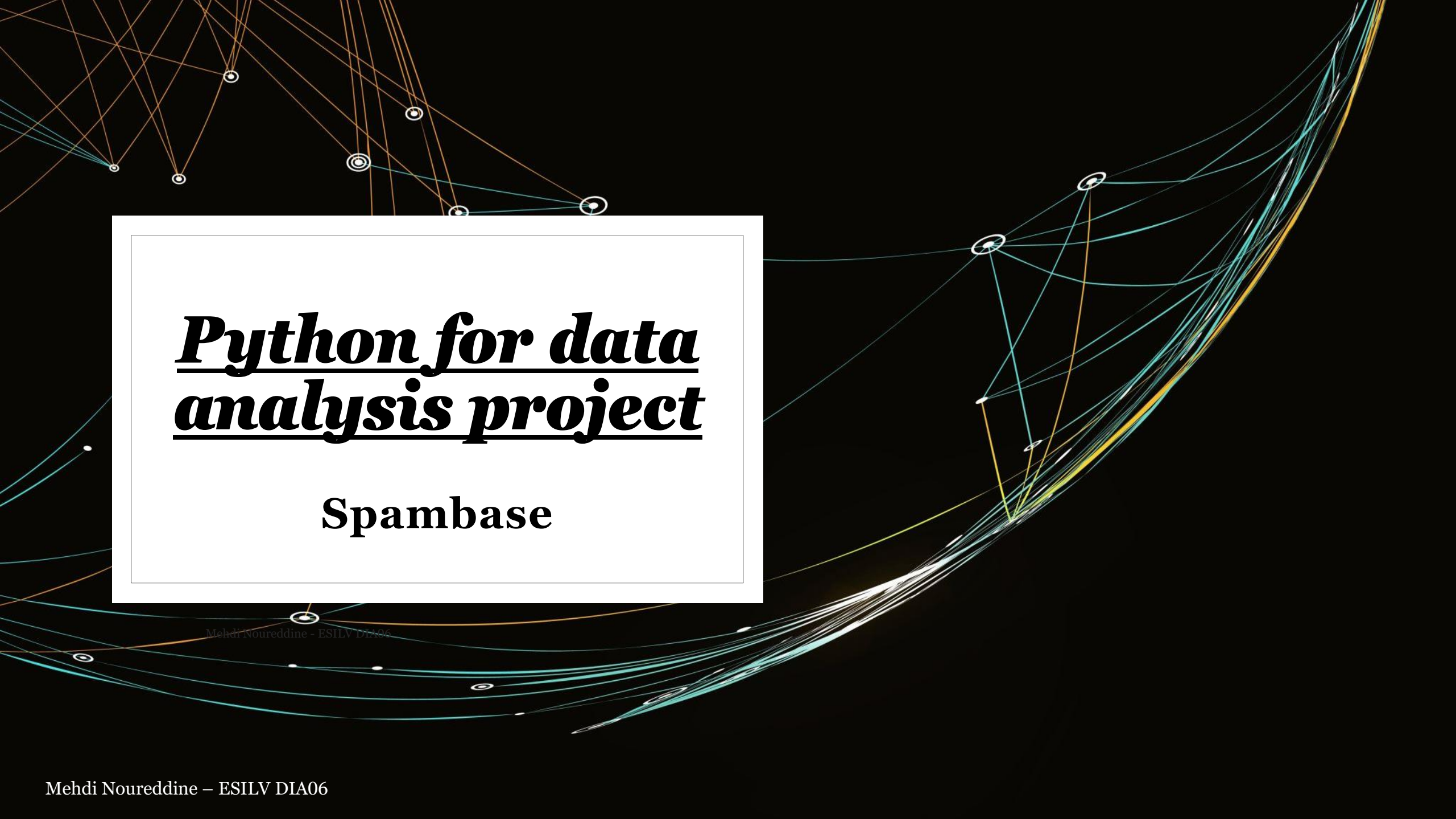

Python for data analysis project

Spambase

Problematic:

Classifying Email as Spam or Non-Spam

Data:

A database containing spam mails and personal or work non-spam mails

The parameters in the database give information on:

- Frequency of certain words or symbols
- Average length of uninterrupted sequences of capital letters
- Length of longest uninterrupted sequence of capital letters
- Total number of capital letters in the e-mail
- If the mail is a spam or not

Interpretation:

- First of all, the presence of certain words such as « project », « meeting » or « conference » are key indicators to tell if a message is a spam or not: these words are most likely to be found in mails that are sent by colleagues.
- The presence of long uninterrupted sequences of capital letters could also be the thing that gives away the spam nature of an email: work mails don’t contain capital words or sentences.

Resolution:

- To treat this problem, the first step was to visualize the dataset and the influence of the different variables, thanks to a PCA among others.
- The next step was to create training and testing sets by dividing the dataset given in order to, first of all, train our model by giving it the answer, and then test it by asking him to predict if an email is a spam or not, only using the parameters he is given. These predictions are then verified thanks to the dataset.
- Finally, the model was improved by changing hyperparameters that influence how the model functions.

Conclusion:

- After testing multiple models and then improving the final one, our model is capable of predicting if an email is a spam or not with a precision of around 95%, which is an acceptable result.