## Sensitive Data Indicator Model

Model aims to detect sensitive data for a given dataset. Model is created by using Convolutional Neural Networks (CNN) and Name Entity Recognition (NER).

1. **Introduction**

This notebook aims to capture Sensitive Data (Name, email address, password, phone number, date of birth etc) for a given dataset. Two different models were used and merged in one notebook.

**First Model** uses Convolutional Neural Networks (CNN), this model is trained with Sensitive and Non Sensitive Datasets. Output gives the probability of sensitivity for a given sentence. This model is best for a text (sentence like) input data.

**Second Model** uses Presidio Analyzer and Presidio Anonymizer. Anonymizing data is optional. Name Entity Recogntion (NER) models are pretrained models and they don't need to be trained. Entities such as CREDIT\_CARD, IBAN\_CODE, EMAIL\_ADDRESS, BANK\_NUMBER can be detected and anonymized(optional) with this model. Output gives detection of sensitive data for chosen entities. This model can be used for text (sentence like) input data or Data Frames as input.

1. **Technologies Used**

Python Libraries

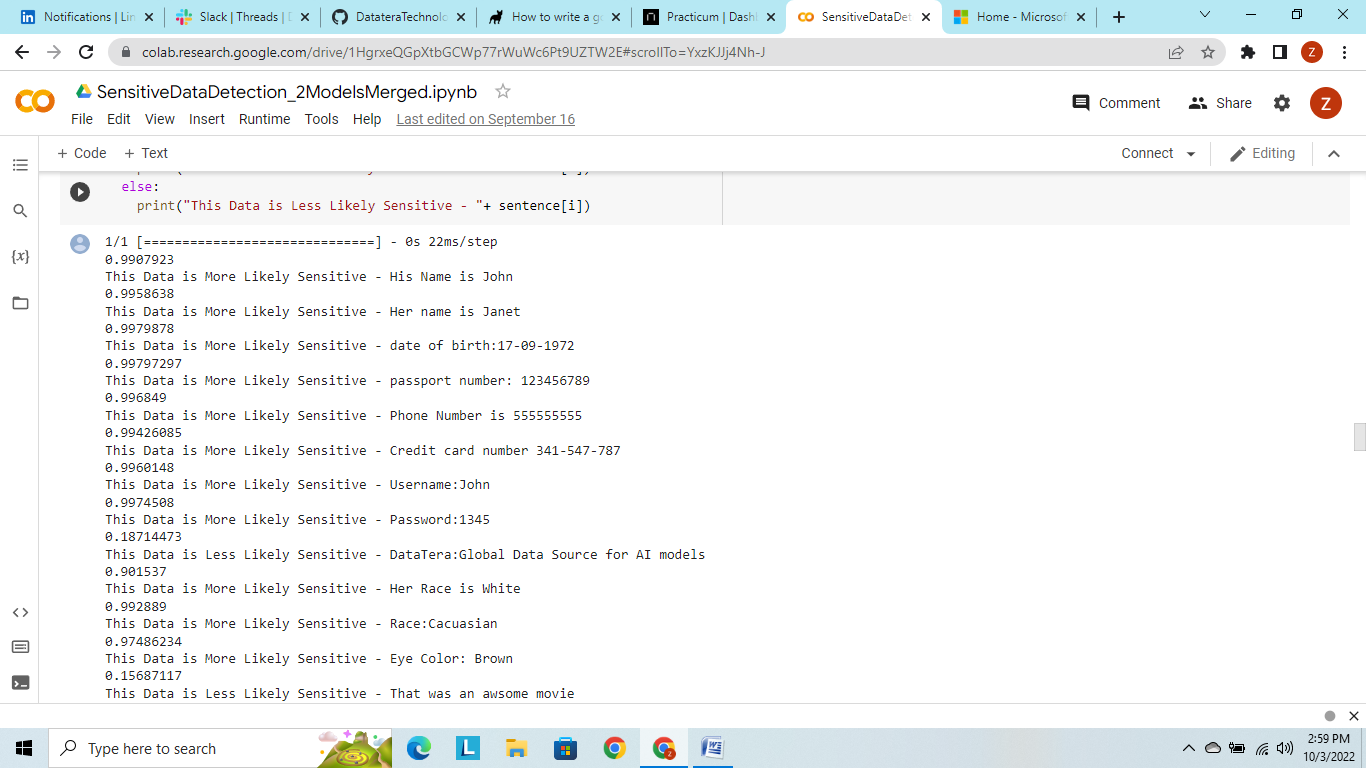
1. **Tensor Flow**-Tensor Flow is a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top of TensorFlow.
2. **Keras**-Keras is a high-level neural network library that runs on top of TensorFlow.
3. **Presidio Analyzer**-The Presidio analyzer is a Python based service for detecting PII entities in text.
4. **Presidio Anonymizer**-The Presidio anonymizer is a Python based module for anonymizing detected PII text entities with desired values.
5. **spaCy** is a free open-source library for Natural Language Processing in **Python**. It features NER, POS tagging, dependency parsing, word vectors and more.

Hugging Face

Hugging Face is **a community and data science platform** that provides: Tools that enable users to build, train and deploy ML models based on open source (OS) code and technologies.

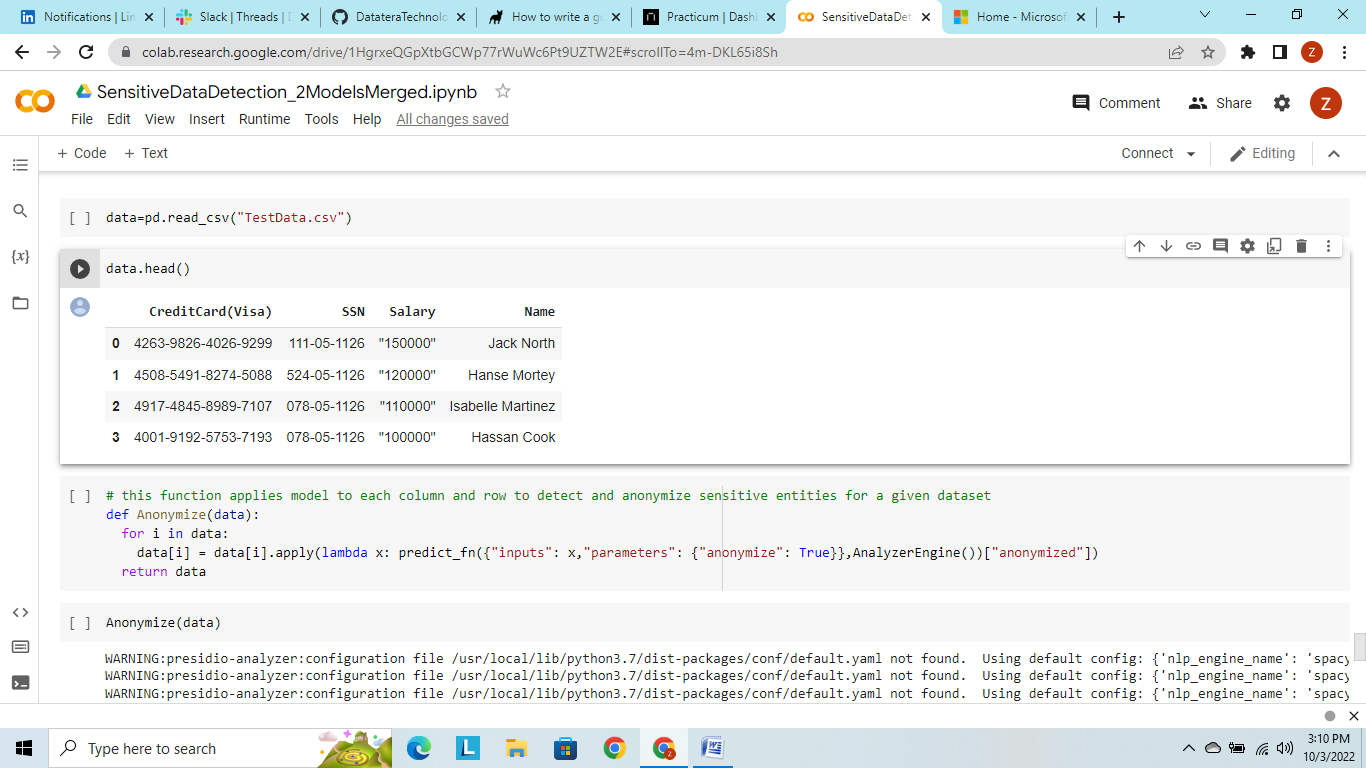
1. **Illustration**

**First model** gives the probability of sensitive data for a given sentence-like input. As seen below, name, date of birth, passport number, phone number, credit card number and password are given as sensitive to the model. Model shows the probability of data sensitivity.

****

**Second Model** detects sensitive data for a given test dataset as shown below;

Input;

****

Output;

