REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

REVA Academy for Corporate Excellence (RACE)

AI/ML Based Sensitive Data Discovery and Classification of Unstructured Data Sources

By,

Name: Shravani

Batch: BA 03

SRN: R18DMO18

Date: 27-08-2022

race.reva.edu.in

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Agenda

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Background Information

4 *Common Challenges of identifying sensitive data*

1. The volume of the data owned by organizations is growing daily and the data management is becoming a huge challenge.
2. It is estimated that **80-90%** of the data is in unstructured format.
3. As per Forbes, almost **95%** businesses are struggling to manage unstructured data.
4. Data leakages, data breaches are also increasing drastically resulting in heavy penalties for the organizations

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Problem Statement

*According to ISO 14001, compliance obligations are **legal requirements that an organization must comply with and other needs that an organization has to, or chooses to comply with**. These needs can include laws and regulations, contracts, codes of practice and voluntary commitments like industry standards. Below are some of the key compliance areas with respect to the data .*

Regulatory Compliance

Comprises the activities that support the coordination, management, and monitoring of various federal, state, and local laws and regulations.

Data Privacy Compliance

Data privacy compliance concerns the proper handling of sensitive data including, notably, personal data but also other confidential data

Data Protection Compliance

Data Protection Compliance is the need to comply with legal requirements regarding data processes

Data Governance

Data governance is an umbrella term that encapsulates the policies and practices implemented to securely manage the data assets within an organization

Data Management

Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively.

Data Security

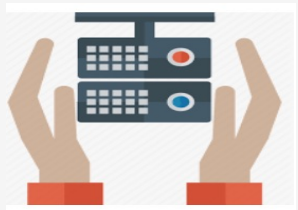
Data security refers to the process of protecting data from unauthorized access and data corruption throughout its lifecycle

Popular Compliance Programs

- *GDPR*
- *CCPA*
- *HIPPA*
- *BCBS 239*

Introduction 1/2

Sensitive Data



*Sensitive data is confidential information that must be protected and can be in both **structured** and **unstructured** formats.*

Sensitive data is very important to locate, identify and protect

Types of Sensitive Data

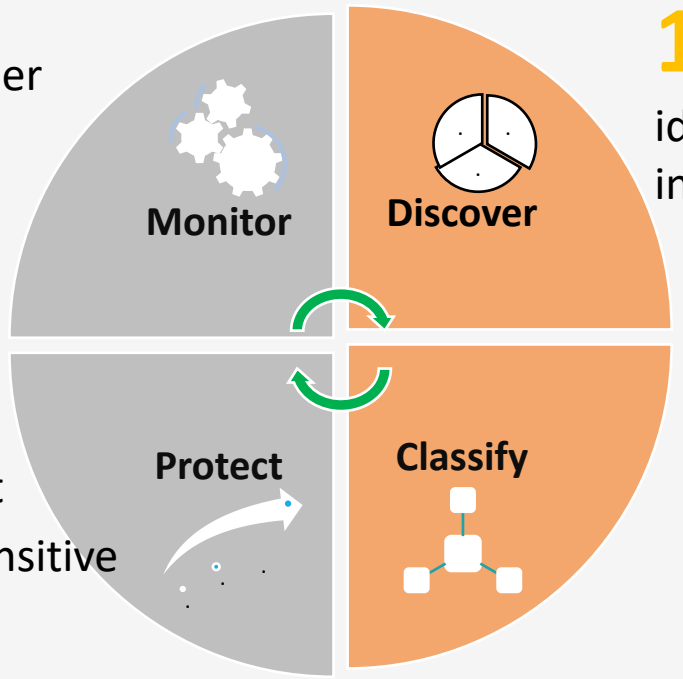
PII	SPI	PHI	NPI
Name, Email, Phone number, DOB, Address	SSN, PAN Drivers license, passport Race/ Ethnicity Religious/Philosophical beliefs Political Opinion Geolocation Email, texts Genetic/ Biometric data	Medical Info Physical or Mental health related info	Bank account number Credit card number
Personally Identifiable Information	**Sensitive Personal Information**	**Protected Health Information**	**Non-public Personal Information**

Introduction 2/2

Data protection lifecycle (DPL)

Data protection lifecycle (DPL) *helps organizations to manage the sensitive data. This illustrates how to discover, classify, protect and monitor the sensitive data . By accurately tracking sensitive data, organizations have a foundation to protect sensitive information and face the future privacy needs.*

4. Monitor end user activity and the protected files.



1. Scans data and identifies sensitive information.

3. Encrypt/ restrict access/mask the sensitive information.

2. Classify the data based on the identified sensitive information.

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Objectives

Part I

Part II

Part III

Part IV

Detect the sensitive data elements	*Document Risk Categorization*	*Auto Classification*	*Visualize the results*
Detect sensitive elements contained in the unstructured document using **Regular Expressions**	**Rule based** document Risk categorization: High, Medium, Low	Auto classify: Public, Internal, Confidential, Restricted using **Multi Class Classification Model**	Visualize the results on an executive dashboard in **Tableau**

Project Methodology 1/3

Part I - Sensitive Data Elements

	Format	Sample
Name	First Name, Last Name	Mark Campbell
		Lisa Thomas

Keywords: Name, full name, full_name, first name, last name

	Format	Sample
SSN	ddd-dd-dddd	986-43-245
	ddd dd dddd	231 24 3168

Social Security Number consists of 9 digits. The first set of three digits is called the Area Number. The second set of two digits is called the Group Number. The final set of four digits is the Serial Number.

Keywords: ssn, social security number, ssn number, social security #, social security no, soc sec

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Project Methodology 2/3

Part II – Document Risk Categorization

The Table illustrates the Risk Categorization criteria:

- If a document contains a name along with SSN, PAN, or DOB is categorized as a High-Risk Document.
- If a document includes either SSN, PAN, Phone number, or e-mail address is categorized as a Medium Risk document.
- If a document contains only a name or DOB is categorized as a Low-Risk document.

Sensitive Data Element	Document Risk Categorization		
	High	Medium	Low
Name			Yes
SSN		Yes	
PAN		Yes	
DOB			Yes
Phone Number		Yes	
Email		Yes	
Name + SSN	Yes		
Name + PAN	Yes		
Name + Phone Number		Yes	
Name + email		Yes	
Name + DOB	Yes		

Project Methodology 3/3

Part III – Auto Classification

The Table shows the type of information contained in each document category. For example, an organization's public document can contain financial statements, press releases, etc. In contrast, a restricted document might contain sensitive information like SSN or Bank Account Numbers.

Since, the documents have more than one category, multi-class classification algorithm is used to predict the class of the document.

Document Classification per Classification Policy

Public	Internal	Confidential	Restricted
		Non Sensitive PII:	**Sensitive PII:**
Financial Statements	Training Materials	Name	SSN/ PAN
Press Release	Instructions	Phone Number	Date of Birth
		E-mail address	Bank Account Number

Data Generation 1/2

Sensitive information (PII) *is not available on open sources due to restrictions.*

Since the project simulation needs the sensitive information, we have generated the ***synthetic data*** *which mimics PII while preserving the format and data type. This dataset will be used as an input to activate and test the model and auto detect the sensitive data elements and classify them.*

```
#Generate randon names
import names
names.get_full_name()
```

```
'Scott Spangler'
```

```
#Generate Random phone no's
import random
random.randint(7000000000,9999999999)
```

```
email = ".".join( Names_list[0].split())+"@gmail.com"
email
```

```
'Gregory.Mccullough@gmail.com'
```

```
#Generate DOB
monthly_days = np.arange(0, 30)
base_date = np.datetime64('1982-12-01')
base_date + np.random.choice(monthly_days)
```

```
numpy.datetime64('1982-12-09')
```

	Name	DOB	email	Mobile Number
0	Gregory Mccullough	1978-06-29	Gregory.Mccullough@gmail.com	917996523744
1	Sophia Villarreal	1978-06-25	Sophia.Villarreal@gmail.com	918244087321
2	Brian Gallipeau	1978-06-25	Brian.Gallipeau@gmail.com	918854475702

Data Generation 2/2

Name Generation

Names is a python library that generates random names.

Figure shows how this library was used for generating full names of different genders.

```
In [5]: #Generate randon names
        import names
        names.get_full_name()

Out[5]: 'Scott Spangler'
```

```
In [11]: names.get_full_name(gender='female')

Out[11]: 'Lisa Thomas'

In [14]: names.get_full_name(gender='male')

Out[14]: 'Robert Hernandez'
```

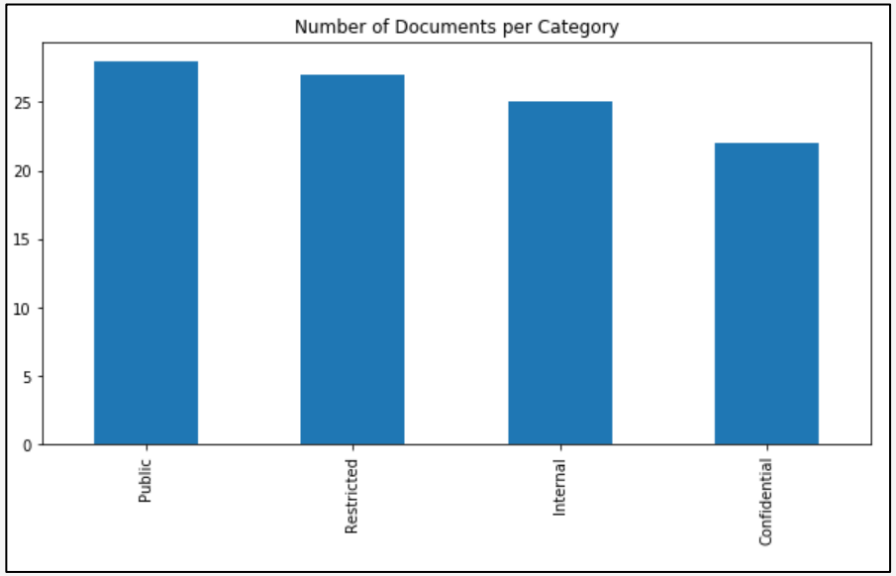
SSN generation

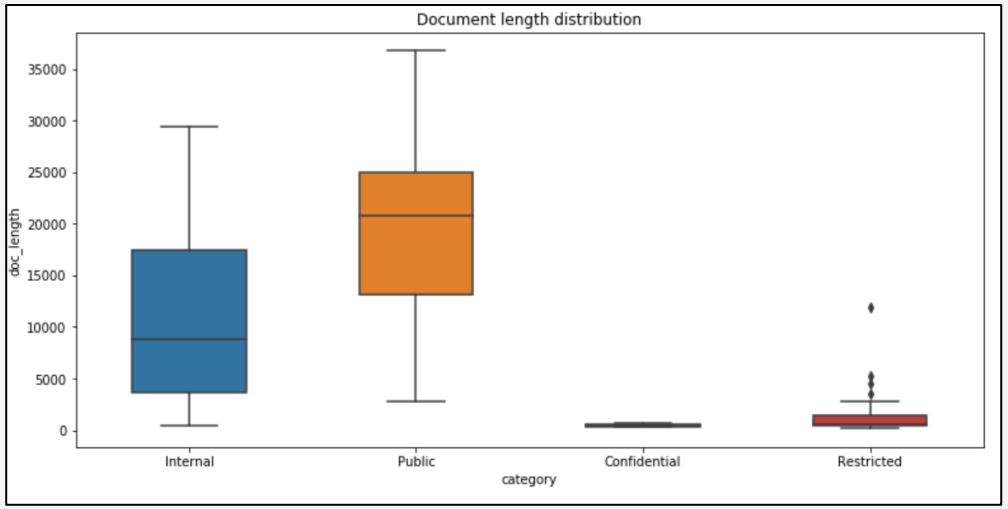
Figure shows how SSN is generated using Python library random, a set of two and four digit number is appended with the area code to generate a SSN.

```
In [121]: import names
          Names_list = []
          i=0
          while i <= 15:
              Names_list.append(names.get_full_name())
              i += 1

In [122]: #Generate SSN
          import random
          Number1_list = []
          Number2_list = []
          i=0
          while i <= 15:
              Number1_list.append(str(random.randint(10,99)))
              Number2_list.append(str(random.randint(1000,9999)))
              i += 1

In [123]: Number3_list = ['-'.join(x) for x in zip(Number1_list, Number2_list)]

In [124]: append_str = '040-'
          ssn = [append_str + sub for sub in Number3_list]
```

Data Exploration

Documents per Category
The figure shows the number of documents per category.





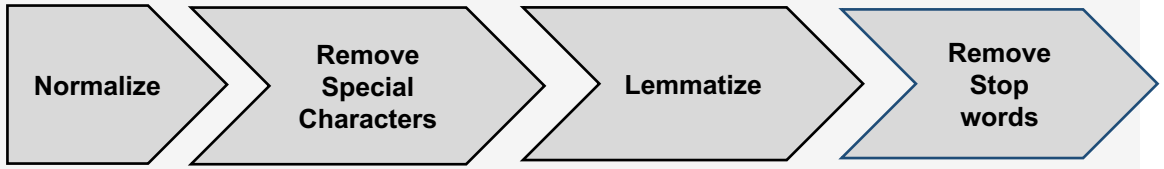
Documents length distribution
The figure shows the document length distribution per category.

Data Preparation

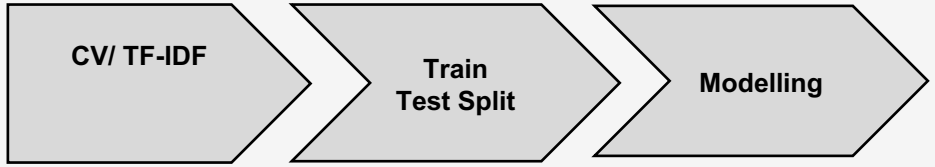
*For **Data Preparation**, we want to create features from the raw text so we can train the machine learning models.*

The steps followed are:

- *Text Cleaning*
- *Text Featurizer*
- Train-test split



Text Cleaning Pipeline



Feature Engineering Pipeline

Data Modelling

Multi-class classification Algorithm is used to Classify the document.

After Data is cleaned and prepared, we use this final dataset for the next step - Data Modeling.

We trained the below multi class classification Algorithms using both Count Vectorizer and TF-IDF features:

- Random Forest
- Multinomial Naïve Bayes
- K-Nearest Neighbors
- Decision Trees

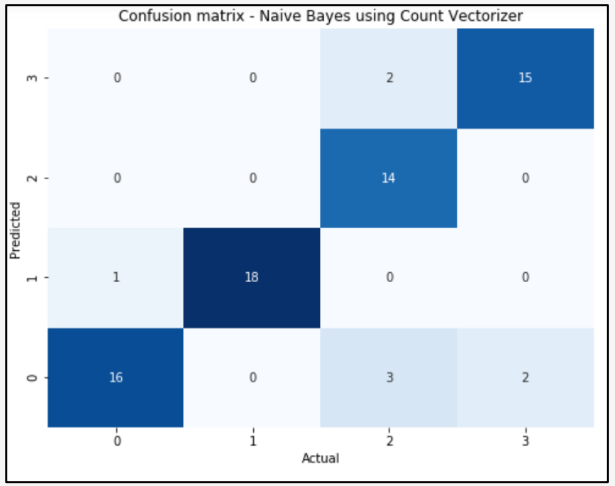
Data Modelling Results

Classifier	Count Vectorizer		TF-IDF	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
Random Forest	98%	80%	97%	68%
Multinomial Naïve Bayes	88%	**90%**	95%	62%
K Neighbors	68%	68%	88%	50%
Decision Tree	100%	83%	100%	56%

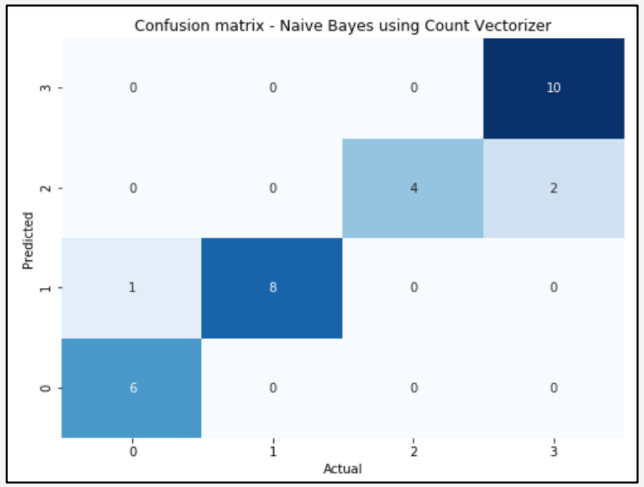
REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Naïve Bayes Model Evaluation

Train



Test



Internal - 0, Public - 1, Confidential - 2, Restricted -3

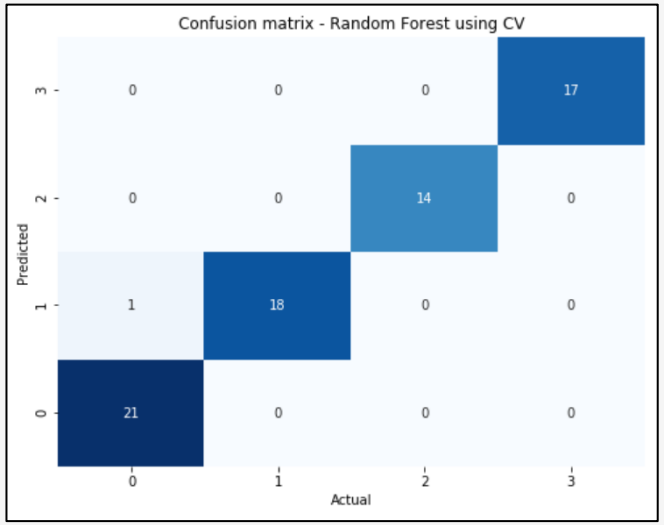
Classification report

	precision	recall	f1-score	support
1	0.86	1.00	0.92	6
2	1.00	0.89	0.94	9
3	1.00	0.67	0.80	6
4	0.83	1.00	0.91	10
accuracy			0.90	31
macro avg	0.92	0.89	0.89	31
weighted avg	0.92	0.90	0.90	31

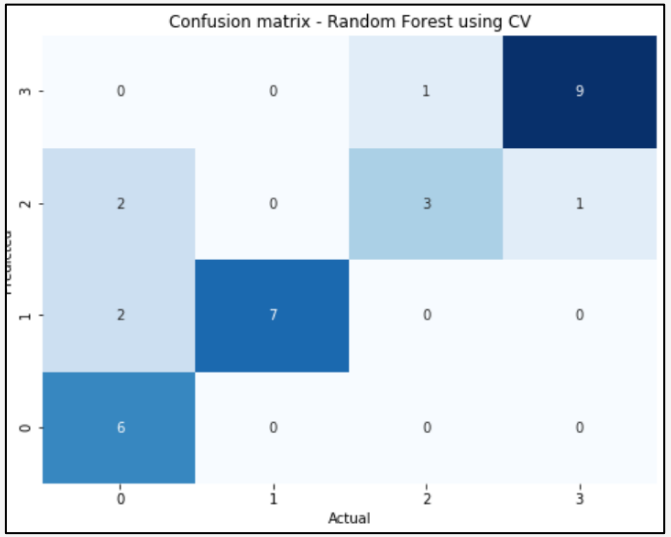
REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Random Forest Model Evaluation

Train



Test



Internal - 0, Public - 1, Confidential - 2, Restricted -3

Classification report

	precision	recall	f1-score	support
1	0.60	1.00	0.75	6
2	1.00	0.78	0.88	9
3	0.75	0.50	0.60	6
4	0.90	0.90	0.90	10
accuracy			0.81	31
macro avg	0.81	0.79	0.78	31
weighted avg	0.84	0.81	0.81	31

Results 1/4

Part I - Sensitive Data Elements

Name
The figure highlights the keyword ‘name’ in the document along with 11 names.

```
--------------------------------------------------
--------------------------------------------------

Results:
Name keyword found. Keyword: name

1. Number of Names found: 11

--------------------------------------------------
```

Phone Numbers
The figure highlights the keyword ‘number’ in the document along with the number of phone numbers found.

```
--------------------------------------------------

Phone number keyword found. Keyword: number
--------------------------------------------------

2.1 Number of Indian Phone Numbers found: 26

2.2 Number of US Phone Numbers found: 11

2.3 Number of UK Phone Numbers found: 9

2.4 Number of Australian Phone Numbers found: 0
--------------------------------------------------
```

Results 2/4

E-mail Address
The figure highlights the keyword ‘email’ found in the document along with the number of email address found.

```
--------------------------------------------------

email keyword found. Keyword: email

4. Number of personal emails found: 4

--------------------------------------------------
```

PAN/SSN
The figure highlights the keyword ‘PAN/SSN’ found in the document along with the number of PAN and SSN numbers found in the document.

```
--------------------------------------------------

PAN keyword found. Keyword: pan

ssn keyword found. Keyword: ssn

5.1 Number of PAN found: 1
5.2 Number of SSN found: 3

--------------------------------------------------
```

Results 3/4

Part II – Document Risk Categorization

As per the rule based risk categorization defined, the document is categorized as a High-Risk document since it contains restricted elements – SSN and PAN along with names and phone numbers.

Part III – Auto Classification

After the document was cleaned and processed. The document was classified as a Restricted Document using the naïve bayes model chosen.

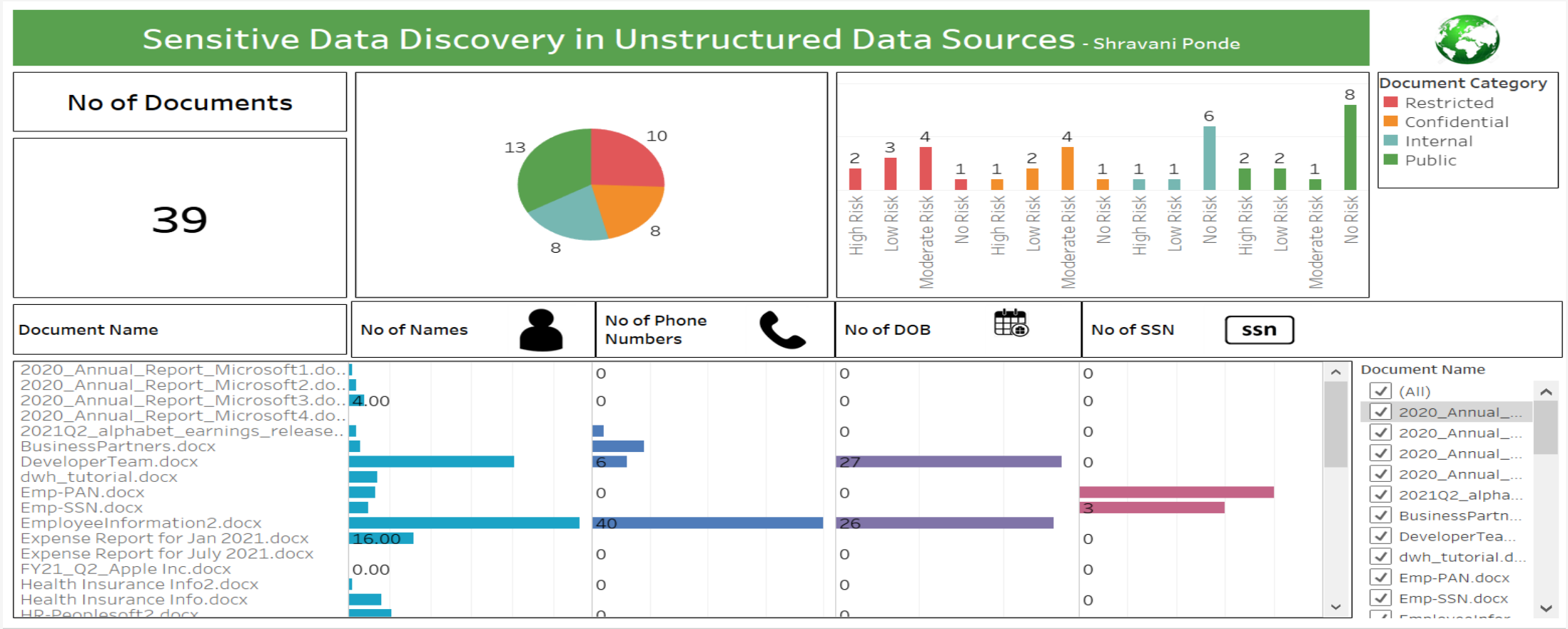
```
Calculating the risk score....
--------------------------------------------------
--------------------------------------------------

Summary:

The Data Security Classficiation of the document is: RESTRCITED
The Document Risk Categorization for the document is: High Risk

--------------------------------------------------
End
```

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Results 4/4

Part IV - Visualize

An Executive Tableau Dashboard was built to summarize the results of the documents scanned. It highlights the number of documents reviewed, Document Category and Risk and the number of sensitive elements found.



Conclusion & Future Scope

Conclusion

The discovery results and the risk classification of the word documents produced as a dashboard, allows the business stakeholders to take necessary actions in protecting their sensitive data assets from heterogenous unstructured sources.

Future Scope

After the Sensitive data discovery and classifying the sensitive data elements the below data privacy and data protection needs can be commenced and the organizations would establish a strong data protection and governance framework to handle regulatory and auditing challenges well in advance.

- Enabling access and security controls : Role based access control, Password protection control, API level Encryptions
- Enabling Data activity monitoring controls
- Data Protection capabilities : Masking, Tokenization, Anonymization, Pseudonymization, Encryption, Data Loss Preventions (DLP), Data Remediation, Data Subject Rights

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

References

Marr, B. (2018, May 21). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=4e4f805860ba

Gartner Top Strategic Technology Trends for 2022. (2022). Retrieved from Gartner: https://www.gartner.com/en/information-technology/insights/top-technology-trends

Goswami, S. (2020, December 14). *The Rising Concern Around Consumer Data And Privacy*. Retrieved from Forbes: https://www.forbes.com/sites/forbestechcouncil/2020/12/14/the-rising-concern-around-consumer-data-and-privacy/?sh=30741b43487e

Hill, M. (2022, August 16). *The 12 biggest data breach fines, penalties, and settlements so far*. Retrieved from CSO: https://www.csoonline.com/article/3410278/the-biggest-data-breach-fines-penalties-and-settlements-so-far.html

Kulkarni, R. (2019, 02 07). *Big Data Goes Big*. Retrieved from Forbes: https://www.forbes.com/sites/rkulkarni/2019/02/07/big-data-goes-big/?sh=278b2aa820d7

References

Wolford, B. (2020). *What is GDPR, the EU's new data protection law?* Retrieved from GDPR.EU: https://gdpr.eu/what-is-gdpr/

Bonta, R. (2022). *California Consumer Privacy Act (CCPA)*. Retrieved from State of California Department of Justice: https://oag.ca.gov/privacy/ccpa

Office for Civil Rights (OCR). (2022, January 19). *Your Rights Under HIPAA*. Retrieved from HHS.gov: https://www.hhs.gov/hipaa/for-individuals/guidance-materials-for-consumers/index.html

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Literature Review

K. Gai, M. Qiu and H. Zhao, "Privacy-Preserving Data Encryption Strategy for Big Data in Mobile Cloud Computing," in IEEE Transactions on Big Data, vol. 7, no. 4, pp. 678-688, 1 Oct. 2021, doi: 10.1109/TBDATA.2017.2705807. Privacy-Preserving Data Encryption Strategy for Big Data in Mobile Cloud Computing | IEEE Journals & Magazine | IEEE Xplore

S. Cha and K. Yeh, "A Data-Driven Security Risk Assessment Scheme for Personal Data Protection," in IEEE Access, vol. 6, pp. 50510-50517, 2018, doi: 10.1109/ACCESS.2018.2868726. A Data-Driven Security Risk Assessment Scheme for Personal Data Protection | IEEE Journals & Magazine | IEEE Xplore

P. Yang, N. Xiong and J. Ren, "Data Security and Privacy Protection for Cloud Storage: A Survey," in IEEE Access, vol. 8, pp. 131723-131740, 2020, doi: 10.1109/ACCESS.2020.3009876. Data Security and Privacy Protection for Cloud Storage: A Survey | IEEE Journals & Magazine | IEEE Xplore

N. B. Truong, K. Sun, G. M. Lee and Y. Guo, "GDPR-Compliant Personal Data Management: A Blockchain-Based Solution," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1746-1761, 2020, doi: 10.1109/TIFS.2019.2948287. GDPR-Compliant Personal Data Management: A Blockchain-Based Solution | IEEE Journals & Magazine | IEEE Xplore

L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren, "Information Security in Big Data: Privacy and Data Mining," in IEEE Access, vol. 2, pp. 1149-1176, 2014, doi: 10.1109/ACCESS.2014.2362522. Information Security in Big Data: Privacy and Data Mining | IEEE Journals & Magazine | IEEE Xplore

Gai, K., Qiu, M., & Zhao, H. (2017). Privacy-Preserving Data Encryption Strategy for Big Data in Mobile Cloud Computing. *IEEE*, 678 - 688.

REVA UNIVERSITY

Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Thank you!

The General Data Protection Regulation (GDPR)

European Union's (EU) GDPR is the law that imposes privacy regulations on any organization that accumulates or processes personal information related to individuals in the EU. Personal information includes but is not limited to names, email, location, ethnicity, gender, biometric data, religious beliefs, etc. All organizations are required to be GDPR compliant as of May 2018. The fines in case of GDPR violations are very high €20miilion or 4% of the global revenue.

The California Consumer Privacy Act (CCPA)

The CCPA of 2018 gives Californian consumers control over how an organization collects their personal information. The personal information includes but is not limited to name, social security number, products purchased, internet browsing history, geolocation data, etc. The CCPA provides consumers with three principal "rights." The first right is the "right to know" how the organization collects, uses, or shares personal information. The second right is the "right to opt-out" of selling personal data. The third right is the "right to delete" personal information collected about the consumer

The Health Insurance Portability and Accountability Act of 1996 (HIPAA)

HIPAA by the Department of Health and Human Services (HHS) gives consumers rights over their health information. Consumers have the right to get a copy of their health information, check who has it, and learn how it is used and shared. These regulations apply to health care providers, insurance companies,

Sl. No	Abbreviation	
1	GDPR	General Data Protection Regulation
2	BCBS	The Basel Committee on Banking Supervision
3	CCPA	The California Consumer Privacy Act of 2018
4	CCAR	The Comprehensive Capital Analysis and Review
5	HIPAA	Health Insurance Portability and Accountability Act