PROJETO INTEGRADOR II

TEMÁTICA: Aplicação e análises nos bancos de dados

Startup: TARGET TEC.

Iniciativa: Unificação de cadastros para o segmento de D2D(Porta a Porta)

Nossa startup possui o objetivo de concentrar as principais empresas do segmento de D2D em uma única plataforma em que, ao se cadastrar, o candidato terá a opção de enviar seus dados a um único player ou optar pela seleção de várias empresas para obter maior sucesso no seu objetivo sem a necessidade de acesso a vários sites;

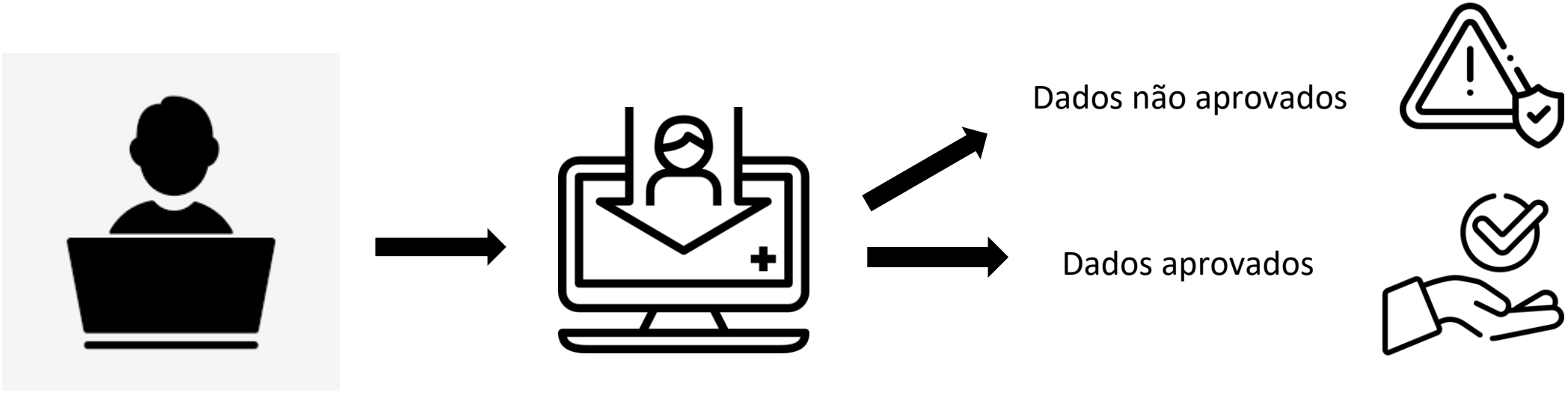
Além disso, para evitar que haja envio de dados incorretos, nossa startup realizará uma primeira análise desses dados como CPF (ativo ou não na Receita Federal) e documentoscopia (recebimento de documento do candidato para análise com o objetivo de prevenção à fraude);

Desta forma, somos facilitadores do processo tanto para o consumidor (candidato) quanto para a empresa, pois já enviaremos os dados analisados e garantindo uma assertividade no processo de análise.

FATEC
Faculdade de Tecnologia

Analise de documentoscopia

A documentoscopia é um instrumento utilizado para fins de análises de documentos com o objetivo de averiguar sua autenticidade para evitar eventuais fraudes.
Ela se mostra fundamental uma vez que consegue detectar se houve ou não adulteração em documentos importantes ou se eles não foram falsificados de alguma forma.
Nossa startup visa agilizar esse processo de análise e verificação para nossos clientes no decorrer de seu processo cadastral, trazendo assim maior segurança na interconexão proporcionada em nossa plataforma.



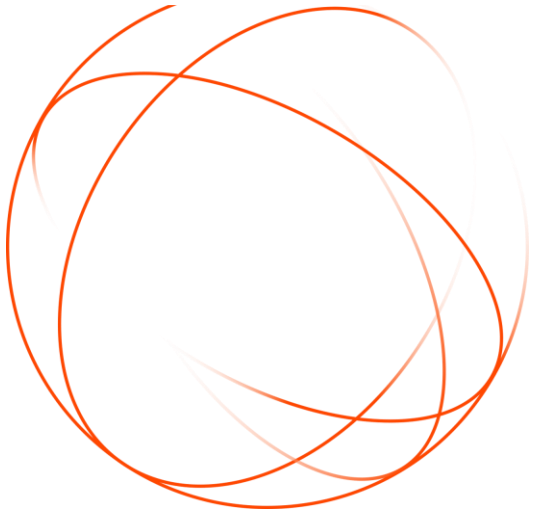
Analise de documentoscopia

•Para embasar e analisar qual seria a importância da implementação de analise de documentoscopia utilizamos uma base de dados do Kaggle que traz um levantamento de dados sobre fraudes de documentos

ndex	Identity Theft Reports by Age	Unnamed: 1
0	NaN	NaN
1	Age Range	# of Reports
2	19 and Under	13,852
3	20 - 29	61,114
4	30 - 39	80,467
5	40 - 49	70,264
6	50 - 59	63,004
7	60 - 69	45,787
8	70 - 79	15,979
9	80 and Over	5,359
10	NaN	NaN
11	NaN	NaN
12	NaN	NaN
13	NaN	NaN
14	NaN	NaN
15	NaN	NaN
16	NaN	NaN
17	NaN	NaN
18	NaN	NaN
19	NaN	NaN
20	NaN	NaN
21	NaN	NaN
22	NaN	NaN
23	NaN	NaN

Analise de documentoscopia

- Utilizando a biblioteca pandas do python no Colab conseguimos verificar os dados de relatórios de fraudes por idade.
- Ao analisa-lo é possivel comparar qual o maior indice de roubo de acordo com a faixa etária do publico que segmentamos para atender.



FATEC
Faculdade de Tecnologia

Analise de documentoscopia

```
Fraud Categories by Total Amount Lost          Internet Services
Unnamed: 1                                          $19,397,434
Unnamed: 2                                                $183
Unnamed: 3                                               45,093
Unnamed: 4                                                  14%
Unnamed: 5                               Consumer Initiated Contact
Unnamed: 6                                          Credit Card
Name: 9, dtype: object
```

```
Fraud Categories by Total Amount Lost    Business and Job Opportunities
Unnamed: 1                                             $46,914,979
Unnamed: 2                                                $1,063
Unnamed: 3                                                18,702
Unnamed: 4                                                   34%
Unnamed: 5                                                 Email
Unnamed: 6                                          Wire Transfer
Name: 6, dtype: object
Fraud Categories by Total Amount Lost          Investment Related
Unnamed: 1                                             $47,697,104
Unnamed: 2                                                  $400
Unnamed: 3                                                15,079
Unnamed: 4                                                   46%
Unnamed: 5                               Consumer Initiated Contact
Unnamed: 6                                          Wire Transfer
Name: 5, dtype: object
```

- Na categoria de fraudes por valor perdido, exibimos e analisamos as linhas 5, 6 e 9 que esta voltando as fraudes para o setor
relacionado à investimentos, oportunidades de emprego e serviços por internet para verificar o total e média de perdas, alem da quantidade de reportagens e métodos de pagamento.

FATEC
Faculdade de Tecnologia

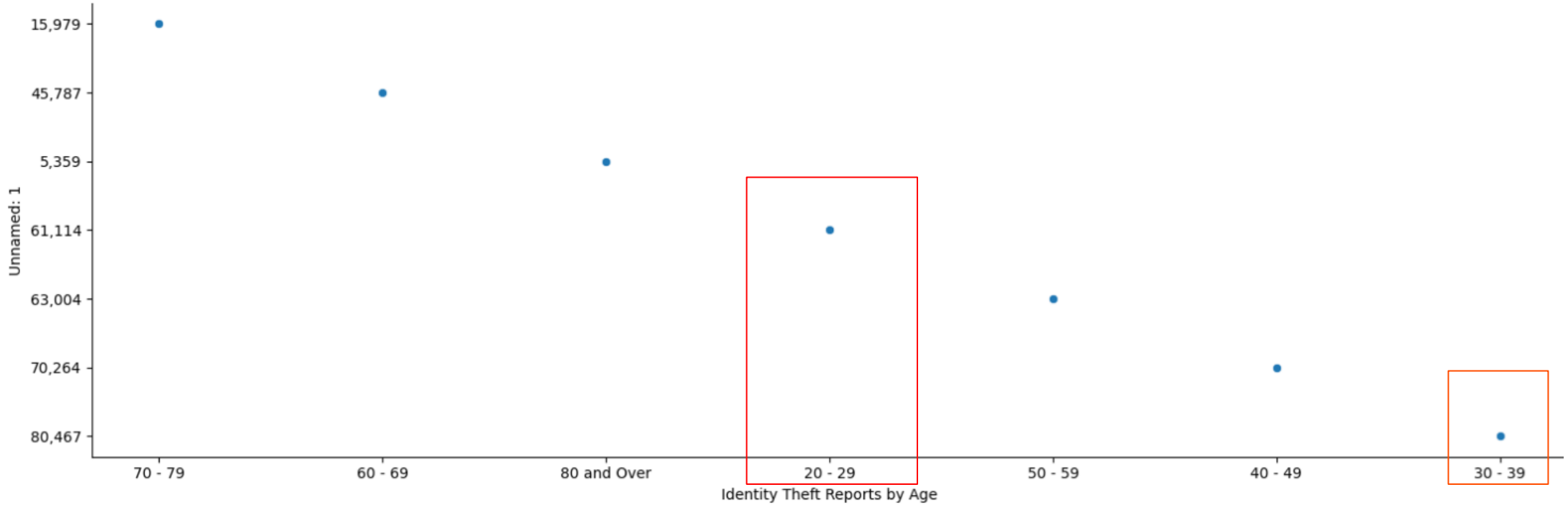
Analise de documentoscopia

•Por fim, analisamos o banco de dados de contagem de realatórios desde 2001 até 2017 para analisar a quantidade do crescimento de reportagens de fraude por ano, para verificar a probabilidade de acertividade de nosso proposito dentro da analise de documentoscopia a longo prazo.

Grafico de fraude por faixa etaria

```
fr = pd.read_csv('id.csv')
fr_ordenado = fr.sort_values(by='Unnamed: 1', ascending=True)
sns.relplot(x ="Identity Theft Reports by Age", y="Unnamed: 1", data=fr_ordenado.iloc[2:17], height = 5, aspect = 3)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fe267c7aa60>
```



FATEC
Faculdade de Tecnologia

Gráfico de quantidade de fraudes por tipo

```
import pandas as pd
t= pd.read_csv('teste.csv', usecols = ['Fraud Categories by Total Amount Lost','Unnamed: 1'])
sns.relplot(x ="Fraud Categories by Total Amount Lost", y="Unnamed: 1", data=df_ordenado.iloc[6:10], height = 2, aspect = 6)
```

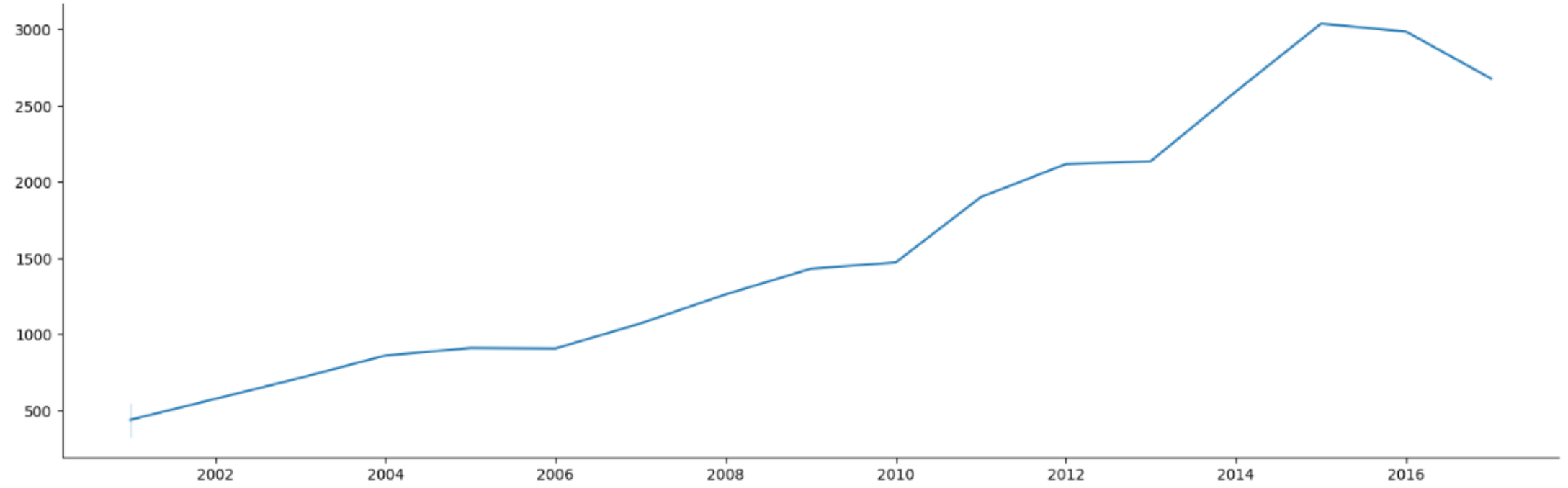
```
<seaborn.axisgrid.FacetGrid at 0x7f7d70eb5430>
```

FATEC
Faculdade de Tecnologia

Gráfico quantidade de fraude por ano

```
gf_ordenado = gf.sort_values(by='Unnamed: 1', ascending=False)
sns.relplot(x ="Number of Fraud, Identity Theft and Other Reports by Year\n", y="Unnamed: 1", data=gf_ordenado.iloc[2:17], height = 5, aspect = 3)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fe2680698b0>
```



FATEC
Faculdade de Tecnologia

Análise de fraude por idade

Contudo, temos que o índice de fraude de identidade é um dos maiores no ranking de tipos de fraude, correspondente a 12% em relação ao total de fraudes.

A faixa etária de 18 a 59 anos correspondem a 67% das vítimas por fraudes.
Levando em consideração os tipos de fraudes, temos uma média de 137,3.

Para identificarmos a extremidade dos números de fraudes temos o desvio padrão aproximado de 291,8.

Média	137,2787679
Erro padrão	39,3508654
Mediana	9,554
Desvio padrão	294,4749124
Variância da amostr	86715,47405
Curtose	2,691822147
Assimetria	2,060664182
Intervalo	994,701
Mínimo	1,299
Máximo	996
Soma	7687,611
Contagem	56

FATEC
Faculdade de Tecnologia

Análise de fraude por ano

Anualmente o índice de fraudes por identidade sobe consideravelmente, tendo em 2001 325,52 casos, e subindo para 2.675.611 em 2017, com uma média anual de 711,1 casos.

E identificando a extremidade dos números de fraudes temos o desvio padrão aproximado de 213,5.

Média	711,10
Variância	45620,5
Desvio Padrão	213,5896

Levantamento de Dados

- **Fonte:** comunidade on-line de cientistas de dados e praticantes de aprendizado de máquina – Kaggle
- **Pesquisa:** Consumer sentinel data book 2017 data

•https://www.kaggle.com/