Progress Tugas Akhir KASDD

- Muhammad Irza Arrizkyputra (1906353744)
- Alifah Azka Nisrina (1906353662)
- Andrew Nehemia H (1906400311)
- Bonifasius Erlangga (1906302850)

Contents of this template

You can delete this slide when you're done editing the presentation

Fonts	To view this template correctly in PowerPoint, download and install the fonts we used
Used and alternative resources	An assortment of graphic resources that are suitable for use in this presentation
Thanks slide	You must keep it so that proper credits for our design are given
Colors	All the colors used in this presentation
Icons and infographic resources	These can be used in the template, and their size and color can be edited
Editable presentation theme	You can edit the master slides easily. For more info, click **here**

Table of contents

01
Problem vs. solution
You can describe the topic of the section here

02
Product
You can describe the topic of the section here

03
Market & competition
You can describe the topic of the section here

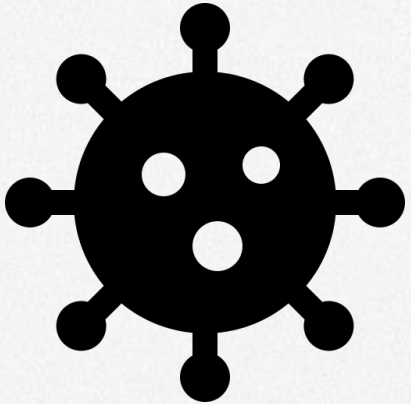
04
Business model
You can describe the topic of the section here

01

Dataset

Deteksi Virus

Sebuah dataset berisi spesifikasi hardware dan software dari sebuah mesin dan meminta kita untuk memprediksi berapa peluang mesin tersebut terinfeksi virus.



02

Exploratory Data Analysis

06

Data Info

#	Column	Non-Null Count	Dtype
0	IdDefaultBrowser	4613 non-null	float64
1	IdSettingAntivirus	134400 non-null	float64
2	BanyakAntivirus	134400 non-null	float64
3	IdNegaraPembuat	149668 non-null	int64
4	IdKotaPembuat	144598 non-null	float64
5	IdOrganisasiPembuat	105521 non-null	float64
6	IdLokasiGeografisMesinSaatIni	149646 non-null	float64
7	Platform	149668 non-null	object
8	Processor	149668 non-null	object
9	OsSuite	149668 non-null	int64
10	OsPlatformSubRelease	149668 non-null	object
11	VersiInternetExplorer	148834 non-null	float64
12	SmartScreenSetting	87660 non-null	object
13	DeviceType	149668 non-null	object
14	IdOEM	148392 non-null	float64
15	IdModelOEM	148310 non-null	float64
16	BanyakCoreProcessor	148638 non-null	float64
17	IdPembuatProcessor	148638 non-null	float64
18	IdModelProcessor	148636 non-null	float64
19	KapasitasDiskMemory	149231 non-null	float64
20	TipeDiskUtama	144981 non-null	object
21	KapasitasVolumeSistem	149231 non-null	float64
22	KapasitasRAM	147415 non-null	float64
23	TipeChassis	149430 non-null	object

Data Info

```
 24  UkuranDiagonalLayar          143543 non-null  float64
 25  UkuranHorisontalLayar        143547 non-null  float64
 26  UkuranVertikalLayar          143547 non-null  float64
 27  TipeBateraiInternal          54931 non-null   object
 28  VersiOS                      149668 non-null  object
 29  ArsitekturOS                 149668 non-null  object
 30  BranchOS                     149668 non-null  object
 31  BuildOS                      149668 non-null  int64
 32  RevisiBuildOS                149668 non-null  int64
 33  EdisiOS                      149668 non-null  object
 34  SkuNameOS                    149668 non-null  object
 35  TipeInstallasiOS             149668 non-null  object
 36  AutoUpdateSetting            149668 non-null  object
 37  IsOSGenuine                  149668 non-null  object
 38  IdPembuatFirmware            145604 non-null  float64
 39  IdVersiFirmware              145763 non-null  float64
 40  IsSecureBootEnabled          149668 non-null  int64
 41  IsTouchScreen                149668 non-null  int64
 42  IsGamer                      148978 non-null  float64
 43  infected_proba               149668 non-null  float64
dtypes: float64(22), int64(6), object(16)
memory usage: 50.2+ MB
```

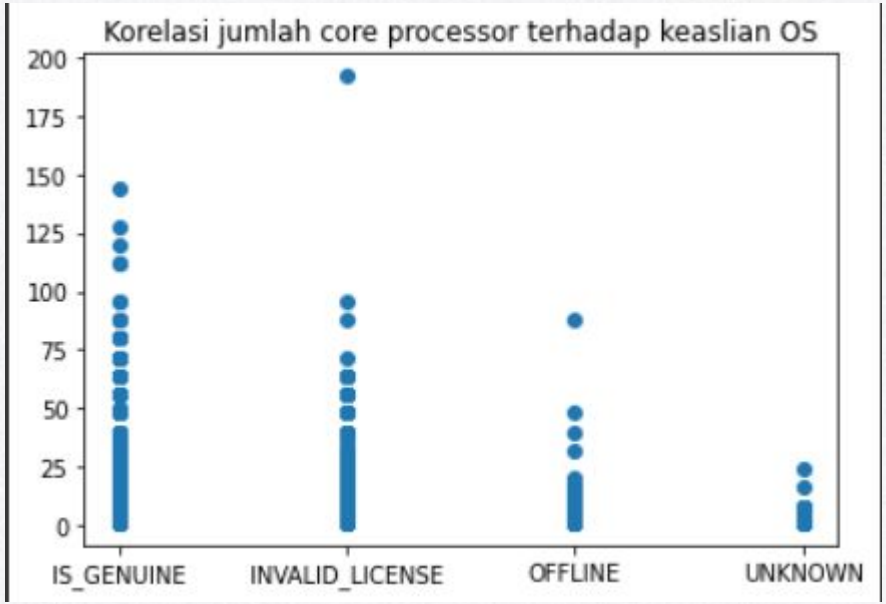
Data Describe

VersiInternetExplorer	IdOEM	...	UkuranHorisontalLayar	UkuranVertikalLayar	BuildOS	RevisiBuildOS	IdPembuatFirmware	IdVersiFirmware	IsSecureBootEna
148834.000000	148392.000000	...	143547.000000	143547.000000	149668.000000	149668.000000	145604.000000	145763.000000	149668.00
240.013028	2381.251516	...	1491.135419	874.431789	13545.320068	814.413188	420.462810	32302.995945	0.38
107.729477	1382.926076	...	329.716481	180.433159	2644.374069	1969.682865	231.248797	21082.281356	0.48
2.000000	74.000000	...	-1.000000	-1.000000	7601.000000	0.000000	13.000000	5.000000	0.00
117.000000	1443.000000	...	1366.000000	768.000000	10586.000000	165.000000	142.000000	13871.000000	0.00
323.000000	2206.000000	...	1366.000000	768.000000	14393.000000	321.000000	500.000000	33026.000000	0.00
333.000000	3035.000000	...	1600.000000	1024.000000	16299.000000	693.000000	556.000000	52402.000000	1.00
429.000000	6143.000000	...	7680.000000	3840.000000	18237.000000	41736.000000	1087.000000	72091.000000	1.00

Data Describe

HorisontalLayar	UkuranVertikalLayar	BuildOS	RevisiBuildOS	IdPembuatFirmware	IdVersiFirmware	IsSecureBootEnabled	IsTouchScreen	IsGamer	infected_proba
143547.000000	143547.000000	149668.000000	149668.000000	145604.000000	145763.000000	149668.000000	149668.000000	148978.000000	149668.000000
1491.135419	874.431789	13545.320068	814.413188	420.462810	32302.995945	0.384090	0.097102	0.101726	0.487458
329.716481	180.433159	2644.374069	1969.682865	231.248797	21082.281356	0.486381	0.296097	0.302289	0.356851
-1.000000	-1.000000	7601.000000	0.000000	13.000000	5.000000	0.000000	0.000000	0.000000	0.005000
1366.000000	768.000000	10586.000000	165.000000	142.000000	13871.000000	0.000000	0.000000	0.000000	0.145000
1366.000000	768.000000	14393.000000	321.000000	500.000000	33026.000000	0.000000	0.000000	0.000000	0.481000
1600.000000	1024.000000	16299.000000	693.000000	556.000000	52402.000000	1.000000	0.000000	0.000000	0.819000
7680.000000	3840.000000	18237.000000	41736.000000	1087.000000	72091.000000	1.000000	1.000000	1.000000	0.995000

Korelasi Variabel



Dalam visualisasi tersebut, didapatkan bahwa semakin banyak jumlah core processor, semakin tinggi kemungkinan OS nya asli

03

Pre-processing

Duplicated Data

```
[ ]  dv_cleaned = dv.copy()
```

```
[ ]  dv_cleaned.duplicated(keep=False).sum()
```

```
101
```

```
[ ]  # dropping duplicate values
     dv_cleaned = dv_cleaned.drop_duplicates()

     dv_cleaned.duplicated(keep=False).sum()
```

```
0
```

Duplicate data

Karena penghapusan pada data-data yang duplikat. Terdapat 101 baris duplikat berhasil di*remove*.

ArsitekturOS	0
AutoUpdateSetting	0
BanyakAntivirus	94
BanyakCoreProcessor	16220
BranchOS	0
BuildOS	0
DeviceType	0
EdisiOS	0
IdDefaultBrowser	0
IdKotaPembuat	0
IdLokasiGeografisMesinSaatIni	0
IdModelOEM	1592
IdModelProcessor	10582
IdNegaraPembuat	0
IdOEM	1233
IdOrganisasiPembuat	2210
IdPembuatFirmware	0
IdPembuatProcessor	18765
IdSettingAntivirus	0
IdVersiFirmware	0
IsGamer	15135
IsOSGenuine	0
IsSecureBootEnabled	0
IsTouchScreen	14531
KapasitasDiskMemory	3183
KapasitasRAM	11677
KapasitasVolumeSistem	2073
OsPlatformSubRelease	0
OsSuite	0
Platform	0
Processor	0
RevisiBuildOS	20404
SkuNameOS	0
SmartScreenSetting	0
TipeBateraiInternal	0
TipeChassis	0
TipeDiskUtama	0
TipeInstallasiOS	0
UkuranDiagonalLayar	4267
UkuranHorisontalLayar	3453
UkuranVertikalLayar	2409
VersiInternetExplorer	0
VersiOS	0
infected_proba	0

Outliers

Terdapat beberapa outliers terdeteksi dari dataset. Namun ada dua fitur/kolom yang tidak dianggap outlier yaitu IsGamer dan IsTouchscreen karena keduanya hanya memiliki dua nilai 0 dan 1

```
[ ] outlier_to_upper_lower(dv_cleaned,"BanyakCoreProcessor")
    outlier saat ini ada sebanyak 0

[ ] outlier_to_upper_lower(dv_cleaned,"RevisiBuildOS")
    outlier saat ini ada sebanyak 0
```

```
[ ] outlier_to_upper_lower(dv_cleaned,"UkuranHorisontalLayar")
    outlier saat ini ada sebanyak 0

[ ] outlier_to_upper_lower(dv_cleaned,"UkuranVertikalLayar")
    outlier saat ini ada sebanyak 0
```

Outliers

Pada beberapa fitur/kolom kami sudah menerapkan data upper/lower untuk menghilangkan outliernya.

```
[ ] outlier_to_upper_lower(dv_cleaned,"BanyakAntivirus")
    outlier saat ini ada sebanyak 0

[ ] outlier_to_upper_lower(dv_cleaned,"KapasitasDiskMemory")
    outlier saat ini ada sebanyak 0

[ ] outlier_to_upper_lower(dv_cleaned,"KapasitasRAM")
    outlier saat ini ada sebanyak 0

[ ] outlier_to_upper_lower(dv_cleaned,"KapasitasVolumeSistem")
    outlier saat ini ada sebanyak 0

[ ] outlier_to_upper_lower(dv_cleaned,"UkuranDiagonalLayar")
    outlier saat ini ada sebanyak 0
```

IdLokasiGeografisMesinSaatIni	0
IdModelOEM	1592
IdModelProcessor	10582
IdNegaraPembuat	0
IdOEM	1233
IdOrganisasiPembuat	2210
IdPembuatFirmware	0
IdPembuatProcessor	18765
IdSettingAntivirus	0
IdVersiFirmware	0

Outliers

Disini terdapat outlier pada atribut yang sifatnya Id, namun tidak kami tangani. Alasannya karena kami menganggap Id merupakan suatu “nama” yang sudah tidak berbentuk data kategorikal. Jadi, untuk sementara kami biarkan terlebih dahulu.

	Total	Percent
IdDefaultBrowser	144992	0.969165
TipeBateraiInternal	94699	0.632994
SmartScreenSetting	61987	0.414338
IdOrganisasiPembuat	44123	0.294930
BanyakAntivirus	15235	0.101835
IdSettingAntivirus	15235	0.101835
UkuranDiagonalLayar	6120	0.040908
UkuranVertikalLayar	6116	0.040881
UkuranHorisontalLayar	6116	0.040881
IdKotaPembuat	5067	0.033869
TipeDiskUtama	4687	0.031329
IdPembuatFirmware	4061	0.027145
IdVersiFirmware	3902	0.026082
KapasitasRAM	2249	0.015033
IdModelOEM	1358	0.009077
IdOEM	1276	0.008529
IdModelProcessor	1032	0.006898
BanyakCoreProcessor	1030	0.006885
IdPembuatProcessor	1030	0.006885
VersiInternetExplorer	834	0.005575
IsGamer	690	0.004612
KapasitasDiskMemory	437	0.002921
KapasitasVolumeSistem	437	0.002921
TipeChassis	238	0.001591
IdLokasiGeografisMesinSaatIni	22	0.000147

Null Value

Terdapat banyak null value dari tiap kolom/fitur. Pada fitur/kolom IdDefaultBrowser akan di drop karena hanya sebagai identifikasi dan null nya sangat besar yaitu 96%. Selain kolom itu, semuanya dilakukan method fillna() berdasarkan median untuk data numerik, dan 0 untuk data kategorikal

Introduction

You can give a brief description of the topic you want to talk about here. If you want to talk about Mercury, you can say that it’s the smallest planet in the Solar System

```
Data columns (total 41 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   BanyakAntivirus                 149605 non-null  float64
 1   IdNegaraPembuat                 149605 non-null  int64
 2   IdKotaPembuat                   149605 non-null  float64
 3   IdLokasiGeografisMesinSaatIni   149605 non-null  float64
 4   Platform                        149605 non-null  object
 5   Processor                       149605 non-null  object
 6   OsSuite                         149605 non-null  int64
 7   OsPlatformSubRelease            149605 non-null  object
 8   VersiInternetExplorer           149605 non-null  float64
 9   SmartScreenSetting              149605 non-null  object
 10  DeviceType                      149605 non-null  object
 11  IdOEM                           149605 non-null  float64
 12  IdModelOEM                      149605 non-null  float64
 13  BanyakCoreProcessor             149605 non-null  float64
 14  IdPembuatProcessor              149605 non-null  float64
 15  IdModelProcessor                149605 non-null  float64
 16  KapasitasDiskMemory             149605 non-null  float64
 17  TipeDiskUtama                   149605 non-null  object
 18  KapasitasVolumeSistem           149605 non-null  float64
 19  KapasitasRAM                    149605 non-null  float64
 20  TipeChassis                     149605 non-null  object
 21  UkuranDiagonalLayar             149605 non-null  float64
 22  UkuranHorisontalLayar           149605 non-null  float64
 23  UkuranVertikalLayar             149605 non-null  float64
 24  TipeBateraiInternal             149605 non-null  object
 25  VersiOS                         149605 non-null  object
 26  ArsitekturOS                    149605 non-null  object
 27  BranchOS                        149605 non-null  object
 28  BuildOS                         149605 non-null  int64
 29  RevisiBuildOS                   149605 non-null  int64
```

Encoding

Terdapat beberapa data yang sifatnya data kategorikal. Pada bagian ini, kami ingin melakukan encoding pada data-data kategorikal tersebut.

```
 30  EdisiOS                         149605 non-null  object
 31  SkuNameOS                       149605 non-null  object
 32  TipeInstallasiOS                149605 non-null  object
 33  AutoUpdateSetting               149605 non-null  object
 34  IsOSGenuine                     149605 non-null  object
 35  IdPembuatFirmware               149605 non-null  float64
 36  IdVersiFirmware                 149605 non-null  float64
 37  IsSecureBootEnabled             149605 non-null  int64
 38  IsTouchScreen                   149605 non-null  int64
 39  IsGamer                         149605 non-null  float64
 40  infected_proba                  149605 non-null  float64
dtypes: float64(19), int64(6), object(16)
memory usage: 47.9+ MB
```

```
dv_cleaned['Platform'].value_counts()
```

```
windows10      47820
windows7       44396
windows8       43051
windows2016    14338
Name: Platform, dtype: int64
```

```
from sklearn.preprocessing import OneHotEncoder

encoder = OneHotEncoder(sparse=False)
encoder = encoder.fit_transform(dv_cleaned[['Platform']])
dv_cleaned_platform = pd.DataFrame(encoder)
```

```
dv_cleaned_platform.value_counts()
```

```
0    1    2    3
1.0  0.0  0.0  0.0    47820
0.0  0.0  1.0  0.0    44396
          0.0  1.0    43051
     1.0  0.0  0.0    14338
dtype: int64
```

Encoding

Disini kami melakukan OneHotEncoder pada atribut Platform. Alasan kami menggunakan jenis encode tersebut adalah karena value yang ada pada platform tidak perlu diurutkan dan valuenya juga tidak banyak, jadi sepertinya tidak terlalu masalah apabila kita melakukan One Hot yang cara encodingnya membuat kolom baru

```
dv_cleaned_platform.rename(columns = {0:'Windows 10', 1:'Windows 7', 2:'Windows 8', 3:'Windows 2016'}, inplace = True)

dv_cleaned_platform
```

```
[934] dv_cleaned = dv_cleaned.join(dv_cleaned_platform)
      dv_cleaned
```

```
dv_cleaned.drop('Platform', inplace=True, axis=1)
dv_cleaned
```

Encoding

Setelah dilakukan One Hot Encoding, kami melakukan join supaya dapat terhubung menjadi 1 dataframe, kemudian membuang kolom yang lama yang isinya masih berupa data kategorikal

Encoding

```
[934] dv_cleaned = dv_cleaned.join(dv_cleaned_platform)
      dv_cleaned
```

```
dv_cleaned.drop('Platform', inplace=True, axis=1)
dv_cleaned
```

Setelah dilakukan One Hot Encoding, kami melakukan join supaya dapat terhubung menjadi 1 dataframe, kemudian membuang kolom yang lama yang isinya masih berupa data kategorikal

Cara yang sama kami lakukan pada atribut Processor. Ada pada slide selanjutnya

Encoding

Processor Encoding

```
[936] encoder = OneHotEncoder(sparse=False)
      encoder = encoder.fit_transform(dv_cleaned[['Processor']])
      dv_cleaned_processor = pd.DataFrame(encoder)
```

```
[937] print(dv_cleaned_processor.value_counts())
      print(dv_cleaned['Processor'].value_counts())
```

```
0    1    2
0.0  1.0  0.0    129639
     0.0  1.0     19963
1.0  0.0  0.0         3
dtype: int64
x64      129639
x86       19963
arm64         3
Name: Processor, dtype: int64
```

```
dv_cleaned_processor.rename(columns = {0:'x64', 1:'x86', 2:'arm64'}, inplace = True)
dv_cleaned = dv_cleaned.join(dv_cleaned_processor)
dv_cleaned.drop('Processor', inplace=True, axis=1)
dv_cleaned
```

›Release	VersiInternetExplorer	SmartScreenSetting	DeviceType	IdOEM	...	IsTouchScreen	IsGamer	infected_proba	Windows 10	Windows 7	Windows 8	Windows 2016	x64	x86	arm64
th2	85.0	RequireAdmin	Notebook	2102.0	...	0	0.0	0.626	1.0	0.0	0.0	0.0	0.0	1.0	0.0
prers5	163.0	RequireAdmin	Notebook	1443.0	...	1	1.0	0.995	1.0	0.0	0.0	0.0	0.0	1.0	0.0
rs3	135.0	RequireAdmin	Notebook	2206.0	...	0	0.0	0.937	1.0	0.0	0.0	0.0	0.0	1.0	0.0
rs2	108.0	RequireAdmin	Notebook	3799.0	...	0	0.0	0.661	1.0	0.0	0.0	0.0	0.0	0.0	1.0

Encoding

- Terdapat juga beberapa atribut yang kami lakukan encode menggunakan LabelEncoder. Alasannya adalah karena kami melihat value yang dimiliki pada atribut tersebut sangat banyak, jadi apabila menggunakan OneHotEncoder akan membuat dataframe tersebut menambah banyak sekali kolom baru hasil encoding

```
[942] from sklearn.preprocessing import LabelEncoder

      # TipeBateraiInternal
      labelencoder = LabelEncoder()
      # Assigning numerical values and storing in another column
      dv_cleaned['TipeBateraiInternal_encode'] = labelencoder.fit_transform(dv_cleaned['TipeBateraiInternal'])
```

```
dv_cleaned['EdisiOS'].unique()
labelencoder = LabelEncoder()
# Assigning numerical values and storing in another column
dv_cleaned['EdisiOS_encode'] = labelencoder.fit_transform(dv_cleaned['EdisiOS'])
```

Encoding


```
dv_cleaned['SkuNameOS'].unique()
dv_cleaned['SkuNameOS_encode'] = labelencoder.fit_transform(dv_cleaned['SkuNameOS'])
```

```
dv_cleaned.drop('SkuNameOS', inplace=True, axis=1)
```

```
dv_cleaned['TipeInstallasiOS'].unique()
dv_cleaned['TipeInstallasiOS_encode'] = labelencoder.fit_transform(dv_cleaned['TipeInstallasiOS'])
```

```
dv_cleaned.drop('TipeInstallasiOS', inplace=True, axis=1)
```

Encoding


```
[ ]  dv_cleaned['AutoUpdateSetting'].value_counts()
```

```
FullAuto                                95101
Notify                                  21680
UNKNOWN                                 16235
DownloadNotify                          14338
AutoInstallAndRebootAtMaintenanceTime    2096
Off                                       155
Name: AutoUpdateSetting, dtype: int64
```

```
[ ]  dv_cleaned['AutoUpdateSetting'].unique()
     dv_cleaned['AutoUpdateSetting_encode'] = labelencoder.fit_transform(dv_cleaned['AutoUpdateSetting'])
```

```
[ ]  print(dv_cleaned['AutoUpdateSetting'].value_counts())
     print(dv_cleaned['AutoUpdateSetting_encode'].value_counts())
```

```
FullAuto                                95101
Notify                                  21680
UNKNOWN                                 16235
DownloadNotify                          14338
AutoInstallAndRebootAtMaintenanceTime    2096
Off                                       155
Name: AutoUpdateSetting, dtype: int64
2    95101
3    21680
5    16235
1    14338
0     2096
4      155
Name: AutoUpdateSetting_encode, dtype: int64
```

```
[ ]  dv_cleaned.drop('AutoUpdateSetting', inplace=True, axis=1)
```

Encoding

IsOSGenuine

```
[ ] dv_cleaned['IsOSGenuine'].value_counts()
```

```
IS_GENUINE         134941
INVALID_LICENSE     10039
OFFLINE              4382
UNKNOWN               243
Name: IsOSGenuine, dtype: int64
```

```
[ ] encoder = OneHotEncoder(sparse=False)
    encoder = encoder.fit_transform(dv_cleaned[['IsOSGenuine']])
    dv_cleaned_genuine = pd.DataFrame(encoder)
```

```
    print(dv_cleaned_genuine.value_counts())
    print(dv_cleaned['IsOSGenuine'].value_counts())
```

```
0    1    2    3
0.0  1.0  0.0  0.0    134941
1.0  0.0  0.0  0.0     10039
0.0  0.0  1.0  0.0      4382
          0.0  1.0       243
dtype: int64
IS_GENUINE         134941
INVALID_LICENSE     10039
OFFLINE              4382
UNKNOWN               243
Name: IsOSGenuine, dtype: int64
```

```
[ ] dv_cleaned_genuine.rename(columns = {0:'IS_GENUINE', 1:'INVALID_LICENSE', 2:'OFFLINE', 3:'UNKNOWN'}, inplace = True)
    dv_cleaned = dv_cleaned.join(dv_cleaned_genuine)
    dv_cleaned.drop('IsOSGenuine', inplace=True, axis=1)
    dv_cleaned
```

Encoding


```
[1182] dv_cleaned['VersiOS'].unique()
       dv_cleaned['VersiOS_encode'] = labelencoder.fit_transform(dv_cleaned['VersiOS'])

       print(dv_cleaned['VersiOS'].value_counts())
       print(dv_cleaned['VersiOS_encode'].value_counts())

       10.0.10586.318     10938
       10.0.17134.228      8836
       10.0.10586.164      7175
       10.0.17134.165      6185
       10.0.10586.494      5181
                          ...
       10.0.16294.1           1
       10.0.17046.1000        1
       10.0.16281.1000        1
       10.0.16288.1           1
       10.0.14393.2311        1
       Name: VersiOS, Length: 308, dtype: int64
       60     10938
       265     8836
       55      7175
       262     6185
       66      5181
              ...
       225        1
       255        1
       223        1
       224        1
       130        1
       Name: VersiOS_encode, Length: 308, dtype: int64

[1184] dv_cleaned.drop('VersiOS', inplace=True, axis=1)
```

Encoding


```
[1185] encoder = OneHotEncoder(sparse=False)
       encoder = encoder.fit_transform(dv_cleaned[['ArsitekturOS']])
       dv_cleaned_ArsitekturOS = pd.DataFrame(encoder)
```

```
[1186] print(dv_cleaned_ArsitekturOS.value_counts())
       print(dv_cleaned['ArsitekturOS'].value_counts())
```

```
0    1    2
1.0  0.0  0.0    129500
0.0  0.0  1.0     20102
     1.0  0.0         3
dtype: int64
amd64    129500
x86       20102
arm64         3
Name: ArsitekturOS, dtype: int64
```

```
dv_cleaned_ArsitekturOS.rename(columns = {0:'Arsitektur amd64', 1:'Arsitektur x86', 2:'Arsitektur arm64'}, inplace = True)
dv_cleaned = dv_cleaned.join(dv_cleaned_ArsitekturOS)
dv_cleaned.drop('ArsitekturOS', inplace=True, axis=1)
dv_cleaned
```

ceType	IdOEM	...	TipeInstallasiOS_encode	AutoUpdateSetting_encode	IS_GENUINE	INVALID_LICENSE	OFFLINE	UNKNOWN	VersiOS_encode	Arsitektur amd64	Arsitektur x86	Arsitektur arm64
otebook	2102.0	...	7	5	0.0	1.0	0.0	0.0	71	1.0	0.0	0.0
otebook	1443.0	...	6	2	0.0	1.0	0.0	0.0	281	1.0	0.0	0.0
otebook	2206.0	...	4	2	0.0	1.0	0.0	0.0	227	1.0	0.0	0.0
otebook	3799.0	...	8	3	0.0	1.0	0.0	0.0	203	0.0	0.0	1.0
otebook	525.0	...	6	0	0.0	1.0	0.0	0.0	247	1.0	0.0	0.0

0 d selesai pada 23.18

Encoding


```
[1185] encoder = OneHotEncoder(sparse=False)
       encoder = encoder.fit_transform(dv_cleaned[['ArsitekturOS']])
       dv_cleaned_ArsitekturOS = pd.DataFrame(encoder)
```

```
[1186] print(dv_cleaned_ArsitekturOS.value_counts())
       print(dv_cleaned['ArsitekturOS'].value_counts())
```

```
0    1    2
1.0  0.0  0.0    129500
0.0  0.0  1.0     20102
     1.0  0.0         3
dtype: int64
amd64    129500
x86       20102
arm64         3
Name: ArsitekturOS, dtype: int64
```

```
dv_cleaned_ArsitekturOS.rename(columns = {0:'Arsitektur amd64', 1:'Arsitektur x86', 2:'Arsitektur arm64'}, inplace = True)
dv_cleaned = dv_cleaned.join(dv_cleaned_ArsitekturOS)
dv_cleaned.drop('ArsitekturOS', inplace=True, axis=1)
dv_cleaned
```

ceType	IdOEM	...	TipeInstallasiOS_encode	AutoUpdateSetting_encode	IS_GENUINE	INVALID_LICENSE	OFFLINE	UNKNOWN	VersiOS_encode	Arsitektur amd64	Arsitektur x86	Arsitektur arm64
otebook	2102.0	...	7	5	0.0	1.0	0.0	0.0	71	1.0	0.0	0.0
otebook	1443.0	...	6	2	0.0	1.0	0.0	0.0	281	1.0	0.0	0.0
otebook	2206.0	...	4	2	0.0	1.0	0.0	0.0	227	1.0	0.0	0.0
otebook	3799.0	...	8	3	0.0	1.0	0.0	0.0	203	0.0	0.0	1.0
otebook	525.0	...	6	0	0.0	1.0	0.0	0.0	247	1.0	0.0	0.0

0 d selesai pada 23.18

Encoding


```
dv_cleaned['BranchOS_encode'] = labelencoder.fit_transform(dv_cleaned['BranchOS'])
print(dv_cleaned['BranchOS'].value_counts())
print(dv_cleaned['BranchOS_encode'].value_counts())
```

```
th2_release                 55541
rs1_release                 36083
rs4_release                 25160
rs2_release                  9715
rs3_release                  9509
rs3_release_svc_escrow       7354
th2_release_sec              4165
th1_st1                      1257
th1                           599
rs5_release                    96
rs_prerelease                  55
rs3_release_svc_escrow_im      38
rs_prerelease_flt              19
rs1_release_srvmedia            9
win7sp1_ldr_escrow              2
winblue_ltsb_escrow             2
win7sp1_ldr                     1
Name: BranchOS, dtype: int64
12    55541
0     36083
6     25160
2      9715
3      9509
4      7354
13     4165
11     1257
10      599
7        96
8        55
5        38
9        19
1         9
15        2
16        2
14        1
Name: BranchOS_encode, dtype: int64
```

```
[1189] dv_cleaned.drop('BranchOS', inplace=True, axis=1)
```

Encoding


```
dv_cleaned['TipeChassis_encode'] = labelencoder.fit_transform(dv_cleaned['TipeChassis'])
print(dv_cleaned['TipeChassis'].value_counts())
print(dv_cleaned['TipeChassis_encode'].value_counts())
```

```
Notebook                78534
Desktop                 31733
Laptop                  10218
Portable                 7583
Other                    5595
AllinOne                 3578
RackMountChassis         2577
MiniTower                2408
Tower                    1800
MainServerChassis        1307
LowProfileDesktop        1072
SpaceSaving               844
HandHeld                  608
UNKNOWN                   593
Convertible               480
Detachable                303
Unknown                    93
LunchBox                   87
Tablet                     76
MiniPC                     29
SubNotebook                26
Blade                      17
MultisystemChassis         16
SealedCasePC               11
ExpansionChassis            6
BusExpansionChassis         4
0                           3
25                          1
StickPC                     1
31                          1
BladeEnclosure              1
Name: TipeChassis, dtype: int64
19    78534
8     31733
12    10218
21     7583
20     5595
3      3578
22     2577
17     2408
28     1800
15     1307
13     1072
24      844
11      608
29      593
7       480
9       303
30       93
14       87
27       76
16       29
26       26
4        17
18       16
23       11
10        6
6         4
```

```
[1191] dv_cleaned.drop('TipeChassis', inplace=True, axis=1)
```

Encoding


```
[1192] encoder = OneHotEncoder(sparse=False)
       encoder = encoder.fit_transform(dv_cleaned[['TipeDiskUtama']])
       dv_cleaned_TipeDiskUtama = pd.DataFrame(encoder)
```

```
[1193] print(dv_cleaned_TipeDiskUtama.value_counts())
       print(dv_cleaned['TipeDiskUtama'].value_counts())
```

```
0    1    2    3
1.0  0.0  0.0  0.0    101729
0.0  1.0  0.0  0.0     23431
     0.0  1.0  0.0     12938
          0.0  1.0     11507
dtype: int64
HDD            101729
SSD             23431
UNKNOWN         12938
Unspecified     11507
Name: TipeDiskUtama, dtype: int64
```

```
dv_cleaned_TipeDiskUtama.rename(columns = {0:'Disk HDD', 1:'Disk SSD', 2:'Disk UNKNOWN', 3:'Disk Unspecified'}, inplace = True)
dv_cleaned = dv_cleaned.join(dv_cleaned_TipeDiskUtama)
dv_cleaned.drop('TipeDiskUtama', inplace=True, axis=1)
dv_cleaned
```

er	SmartScreenSetting	DeviceType	IdOEM	...	VersiOS_encode	Arsitektur amd64	Arsitektur x86	Arsitektur arm64	BranchOS_encode	TipeChassis_encode	Disk HDD	Disk SSD	Disk UNKNOWN	Disk Unspecified
.0	RequireAdmin	Notebook	2102.0	...	71	1.0	0.0	0.0	12	19	0.0	1.0	0.0	0.0
.0	RequireAdmin	Notebook	1443.0	...	281	1.0	0.0	0.0	9	19	1.0	0.0	0.0	0.0
.0	RequireAdmin	Notebook	2206.0	...	227	1.0	0.0	0.0	3	19	1.0	0.0	0.0	0.0
.0	RequireAdmin	Notebook	3799.0	...	203	0.0	0.0	1.0	2	19	1.0	0.0	0.0	0.0

Encoding


```
# SmartScreenSetting
dv_cleaned['SmartScreenSetting_encode'] = labelencoder.fit_transform(dv_cleaned['SmartScreenSetting'])
print(dv_cleaned['SmartScreenSetting'].value_counts())
print(dv_cleaned['SmartScreenSetting_encode'].value_counts())
```

```
RequireAdmin    120795
ExistsNotSet     15466
Off              10434
Prompt            1978
Warn               778
Block              118
off                 19
&#x01;               8
On                   4
&#x02;               3
on                   2
Name: SmartScreenSetting, dtype: int64
7     120795
3      15466
4      10434
6       1978
8        778
2        118
9         19
0          8
5          4
1          3
10         2
Name: SmartScreenSetting_encode, dtype: int64
```

```
[1201] dv_cleaned.drop('SmartScreenSetting', inplace=True, axis=1)
```

Encoding


```
# DeviceType
dv_cleaned['DeviceType_encode'] = labelencoder.fit_transform(dv_cleaned['DeviceType'])
print(dv_cleaned['DeviceType'].value_counts())
print(dv_cleaned['DeviceType_encode'].value_counts())
```

```
Notebook        90553
Desktop         32996
SmallServer      7608
AllInOne         5239
Convertible      4456
MediumServer     2905
Detachable       2677
LargeTablet       987
PCOther           923
LargeServer       865
SmallTablet       366
ServerOther        30
Name: DeviceType, dtype: int64
7     90553
2     32996
10     7608
0      5239
1      4456
6      2905
3      2677
5       987
8       923
4       865
11      366
9        30
Name: DeviceType_encode, dtype: int64
```

```
[1203] dv_cleaned.drop('DeviceType', inplace=True, axis=1)
```

Encoding


```
# OsPlatformSubRelease
dv_cleaned['OsPlatformSubRelease_encode'] = labelencoder.fit_transform(dv_cleaned['OsPlatformSubRelease'])
print(dv_cleaned['OsPlatformSubRelease'].value_counts())
print(dv_cleaned['OsPlatformSubRelease_encode'].value_counts())
```

```
windows7      44396
windows8.1    43051
rs4           21254
rs1           18438
rs3           14114
rs2            4402
th2            2289
th1            1532
prers5          129
Name: OsPlatformSubRelease, dtype: int64
7    44396
8    43051
4    21254
1    18438
3    14114
2     4402
6     2289
5     1532
0      129
Name: OsPlatformSubRelease_encode, dtype: int64
```

```
[1207] dv_cleaned.drop('OsPlatformSubRelease', inplace=True, axis=1)
```

Thanks!

Do you have any questions?
youremail@freepik.com
+91 620 421 838
yourwebsite.com