```
import pandas as pd
df=pd.read_csv("/content/data.csv")
```

```
/usr/local/lib/python3.7/dist-packages/IPython/core/interactiveshell.py:2718: DtypeWarning: Columns (183,378)
  interactivity=interactivity, compiler=compiler, result=result)
```

```
df.head()
```

	CHECKUP1	_INCOMG	_EDUCAG	_SMOKER3
0	Within past 2 years (1 year but less than 2 ye...	$50,000 or more	Graduated from College or Technical School	Never smoked
1	Within past year (anytime less than 12 months ...	Don't know/Not sure/Missing	Attended College or Technical School	Never smoked
2	Within past year (anytime less than 12 months ...	Don't know/Not sure/Missing	Graduated from College or Technical School	Never smoked
3	Within past year (anytime less	Don't know/Not	Graduated High School	Never

Getting all the unique values of checkup column

```
df.CHECKUP1.unique()
```

```
array(['Within past 2 years (1 year but less than 2 years ago)',
       'Within past year (anytime less than 12 months ago)',
       '5 or more years ago',
       'Within past 5 years (2 years but less than 5 years ago)',
       "Don't know/Not sure", 'Never', 'Refused'], dtype=object)
```

Removing all the values which have values Don't know/Not sure/Missing for Income column

```
df_filtered = df[df['_INCOMG'] !=  "Don't know/Not sure/Missing"]
```

Removing all the rows which have values Don't know/Refused/Missing for Smoker Column

```
df_filtered = df_filtered[df_filtered['_SMOKER3'] !=  "Don't know/Refused/Missing"]
```

Removing all the values which have values Don't know/Not sure/Missing for Education Column

```
df_filtered = df_filtered[df_filtered['_EDUCAG'] !=  "Don't know/Not sure/Missing"]
df_filtered.shape
```

```
    (9961, 4)
```

Removing all the values which have values Don't know/Not sure/Refused for Checkup Column

```
df_filtered = df_filtered[df_filtered['CHECKUP1'] !=  "Don't know/Not sure"  ]
df_filtered.shape
```

```
    (9931, 4)
```

```
df_filtered = df_filtered[df_filtered['CHECKUP1'] !=  "Refused"  ]
df_filtered.shape
```

```
    (9897, 4)
```

```
df_filtered.CHECKUP1.unique()
```

```
    array(['Within past 2 years (1 year but less than 2 years ago)',
           'Within past year (anytime less than 12 months ago)',
           '5 or more years ago',
           'Within past 5 years (2 years but less than 5 years ago)', 'Never'],
          dtype=object)
```

Making a copy of dataset

```
df_test=df_filtered
```

For checkup column dividing values into Recent and Not Recent for easier processing.

```
df_test['CHECKUP1']= df_test['CHECKUP1'].replace(['Within past 2 years (1 year but less than 2 years ago)',
       'Within past year (anytime less than 12 months ago)',
       '5 or more years ago',
       'Within past 5 years (2 years but less than 5 years ago)', 'Never'],['Not Recent','Recent','Not Rece
```

```
df_test.CHECKUP1.unique()
```

```
    array(['Not Recent', 'Recent'], dtype=object)
```

```
df_filtered.head()
```

	CHECKUP1	_INCOMG	_EDUCAG	_SMOKER3
0	Not Recent	$50,000 or more	Graduated from College or Technical School	Never smoked
4	Recent	$35,000 to less than $50,000	Graduated from College or Technical School	Never smoked
5	Recent	$50,000 or more	Graduated from College or Technical School	Current smoker - now smokes some days
	Not		Graduated from College or	

```
df_test._SMOKER3.unique()
```

```
    array(['Never smoked', 'Current smoker - now smokes some days',
           'Former smoker', 'Current smoker - now smokes every day'],
          dtype=object)
```

Convert smoker column values to 0 and 1 where 0 means doesn't smoke and 1 means they smoke so that machine learning algorthim can process it easily

```
df_test['_SMOKER3']= df_test['_SMOKER3'].replace(['Never smoked', 'Current smoker - now smokes some days',
       'Former smoker', 'Current smoker - now smokes every day'],['0','1','0','1'])
```

```
df_test._SMOKER3.unique()
```

```
    array(['0', '1'], dtype=object)
```

Taking final look at the datset we have preprocessed

```
df_test.head()
```

	CHECKUP1	_INCOMG	_EDUCAG	_SMOKER3
0	Not Recent	$50,000 or more	Graduated from College or Technical School	0
4	Recent	$35,000 to less than $50,000	Graduated from College or Technical School	0
5	Recent	$50,000 or more	Graduated from College or Technical School	1
6	Not Recent	$50,000 or more	Graduated from College or Technical School	0
7	Recent	$50,000 or more	Attended College or Technical School	1

Saving file to desk so that it can be used by weka for further work

```
df_test.to_csv('cleaned_checkup_edu_smoke.csv',index=False)
```