dataset

July 21, 2024

1

dataset	description	Type	Data Points
BACE (Classification)	Provides bindings results for a set of inhibitors of human beta-secretase (BACE-1)	SMILES	1513
BBBP	Blood-Brain Barrier Penetration designed for the modeling and prediction of barrier permeability	SMILES	2000
QM9	Dataset that provides geometric/energetic/electronic and thermodynamic properties for a subset of GDB-17 database	SMILES, 3D coordinates	134 000
HIV	A dataset wich tested the ability to inhibit HIV replication	SMILES	40 000
Tox21	The "Toxicology in the 21st Century" (Tox21) initiative created a public database measuring the toxicity of compounds	SMILES	8 000
MUV	Benchmark dataset selected from PubChem BioAssay by applying a refined nearest neighbor analysis	SMILES	9000
SIDER	The Side Effect Resource (SIDER) is a database of marketed drugs and adverse drug reactions (ADR)	SMILES	1 427

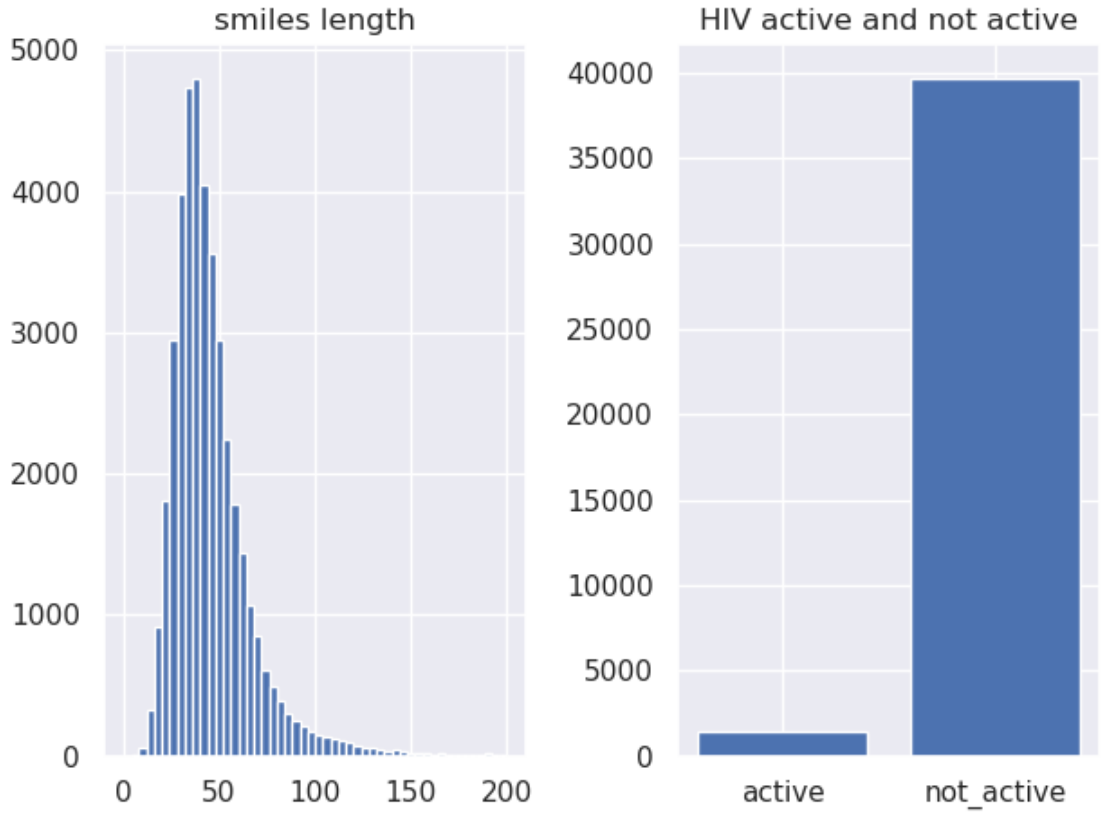
[1]:
```
#
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

1.1 HIV

HIV

[34]:
```
df = pd.read_csv('data/HIV.csv')
length = df.iloc[:,0].apply(lambda x: len(x))
df["length"] = length # smiles
plt.subplot(121)
_ = plt.hist(df["length"], bins=50, range=(0,200))
```

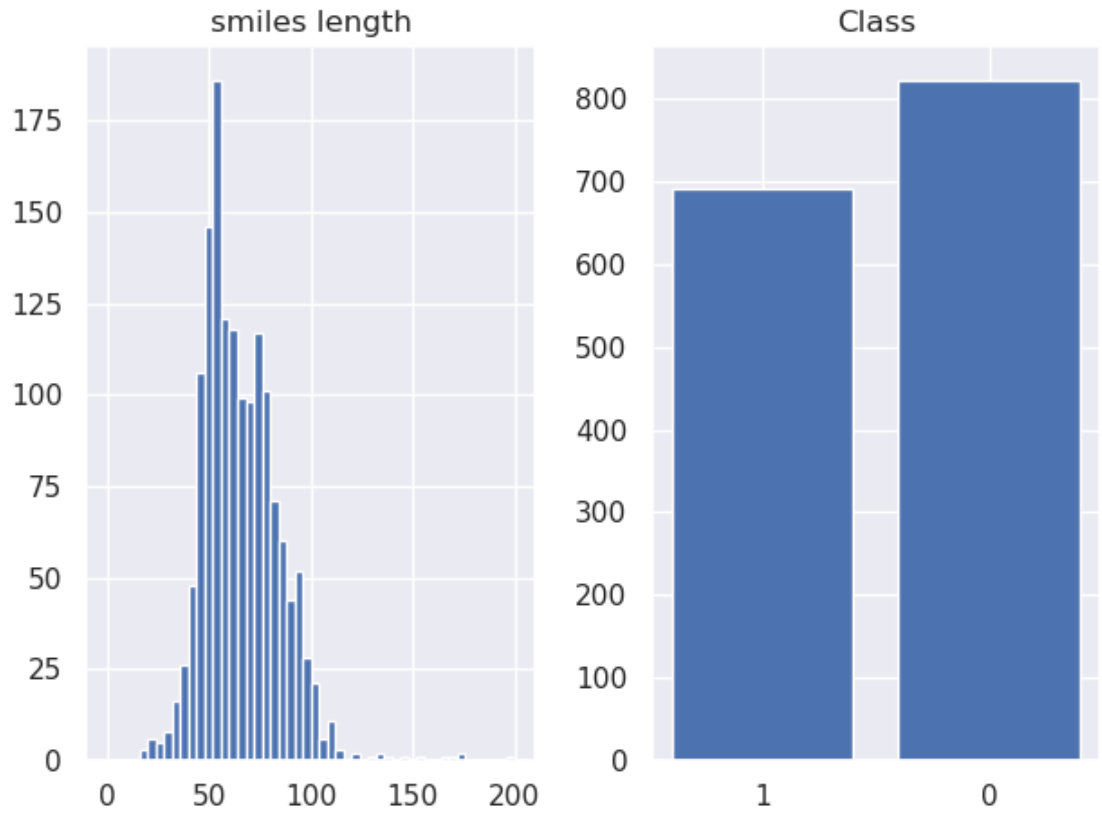
```
plt.title("smiles length")
plt.subplot(122)
plt.bar(["active","not_active"],[(df["HIV_active"]==1).
  ↪sum(),(df["HIV_active"]==0).sum()])
plt.title("HIV active and not active")
plt.tight_layout()
```



1.2 bace dataset

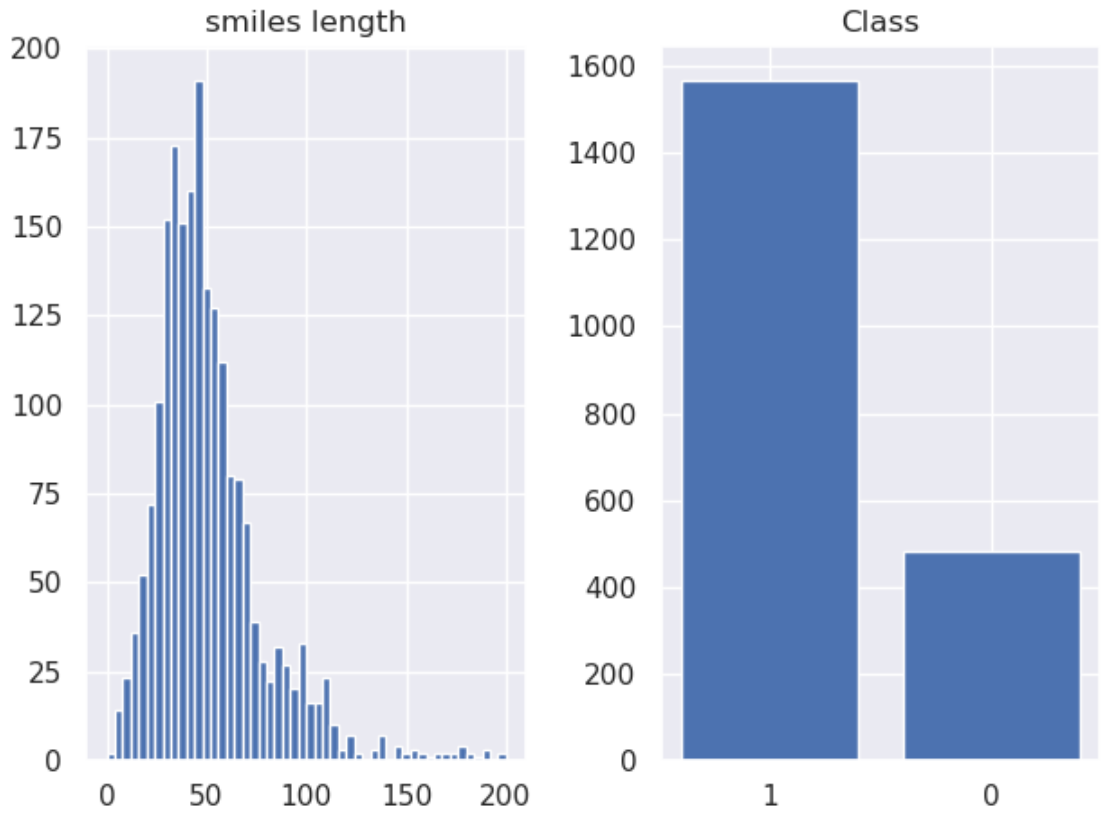
keys mol Class keys bace-key.md.

[50]:
```
df = pd.read_csv('data/bace.csv')
length = df["mol"].apply(lambda x: len(x))
df["length"] = length # smiles
plt.subplot(121)
_ = plt.hist(df["length"], bins=50, range=(0,200))
plt.title("smiles length")
plt.subplot(122)
plt.bar(["1","0"],[(df["Class"]==1).sum(),(df["Class"]==0).sum()])
plt.title("Class")
plt.tight_layout()
```



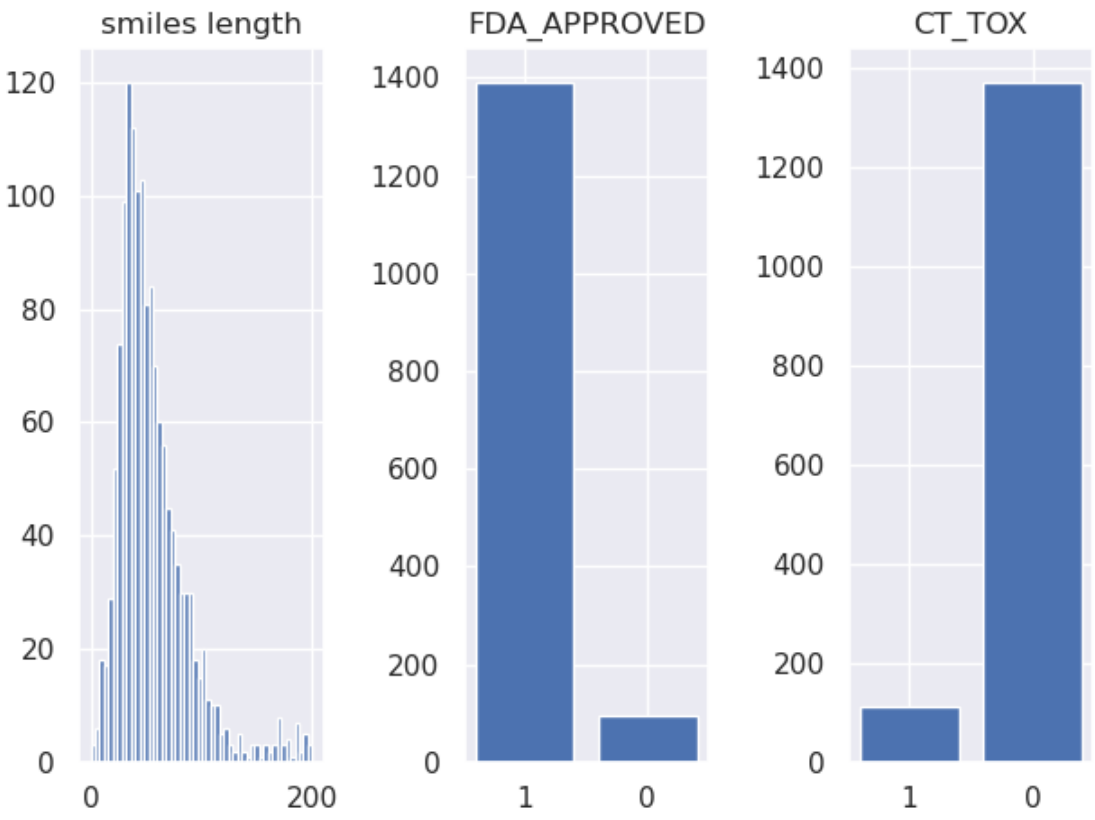
1.3 bbbp dataset

[57]:
```
df = pd.read_csv('data/bbbp.csv')
length = df["smiles"].apply(lambda x: len(x))
df["length"] = length # smiles
plt.subplot(121)
_ = plt.hist(df["length"], bins=50, range=(0,200))
plt.title("smiles length")
plt.subplot(122)
plt.bar(["1","0"],[(df["p_np"]==1).sum(),(df["p_np"]==0).sum()])
plt.title("Class")
plt.tight_layout()
```



1.4 ClinTox

[61]:
```
df = pd.read_csv('data/clintox.csv')
length = df["smiles"].apply(lambda x: len(x))
df["length"] = length # smiles
plt.subplot(131)
_ = plt.hist(df["length"], bins=50, range=(0,200))
plt.title("smiles length")
plt.subplot(132)
plt.bar(["1","0"],[(df["FDA_APPROVED"]==1).sum(),(df["FDA_APPROVED"]==0).sum()])
plt.title("FDA_APPROVED")
plt.subplot(133)
plt.bar(["1","0"],[(df["CT_TOX"]==1).sum(),(df["CT_TOX"]==0).sum()])
plt.title("CT_TOX")
plt.tight_layout()
```


smiles length
120
100
80
60
40
20
0
0
200
FDA_APPROVED
1400
1200
1000
800
600
400
200
0
1
0
CT_TOX
1400
1200
1000
800
600
400
200
0
1
0