CDS Mini Project

Predicting environmental carcinogens

Submitted By:

Palak Arora (16103046)

Anjali Sharma (16103015)

Aman Parmar (16103221)

Sameer Kumar Bairwa (16103309)

Submitted To:

Megha Rathi

Objective

To create a wholesome project that can help us predict the nature of the elements occurring in various natural substances, whether they are carcinogenic or not with the help of the globally available tox21 dataset and biological indicator p53 protein.

Introduction

What are carcinogens?

A carcinogen is any substance, radionuclide, or radiation that promotes carcinogenesis, the formation of cancer. This may be due to the ability to damage the genome or to the disruption of cellular metabolic processes. Detecting carcinogens is a must as cancer being a so widespread disease is affecting a wide lot of the population of the world on the daily basis.

What is tox 21?

The Toxicology in the 21st Century program, or Tox21, is a unique collaboration between several federal agencies to develop new ways to rapidly test whether substances adversely affect human health. Substances assayed in Tox21 include a diverse range of products such as: commercial chemicals, pesticides, food additives/contaminants, and medical compounds.

Why the p53 protein?

The p53 gene encodes a protein of the same name and is known as a tumor-suppressor protein. The p53 protein is expressed in cells when they undergo DNA damage — which can transform a normal cell into a cancerous one. To counteract the effects, p53 can cause growth arrest, repair DNA, or begin the process of cell death. Therefore, when DNA damage occurs, there is a significant increase in p53 expression. This increase in protein expression is a good indicator of irregular cell health. The Tox21 data was generated by testing cell lines which produce a florescent reporter gene product under the control of p53 cellular machinery. By measuring levels of the reporter gene product against various compounds, researchers were able to determine whether a compound was an agonist (activator) of the p53 pathway or not.

What is molecular fingerprinting?

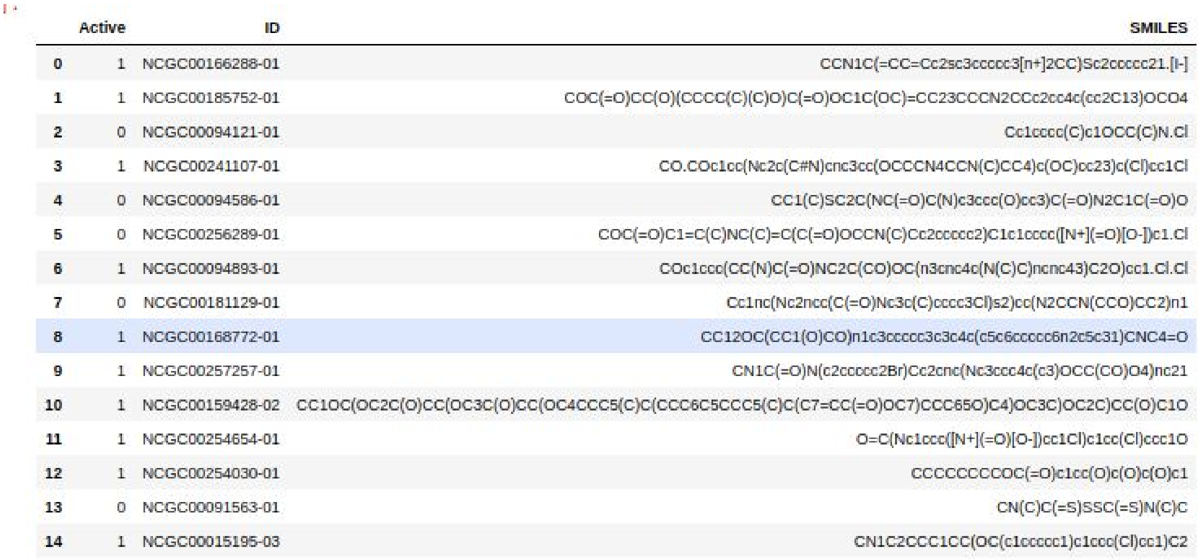
Molecular fingerprints are a way of encoding the structure of a molecule. The most common type of fingerprint is a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule.

What is SMILES notation?

SMILES or Simplified Molecular Input Line Entry System is a chemical notation that allows a user to represent a chemical structure in a way that can be used by the computer. SMILES is an easily learned and flexible notation

Dataset description

The dataset used for the project is the worldwide renowned TOX21 dataset, which contains information about various naturally occurring elements. The information provided is the nature of the element to be carcinogen or not , a binary indicator, the molecular formula of the element and the SMILES notation of the same.



	Active	ID	SMILES
0	1	NCGC00166288-01	CCN1C(=CC=Cc2sc3ccccc3[n+]2CC)Sc2ccccc21.[I-]
1	1	NCGC00185752-01	COC(=O)CC(O)(CCCC(C)(C)O)C(=O)OC1C(OC)=CC23CCCN2CCc2cc4c(cc2C13)OCO4
2	0	NCGC00094121-01	Cc1cccc(C)c1OCC(C)N.Cl
3	1	NCGC00241107-01	CO.COc1cc(Nc2c(C#N)cnc3cc(OCCCN4CCN(C)CC4)c(OC)cc23)c(Cl)cc1Cl
4	0	NCGC00094586-01	CC1(C)SC2C(NC(=O)C(N)c3ccc(O)cc3)C(=O)N2C1C(=O)O
5	0	NCGC00256289-01	COC(=O)C1=C(C)NC(C)=C(C(=O)OCCN(C)Cc2ccccc2)C1c1cccc([N+](=O)[O-])c1.Cl
6	1	NCGC00094893-01	COc1ccc(CC(N)C(=O)NC2C(CO)OC(n3cnc4c(N(C)C)ncnc43)C2O)cc1.Cl.Cl
7	0	NCGC00181129-01	Cc1nc(Nc2ncc(C(=O)Nc3c(C)cccc3Cl)s2)cc(N2CCN(CCO)CC2)n1
8	1	NCGC00168772-01	CC12OC(CC1(O)CO)n1c3ccccc3c3c4c(c5c6ccccc6n2c5c31)CNC4=O
9	1	NCGC00257257-01	CN1C(=O)N(c2ccccc2Br)Cc2cnc(Nc3ccc4c(c3)OCC(CO)O4)nc21
10	1	NCGC00159428-02	CC1OC(OC2C(O)CC(OC3C(O)CC(OC4CCC5(C)C(CCC6C5CCC5(C)C(C7=CC(=O)OC7)CCC65O)C4)OC3C)OC2C)CC(O)C1O
11	1	NCGC00254654-01	O=C(Nc1ccc([N+](=O)[O-])cc1Cl)c1cc(Cl)ccc1O
12	1	NCGC00254030-01	CCCCCCCCOC(=O)c1cc(O)c(O)c(O)c1
13	0	NCGC00091563-01	CN(C)C(=S)SSC(=S)N(C)C
14	1	NCGC00015195-03	CN1C2CCC1CC(OC(c1ccccc1)c1ccc(Cl)cc1)C2

Project Structure

Language used:

Python

Libraries used:

- Numpy
- Pandas
- Os
- Collections
- Rdkit
- Sklearn

Machine learning algorithms used

- Logistic regression
- Gradient boosting
- Cross validation
- K nearest neighbours
- Grid search

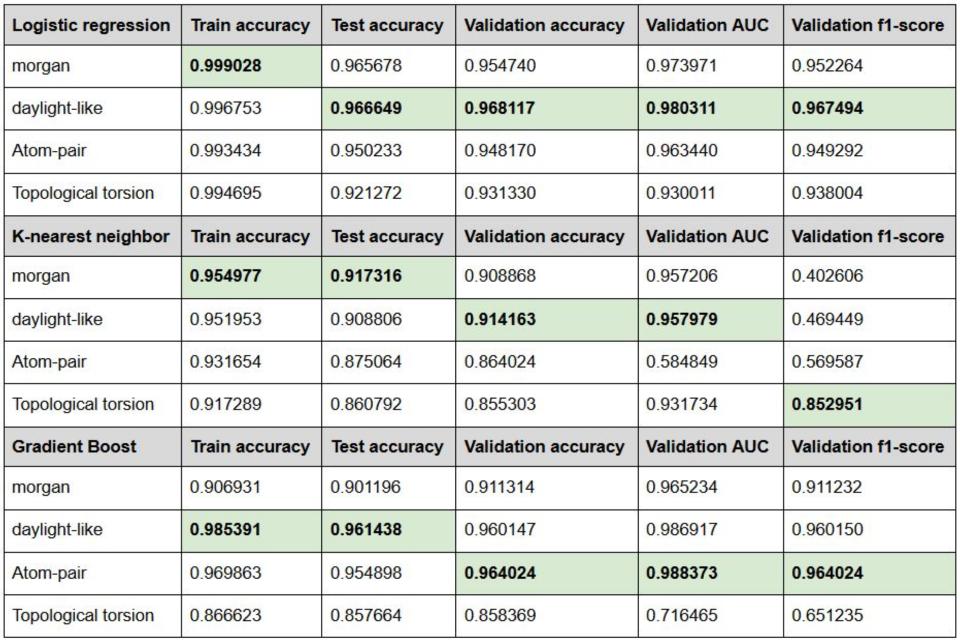
Tasks performed in the project

- Collected information about
    - molecular fingerprinting
    - Smiles
    - Carcinogens
    - Tox21
    - P53
    - different types of fingerprinting:
        - morgan circular fingerprinting
        - daylight-like fingerprinting
        - atom-pair fingerprinting
        - topological torsion fingerprinting
- Data preparation

- Fingerprint generation
- Sampling
- Class balancing using ADASYN
- Creation of training, testing and validation testsets
- Using a logistic regression model
- Cross validation and model fitting
- Using a KNN model
- Applying grid searching and model fitting
- Gradient boosting

Results and conclusion

The table below shows the predictive power of each classifier when using the four fingerprints. Metrics included are accuracy scores are shown for training, test, and validation data sets, AUC scores for validation data, and the f1 scores of the validation data. The highest scores are highlighted in green for each fingerprint and classifier.



Logistic regression	Train accuracy	Test accuracy	Validation accuracy	Validation AUC	Validation f1-score
morgan	**0.999028**	0.965678	0.954740	0.973971	0.952264
daylight-like	0.996753	**0.966649**	**0.968117**	**0.980311**	**0.967494**
Atom-pair	0.993434	0.950233	0.948170	0.963440	0.949292
Topological torsion	0.994695	0.921272	0.931330	0.930011	0.938004
K-nearest neighbor	**Train accuracy**	**Test accuracy**	**Validation accuracy**	**Validation AUC**	**Validation f1-score**
morgan	**0.954977**	**0.917316**	0.908868	0.957206	0.402606
daylight-like	0.951953	0.908806	**0.914163**	**0.957979**	0.469449
Atom-pair	0.931654	0.875064	0.864024	0.584849	0.569587
Topological torsion	0.917289	0.860792	0.855303	0.931734	**0.852951**
Gradient Boost	**Train accuracy**	**Test accuracy**	**Validation accuracy**	**Validation AUC**	**Validation f1-score**
morgan	0.906931	0.901196	0.911314	0.965234	0.911232
daylight-like	**0.985391**	**0.961438**	0.960147	0.986917	0.960150
Atom-pair	0.969863	0.954898	**0.964024**	**0.988373**	**0.964024**
Topological torsion	0.866623	0.857664	0.858369	0.716465	0.651235

On running various classifiers we found out the most appropriate classifier is the ***LOGISTIC REGRESSION*** and the best fingerprinting method for the task is ***DAYLIGHT-LIKE FINGERPRINTING.***