Genetic Risk Profiling

Introduction

Genetic risk profiling involves using a patient's genetic information to assess their predisposition to certain diseases. This case study focuses on using unsupervised machine learning to cluster patients based on their genetic data, with the goal of identifying groups with similar genetic profiles and, potentially, shared health risks. The ability to identify these patient subgroups is a cornerstone of precision medicine, allowing for more targeted and personalized healthcare interventions and preventive strategies. This project demonstrates how genetic data, often complex and high-dimensional, can be leveraged to gain actionable insights into population health.

Dataset

The dataset for this project was sourced from Kaggle and contains a mix of patient demographics and genetic markers. Kaggle Dataset Link: https://www.kaggle.com/datasets/aibuzz/predict-the-genetic-disorders-datasetof-genomes/data

The dataset was curated by Amit Kumar and posted on Kaggle. It has been divided into train and test. For our case study, we will only work with the train dataset.

Dataset Schema

Column name	Column description
Patient Id	Represents the unique identification number of a patient
Patient Age	Represents the age of a patient
Genes in mother's side	Represents a gene defect in a patient's mother
Inherited from father	Represents a gene defect in a patient's father
Maternal gene	Represents a gene defect in the patient's maternal side of the family
Paternal gene	Represents a gene defect in a patient's paternal side of the family
Blood cell count (mcL)	Represents the blood cell count of a patient
Patient First Name	Represents a patient's first name
Family Name	Represents a patient's family name or surname
Father's name	Represents a patient's father's name
Mother's age	Represents a patient's mother's name
Father's age	Represents a patient's father's age
Institute Name	Represents the medical institute where a patient was born
Location of Institute	Represents the location of the medical institute
Status	Represents whether a patient is deceased
Respiratory Rate (breaths/min)	Represents a patient's respiratory breathing rate
Heart Rate (rates/min)	Represents a patient's heart rate
Test 1 - Test 5	Represents different (masked) tests that were conducted on a patient
Parental consent	Represents whether a patient's parents approved the treatment plan
Follow-up	Represents a patient's level of risk (how intense their condition is)
Gender	Represents a patient's gender
Birth asphyxia	Represents whether a patient suffered from birth asphyxia
Autopsy shows birth defect (if applicable)	Represents whether a patient's autopsy showed any birth defects
Place of birth	Represents whether a patient was born in a medical institute or home
Folic acid details (peri-conceptional)	Represents the periconceptional folic acid supplementation details of a patient
H/O serious maternal illness	Represents an unexpected outcome of labor and delivery that resulted in significant short or long-term consequences to a patient's mother
H/O radiation exposure (x-ray)	Represents whether a patient has any radiation exposure history
H/O substance abuse	Represents whether a parent has a history of drug addiction
Assisted conception IVF/ART	Represents the type of treatment used for infertility
History of anomalies in previous pregnancies	Represents whether the mother had any anomalies in her previous pregnancies
No. of previous abortion	Represents the number of abortions that a mother had
Birth defects	Represents whether a patient has birth defects
White Blood cell count (thousand per microliter)	Represents a patient's white blood cell count
Blood test result	Represents a patient's blood test results
Symptom 1 - Symptom 5	Represents (masked) different types of symptoms that a patient had
Genetic Disorder	Represents the genetic disorder that a patient has
Disorder Subclass	Represents the subclass of the disorder

Libraries

In [2]:

```python
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px
import matplotlib.cm as cm
import plotly.graph_objects as go
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

pd.options.display.max_colwidth = 100
pd.set_option('display.max_columns', None)

from numpy.random import seed
seed(42)

import math

from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder
from sklearn.cluster import DBSCAN, AgglomerativeClustering, KMeans, SpectralClustering
from scipy.cluster.hierarchy import dendrogram
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler, RobustScaler
from sklearn.metrics import silhouette_score, davies_bouldin_score
from sklearn.neighbors import NearestNeighbors
from sklearn.mixture import GaussianMixture
from sklearn.base import clone
from sklearn.manifold import TSNE
from sklearn.feature_selection import SelectKBest, f_classif, SelectFromModel, RFE
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier

import warnings
warnings.filterwarnings('ignore')
```

Data

In [3]:

```python
df = pd.read_csv('../Data/genetic_genome/train_genetic_disorders.csv')
df.head()
```

Out[3]:

	Patient Id	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Patient First Name	Family Name	Father's name	Mother's age	Father's age	Institute Name	Location of Institute	Statu
0	PID0x6418	2.0	Yes	No	Yes	No	4.760603	Richard	NaN	Larre	NaN	NaN	Boston Specialty & Rehabilitation Hospital	55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.06924724545246)	Aliv
1	PID0x25d5	4.0	Yes	Yes	No	No	4.910669	Mike	NaN	Brycen	NaN	23.0	St. Margaret's Hospital For Women	1515 COMMONWEALTH AV\nALLSTON/BRIGHTON, MA 02135\n(42.34665771451756, -71.14136122385321)	Decease
2	PID0x4a82	6.0	Yes	No	No	No	4.893297	Kimberly	NaN	Nashon	41.0	22.0	NaN	-	Aliv
3	PID0x4ac8	12.0	Yes	No	Yes	No	4.705280	Jeffery	Hoelscher	Aayaan	21.0	NaN	NaN	55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.06924724545246)	Decease
4	PID0x1bf7	11.0	Yes	No	NaN	Yes	4.720703	Johanna	Stutzman	Suave	32.0	NaN	Carney Hospital	300 LONGWOOD AV\nFENWAY/ KENMORE, MA 02115\n(42.337592548462226, -71.10472284437952)	Aliv

Exploratory Data Analysis

In [4]:

```python
df.columns
```

Out[4]:

```
Index(['Patient Id', 'Patient Age', 'Genes in mother's side',
       'Inherited from father', 'Maternal gene', 'Paternal gene',
       'Blood cell count (mcL)', 'Patient First Name', 'Family Name',
       'Father's name', 'Mother's age', 'Father's age', 'Institute Name',
       'Location of Institute', 'Status', 'Respiratory Rate (breaths/min)',
       'Heart Rate (rates/min', 'Test 1', 'Test 2', 'Test 3', 'Test 4',
       'Test 5', 'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies',
       'No. of previous abortion', 'Birth defects',
       'White Blood cell count (thousand per microliter)', 'Blood test result',
       'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
       'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

In [5]:

```python
df.shape
```

Out[5]:

```
(22083, 45)
```

In [6]:

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22083 entries, 0 to 22082
Data columns (total 45 columns):
 #   Column                                            Non-Null Count  Dtype  
---  ------                                            --------------  -----  
 0   Patient Id                                        21011 non-null  object 
 1   Patient Age                                       19643 non-null  float64
 2   Genes in mother's side                            21011 non-null  object 
 3   Inherited from father                             20724 non-null  object 
 4   Maternal gene                                     18317 non-null  object 
 5   Paternal gene                                     21011 non-null  object 
 6   Blood cell count (mcL)                            21011 non-null  float64
 7   Patient First Name                                21011 non-null  object 
 8   Family Name                                       11771 non-null  object 
 9   Father's name                                     21011 non-null  object 
 10  Mother's age                                      15293 non-null  float64
 11  Father's age                                      15322 non-null  float64
 12  Institute Name                                    16151 non-null  object 
 13  Location of Institute                             21011 non-null  object 
 14  Status                                            21011 non-null  object 
 15  Respiratory Rate (breaths/min)                    18952 non-null  object 
 16  Heart Rate (rates/min                             18986 non-null  object 
 17  Test 1                                            18992 non-null  float64
 18  Test 2                                            18958 non-null  float64
 19  Test 3                                            18970 non-null  float64
 20  Test 4                                            18962 non-null  float64
 21  Test 5                                            18939 non-null  float64
 22  Parental consent                                  18991 non-null  object 
 23  Follow-up                                         18941 non-null  object 
 24  Gender                                            18948 non-null  object 
 25  Birth asphyxia                                    18953 non-null  object 
 26  Autopsy shows birth defect (if applicable)        16847 non-null  object 
 27  Place of birth                                    18993 non-null  object 
 28  Folic acid details (peri-conceptional)            18998 non-null  object 
 29  H/O serious maternal illness                      18959 non-null  object 
 30  H/O radiation exposure (x-ray)                    18964 non-null  object 
 31  H/O substance abuse                               18921 non-null  object 
 32  Assisted conception IVF/ART                       19007 non-null  object 
 33  History of anomalies in previous pregnancies      18945 non-null  object 
 34  No. of previous abortion                          18957 non-null  float64
 35  Birth defects                                     18959 non-null  object 
 36  White Blood cell count (thousand per microliter)  18965 non-null  float64
 37  Blood test result                                 18977 non-null  object 
 38  Symptom 1                                         18955 non-null  float64
 39  Symptom 2                                         18899 non-null  float64
 40  Symptom 3                                         19008 non-null  float64
 41  Symptom 4                                         18987 non-null  float64
 42  Symptom 5                                         18956 non-null  float64
 43  Genetic Disorder                                  18962 non-null  object 
 44  Disorder Subclass                                 18943 non-null  object 
dtypes: float64(16), object(29)
memory usage: 7.6+ MB
```

In [7]:

```
# Dropping the features
df.drop(columns=[
    'Patient Id','Patient First Name','Family Name','Father\'s name','Institute Name',
    'Location of Institute','Parental consent'],
    axis=1, inplace=True)
```

In [8]:

```
df=df.rename(columns={
    "Genes in mother's side":'defective_mother',
    'Inherited from father':'defective_father',
    'Maternal gene':'maternal_gene',
    'Paternal gene':'paternal_gene',
    'Respiratory Rate (breaths/min)':'respiratory_rate',
    'Heart Rate (rates/min':'heart_rate',
    'Parental consent':'parental_consent',
    'Follow-up':'follow_up',
    'Birth asphyxia':'birth_asphyxia',
    'Autopsy shows birth defect (if applicable)':'birth_defect_autopsy',
    'Place of birth':'birth_place',
    'Folic acid details (peri-conceptional)':'folic_acid_periconceptional',
    'H/O serious maternal illness':'maternal_illness',
    'H/O radiation exposure (x-ray)':'radiation_exposure',
    'H/O substance abuse':'substance_abuse',
    'Assisted conception IVF/ART':'assisted_conception',
    'History of anomalies in previous pregnancies':'previous_pregnancy_anomalies',
    'Birth defects':'birth_defects',
    'Blood test result':'blood_test_result',
    'Genetic Disorder':'genetic_disorder',
    'Disorder Subclass':'disorder_subclass',
    'Patient Age':'patient_age',
    'Blood cell count (mcL)':'blood_cell_count',
    "Mother's age":'mother_age',
    "Father's age":'father_age',
    'No. of previous abortion':'num_previous_abortion',
    'White Blood cell count (thousand per microliter)':'WBC_count'
})
```

In [9]:

```
df.select_dtypes(exclude = 'object').describe()
```

Out[9]:

	patient_age	blood_cell_count	mother_age	father_age	Test 1	Test 2	Test 3	Test 4	Test 5	num_previous_abortion	WBC_count	Symptom 1	Sym
count	19643.000000	21011.000000	15293.000000	15322.000000	18992.0	18958.0	18970.0	18962.0	18939.0	18957.000000	18965.000000	18955.000000	18899.0
mean	6.974851	4.899004	34.522527	41.942436	0.0	0.0	0.0	1.0	0.0	2.000106	7.484680	0.592034	0.5
std	4.322584	0.199829	9.847256	13.027701	0.0	0.0	0.0	0.0	0.0	1.411488	2.653633	0.491470	0.4
min	0.000000	4.092727	18.000000	20.000000	0.0	0.0	0.0	1.0	0.0	0.000000	3.000000	0.000000	0.0
25%	3.000000	4.763230	26.000000	31.000000	0.0	0.0	0.0	1.0	0.0	1.000000	5.419026	0.000000	0.0
50%	7.000000	4.899548	35.000000	42.000000	0.0	0.0	0.0	1.0	0.0	2.000000	7.473071	1.000000	1.0
75%	11.000000	5.033977	43.000000	53.000000	0.0	0.0	0.0	1.0	0.0	3.000000	9.528684	1.000000	1.0
max	14.000000	5.609829	51.000000	64.000000	0.0	0.0	0.0	1.0	0.0	4.000000	12.000000	1.000000	1.0

In [10]:

```
df.select_dtypes(include = 'object').describe()
```

Out[10]:

	defective_mother	defective_father	maternal_gene	paternal_gene	Status	respiratory_rate	heart_rate	follow_up	Gender	birth_asphyxia	birth_defect_autopsy	bir
count	21011	20724	18317	21011	21011	18952	18986	18941	18948	18953	16847	
unique	2	2	2	2	2	2	2	2	3	4	3	
top	Yes	No	Yes	No	Alive	Normal (30-60)	Normal	Low	Ambiguous	Yes	Not applicable	
freq	12509	12508	10125	11887	10572	9595	9715	9564	6385	4839	10572	

In [11]:

```
# checking for count of duplicate records
df.duplicated().sum()
```

Out[11]:

```
1071
```

In [12]:

```
# dropping duplicate records
df.drop_duplicates(inplace=True)
```

In [13]:

```
df.isnull().sum()
```

Out[13]:

```
patient_age                     1369
defective_mother                   1
defective_father                 288
maternal_gene                   2695
paternal_gene                      1
blood_cell_count                   1
mother_age                      5719
father_age                      5690
Status                             1
respiratory_rate                2060
heart_rate                      2026
Test 1                          2020
Test 2                          2054
Test 3                          2042
Test 4                          2050
Test 5                          2073
follow_up                       2071
Gender                          2064
birth_asphyxia                  2059
birth_defect_autopsy            4165
birth_place                     2019
folic_acid_periconceptional     2014
maternal_illness                2053
radiation_exposure              2048
substance_abuse                 2091
assisted_conception             2005
previous_pregnancy_anomalies    2067
num_previous_abortion           2055
birth_defects                   2053
WBC_count                       2047
blood_test_result               2035
Symptom 1                       2057
Symptom 2                       2113
Symptom 3                       2004
Symptom 4                       2025
Symptom 5                       2056
genetic_disorder                2050
disorder_subclass               2069
dtype: int64
```

In [14]:

```
# percentage of missing values
percentage_missing = df.isnull().sum() / df.shape[0] * 100
percentage_missing
```

Out[14]:

```
patient_age                      6.515325
defective_mother                 0.004759
defective_father                 1.370645
maternal_gene                   12.826004
paternal_gene                    0.004759
blood_cell_count                 0.004759
mother_age                      27.217780
father_age                      27.079764
Status                           0.004759
respiratory_rate                 9.803922
heart_rate                       9.642109
Test 1                           9.613554
Test 2                           9.775366
Test 3                           9.718256
Test 4                           9.756330
Test 5                           9.865791
follow_up                        9.856273
Gender                           9.822958
birth_asphyxia                   9.799162
birth_defect_autopsy            19.822006
birth_place                      9.608795
folic_acid_periconceptional      9.584999
maternal_illness                 9.770607
radiation_exposure               9.746811
substance_abuse                  9.951456
assisted_conception              9.542166
previous_pregnancy_anomalies     9.837236
num_previous_abortion            9.780126
birth_defects                    9.770607
WBC_count                        9.742052
blood_test_result                9.684942
Symptom 1                        9.789644
Symptom 2                       10.056158
Symptom 3                        9.537407
Symptom 4                        9.637350
Symptom 5                        9.784885
genetic_disorder                 9.756330
disorder_subclass                9.846754
dtype: float64
```

Genetic Disorder and Discorder subclass can be used for evaluating clusters

In [15]:

```
df['genetic_disorder'].unique()
```

Out[15]:

```
array(['Mitochondrial genetic inheritance disorders', nan,
       'Multifactorial genetic inheritance disorders',
       'Single-gene inheritance diseases'], dtype=object)
```

In [16]:

```
df['disorder_subclass'].unique()
```

Out[16]:

```
array(["Leber's hereditary optic neuropathy", 'Cystic fibrosis',
       'Diabetes', 'Leigh syndrome', 'Cancer', 'Tay-Sachs',
       'Hemochromatosis', 'Mitochondrial myopathy', nan, "Alzheimer's"],
      dtype=object)
```

In [17]:

```
len(df['disorder_subclass'].unique())
```

Out[17]:

```
10
```

In [18]:

```
# removing rows were disorder_subclass is nan
df=df[(df['genetic_disorder'].isnull()!=True)&(df['disorder_subclass'].isnull()!=True)]
df.shape
```

Out[18]:

```
(17160, 38)
```

In [19]:

```
df[['genetic_disorder','disorder_subclass']].isnull().sum()
```

Out[19]:

```
genetic_disorder     0
disorder_subclass    0
dtype: int64
```

In [20]:

```
# printing the unique values of all columns
for col in df.columns:
    print(f"{col}: {df[col].unique()}")
```

```
patient_age: [ 2.  6. 12. 11. 14.  3.  4.  7.  1.  0. nan 10.  5.  8.  9. 13.]
defective_mother: ['Yes' 'No']
defective_father: ['No' 'Yes' nan]
maternal_gene: ['Yes' 'No' nan]
paternal_gene: ['No' 'Yes']
blood_cell_count: [4.76060309 4.89329743 4.70528039 ... 5.21475028 5.22482777 5.13794212]
mother_age: [nan 41. 21. 32. 40. 45. 44. 50. 30. 24. 36. 51. 23. 49. 46. 18. 38. 37.
 42. 48. 28. 25. 19. 47. 34. 35. 22. 33. 20. 29. 26. 31. 27. 43. 39.]
father_age: [nan 22. 63. 44. 42. 56. 20. 24. 57. 48. 30. 55. 62. 32. 41. 52. 28. 31.
 61. 35. 49. 50. 23. 29. 64. 39. 34. 51. 25. 43. 60. 53. 58. 26. 27. 59.
 38. 47. 54. 21. 37. 36. 46. 40. 45. 33.]
Status: ['Alive' 'Deceased']
respiratory_rate: ['Normal (30-60)' 'Tachypnea' nan]
heart_rate: ['Normal' 'Tachycardia' nan]
Test 1: [ 0. nan]
Test 2: [nan  0.]
Test 3: [nan  0.]
Test 4: [ 1. nan]
Test 5: [ 0. nan]
follow_up: ['High' 'Low' nan]
Gender: [nan 'Male' 'Female' 'Ambiguous']
birth_asphyxia: [nan 'No record' 'Not available' 'Yes' 'No']
birth_defect_autopsy: ['Not applicable' 'No' nan 'Yes']
birth_place: ['Institute' nan 'Home']
folic_acid_periconceptional: ['No' 'Yes' nan]
maternal_illness: [nan 'No' 'Yes']
radiation_exposure: ['No' 'Yes' '-' 'Not applicable' nan]
substance_abuse: ['No' nan 'Not applicable' '-' 'Yes']
assisted_conception: ['No' 'Yes' nan]
previous_pregnancy_anomalies: ['Yes' 'No' nan]
num_previous_abortion: [nan  4.  1.  0.  3.  2.]
birth_defects: [nan 'Singular' 'Multiple']
WBC_count: [9.85756248        nan 7.91932098 ... 9.86337418 7.08631173 6.75186636]
blood_test_result: [nan 'normal' 'inconclusive' 'slightly abnormal' 'abnormal']
Symptom 1: [ 1.  0. nan]
Symptom 2: [ 1.  0. nan]
Symptom 3: [ 1.  0. nan]
Symptom 4: [ 1.  0. nan]
Symptom 5: [ 1.  0. nan]
genetic_disorder: ['Mitochondrial genetic inheritance disorders'
 'Multifactorial genetic inheritance disorders'
 'Single-gene inheritance diseases']
disorder_subclass: ["Leber's hereditary optic neuropathy" 'Diabetes' 'Leigh syndrome'
 'Cancer' 'Cystic fibrosis' 'Tay-Sachs' 'Hemochromatosis'
 'Mitochondrial myopathy' "Alzheimer's"]
```

In [21]:

```
df.info()
```