Effects Of Country Specific Features On Summer Olympics Swimming Games Medal Results

Tunç GÜLTEKİN

HABER TÜRK
GÜCÜ ÖZGÜRLÜĞÜNDE

Galeri | Video | Takip et | Beğen

Piyasa USD: 1

Son Dakika Haberleri

G.Saray hazır!

EKONOMİ | SPOR | MAGAZİN | MEDYA | SAĞLIK | GÜNDEM | SİYASET | DÜNYA | SPOR | MEDYA | POLEMİK | YAŞAM

EKONOMİ | MAGAZİN

Türkiye olimpiyatlarda neden madalya kazanamıyor?

BBC TÜRKÇE

Haberler | Multimedya | Fotoğraflarla | İzlenim | Özel Dosyalar | İngilizce Dersleri

Ekonomi | Spor | Bilim-Teknoloji | Sağlık | Yaşam | Basın Özeti

Türkiye neden madalya alamıyor?

Türk Eğitim Sen ve Kamu Sen Genel Başkanı İsmail KONCUK:

Üç tarafı denizlerle çevrili bir ülkede bir yüzücü dahi çıkaramamış olmamız son derece ilginç. Eğitim sistemimizi bu yönüyle masaya yatırmalıyız. Başarısızlığı, eğitim sisteminden bağımsız düşünemeyiz.

Ülkemizi uluslararası alanda temsil edecek yüzücü çok yetişmiyor. Oysa üç yanı denizlerle çevrili bir ülkeyiz. Bu ironinin nedeni nedir?

Derya Büyükuncu: Mantıken dediğiniz gibi olması lazım. Her tarafımız deniz, suyu sevmeliyiz. Ama maalesef biz sporu sevmiyoruz.

VATAN Spor

13.08.2011 - 01:19 | A+ A-

"Üç tarafı denizlerle çevrili bir ülkenin çocukları Derya Büyükuncu dışında bir isim çıkaramıyorsa bu büyük bir kayıptır"

Spor Bakanı Suat Kılıç ilk röportajını VATAN'a verdi:

Özgür BOLAT
ozgurbolat@hurriyet.com.tr

2 Ağustos 2012

Türkiye neden olimpiyatlarda başarısız?

Outline

- Some Statistics about Swimming on Olympics.
- Data Collection
- Data Integration
- Data Preperation
- Data Analysis
  - Simple Visualization
  - Clustering
  - Classification
- Conclusion

Some Information

- Michael Phelps is the most decorated Olympian of all time, with a total of 22 medals.

  (18 gold, 2 silver, 2 bronze)

Some Information

- Derya Büyükuncu is the first ever Turkish swimmer to win an international competition. (1997 Mediterranean Games 3rd)

- Also he has been a Turkish national team member for more than 25 years

Some Statistics

What about countries?

Some Statistics

Rank	Nation	Gold	Silver	Bronze	Total
1	United States (USA)	230	164	126	520
2	Australia (AUS)	57	60	61	178
3	*East Germany (GDR)*	38	32	22	92
4	Hungary (HUN)	25	23	18	66
5	Japan (JPN)	20	24	29	73
6	Netherlands (NED)	19	18	19	56
7	Great Britain (GBR)	15	22	30	67
8	Germany (GER)	13	18	28	59
9	*Soviet Union (URS)*	12	21	26	59
10	China (CHN)	12	17	8	37

Study

- Two different data was used for this;
  - Athlete medal information related to specific olympic swimming event.
  - Country specific features (Coastline Length GNI…)

Athlete Medal Information

- Contains these features;
  - Athlete Name
  - Country Name
  - Sex
  - Race Cource Length
  - Race Type
  - Race Time
  - Year,
  - Medal Type

- Example: *Larsen Jensen,USA, Men,1500,Freestyle,2004,14:45.3, SILVER*

Country Specific Features

- **CoastLine:** Numeric, Defines the coastline length of that country.
- **EducationIndex :** Numeric, Defines the adult literacy rate of that country. It is a value which between 0 and 1.
- **GNI :** Numeric, Defines the Gross National Income of that country.
- **Population :** Numeric, Defines the Population of that country.

Country Specific Features

- **ForestArea :** Numeric, Defines Forest Area of that country.
- **Fishing :** Numeric, Defines Fisheries Production of that country.
- **BirthRate :** Numeric, Defines Birth Rate of that country. It is a value which between 0 and 1.
- **GDP :** Numeric, Defines Gross Domestic Product of that country. It is often considered an indicator of a country's standard of living.

Country Specific Features

- **CO2Rate :** Numeric, Defines CO2 Emission (metric tones) of that country.
- **DeathRate :** Numeric, Defines Death Rate of that country. It is a value which between 0 and 1.
- **HDI:** Numeric, Defines the Human Development Index of that country.

Data Integration

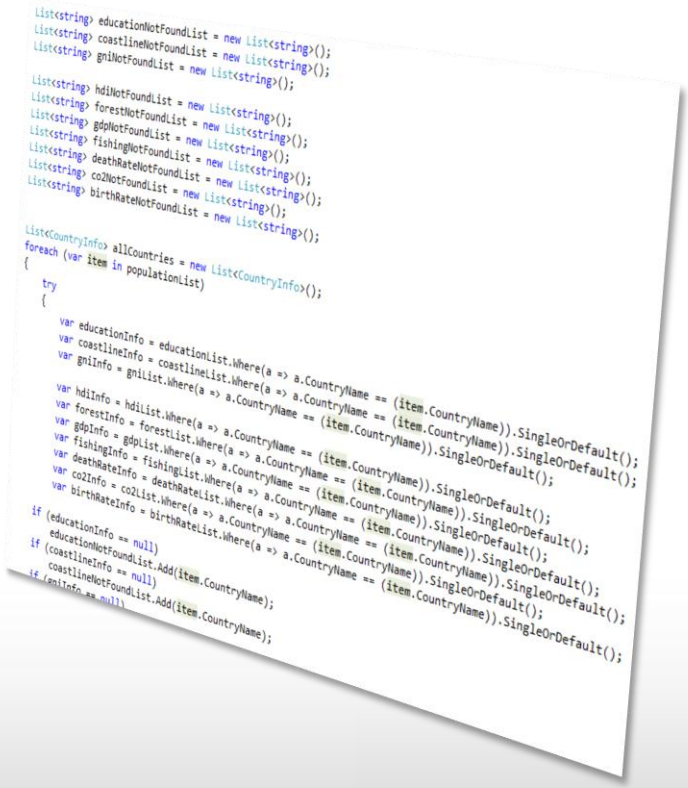
- Country medal information was taken from *www.databaseolympics.com*
- Country specific features were taken from different pages of wikipedia.
- This method created many different tables

Data Integration

Horizontal Data Integration is Required

Data Integration

- To achive this goal, a .net application was developed.



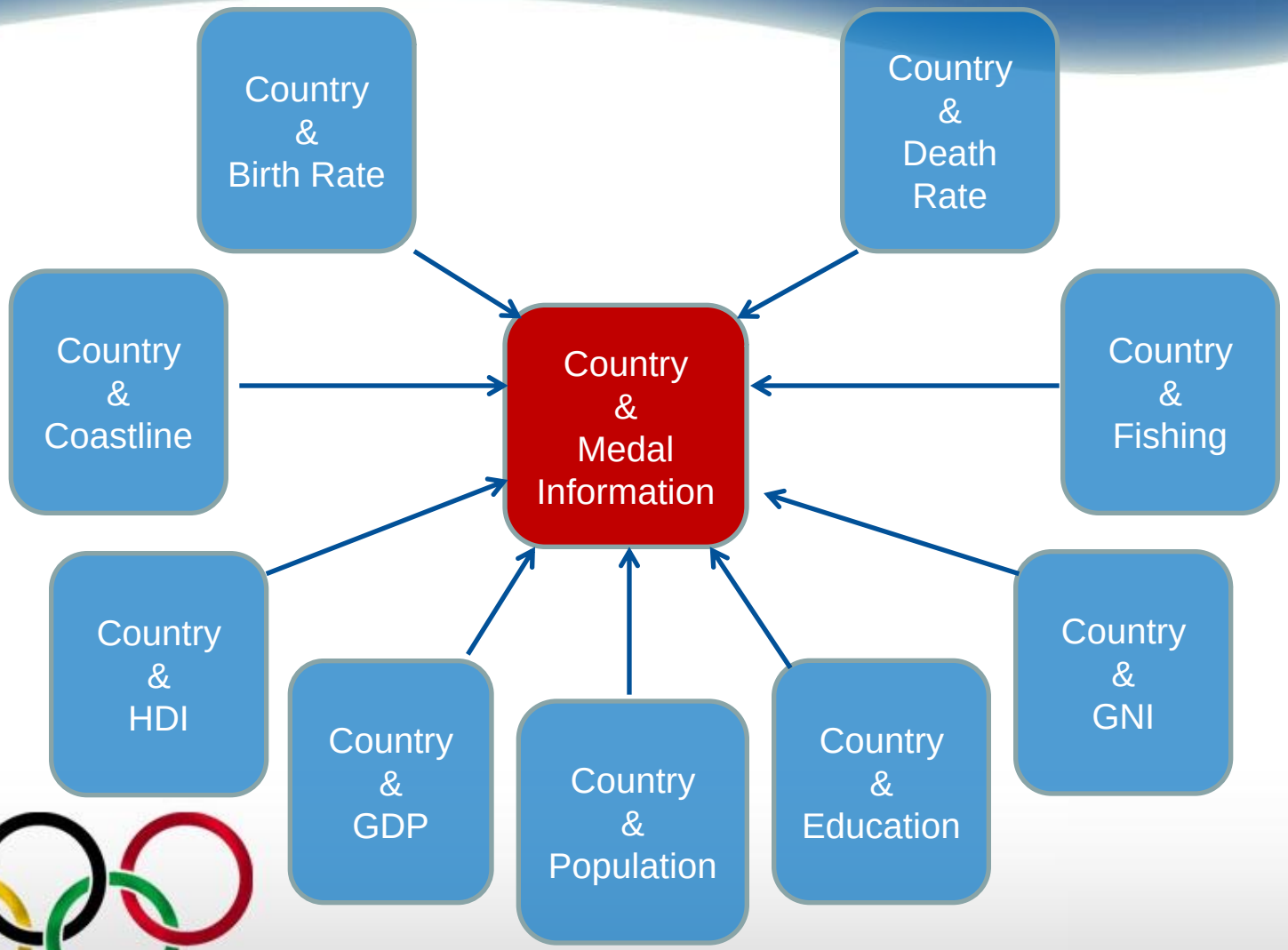
```
List<string> educationNotFoundList = new List<string>();
List<string> coastlineNotFoundList = new List<string>();
List<string> gniNotFoundList = new List<string>();

List<string> hdiNotFoundList = new List<string>();
List<string> forestNotFoundList = new List<string>();
List<string> gdpNotFoundList = new List<string>();
List<string> fishingNotFoundList = new List<string>();
List<string> deathRateNotFoundList = new List<string>();
List<string> co2NotFoundList = new List<string>();
List<string> birthRateNotFoundList = new List<string>();

List<CountryInfo> allCountries = new List<CountryInfo>();
foreach (var item in populationList)
{
    try
    {
        var educationInfo = educationList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        var coastlineInfo = coastlineList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        var gniInfo = gniList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();

        var hdiInfo = hdiList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        var forestInfo = forestList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        var gdpInfo = gdpList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        var fishingInfo = fishingList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        var deathRateInfo = deathRateList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        var co2Info = co2List.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        var birthRateInfo = birthRateList.Where(a => a.CountryName == (item.CountryName)).SingleOrDefault();
        if (educationInfo == null)
            educationNotFoundList.Add(item.CountryName);
        if (coastlineInfo == null)
            coastlineNotFoundList.Add(item.CountryName);
        if (gniInfo == null)
```

Data Integration



Data Integration

- Data was integrated over country name

	A	B	C	D	E	F	G	H	I
1	AthletName	CountryName	Coastline	EducationRate	GNI	Population	HDI	GDP	ForestArea
2	Larsen Jensen	United States	133312	0.968	48890	314419000	91	14447100	3030890

J	K	L	M	N	O	P	Q	R	S
Fishing	BirthRate	CO2	DeathRate	Gender	RaceCourceLength	RaceType	Time	Year	MedalType
100000	13.7	5461014	8.2	Men	1500	Freestyle	14:45.3	2004	SILVER

Data Preparation & Cleaning

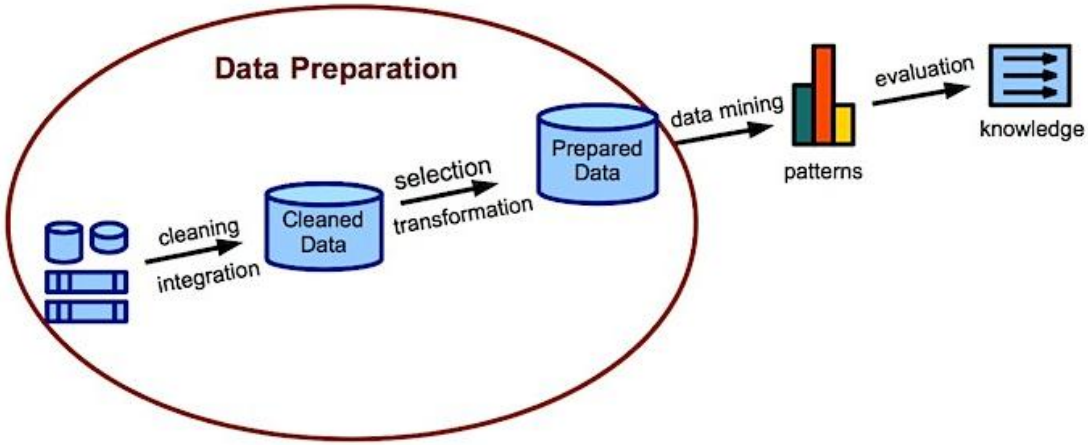
- Country name for some feature sets are inconsistent or misspelled.
  - Ex: USA → United States → America
    China, Ch, PRC …
  - Solved programmaticaly.
- Some countries do not exist anymore.
  - Ex: Soviet Union, West Germany…
  - These ones are eliminated.

Data Preparation & Cleaning

- Some country specific feature information is incomplete for some countries.
  - Ex: Fishery information below 100K Tones.
  - Assigned constant value (100K), and deleted
- Country specific informations belongs to 2000's but medal informations cover all olympics of last century.
  - Solved by using recent years' results (such that 1996 and later.)

Data Preparation & Cleaning

- Some other data preparation techniques were used before performing different analysis methods.



Data Transformation & Analysis

- Three different approaches were followed to analyse
  - Simple visualization of data
  - Clustering
  - Classification

Simple Visualization

- To analyse distribution of data, Weka's Visualization tool was used without performing any special operation.

- Some results were obtained.

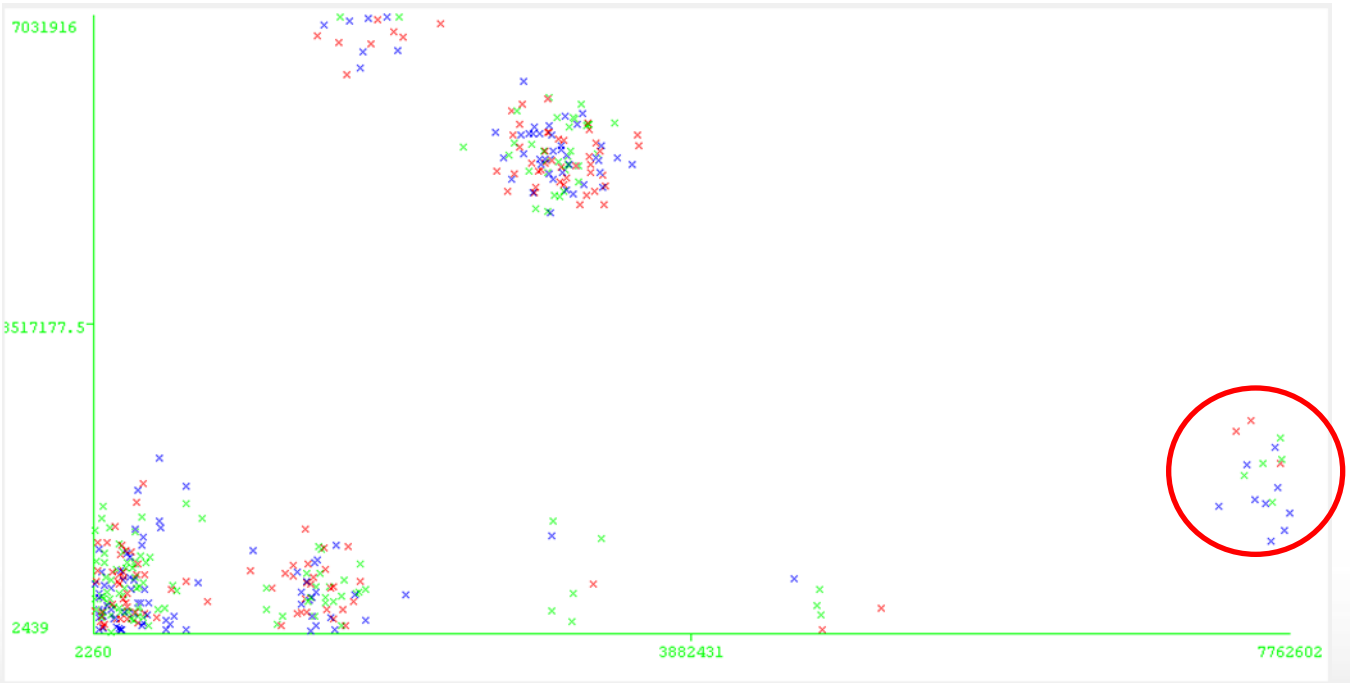
Simple Visualization

- Probability of getting Gold Medal is higher than other medals for the countries which have very high forest area and low CO2 emission.

x: Gold

x: Silver

x: Bronze



X-Axis: Forest Area

Y-Axis: CO2

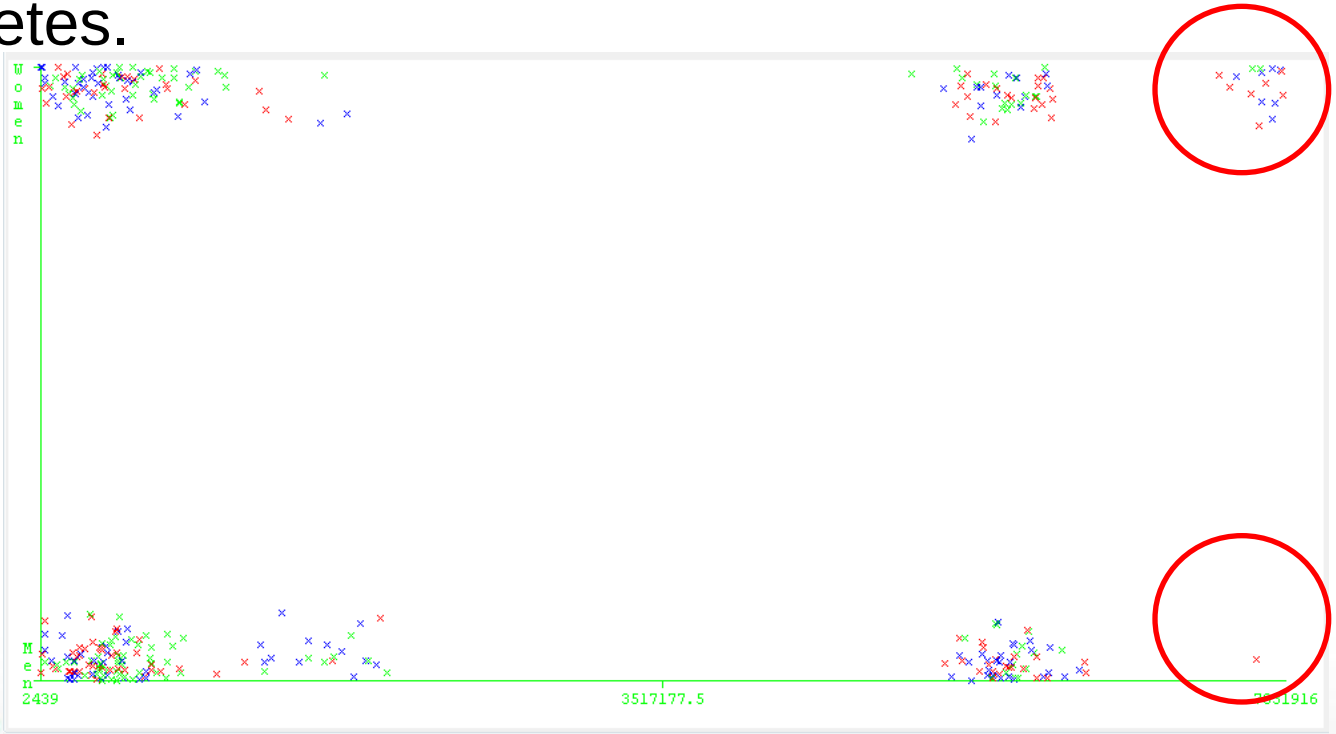
Simple Visualization

- Probability of getting a Medal is high for the countries which have very high CO2 emission and races with women athletes.

x: Gold

x: Silver

x: Bronze



X-Axis: CO2

Y-Axis: Gender

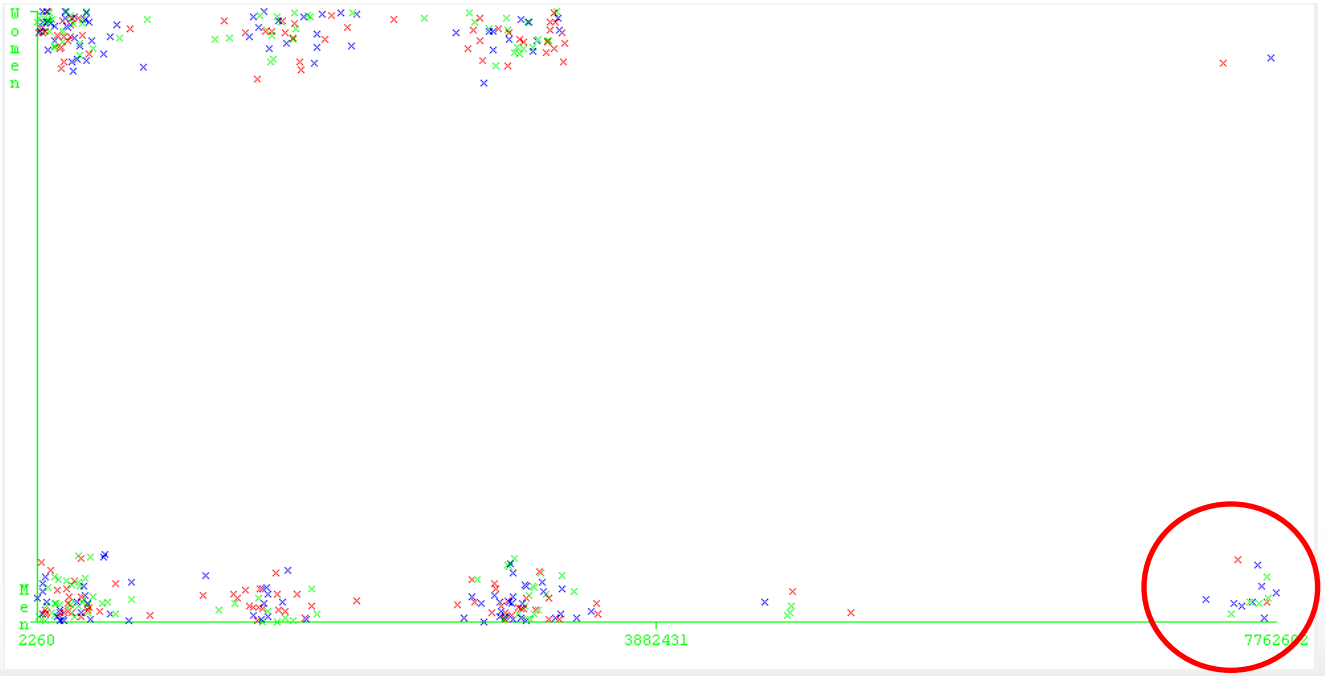
Simple Visualization

- Probability of getting a Medal (especially gold) is higher than other medals for the countries which have very high forest area and races with male athletes.

x: Gold

x: Silver

x: Bronze



X-Axis: Forest Area

Y-Axis: Gender

Simple Visualization

- Probability of getting Silver Medal is lower than other medals for the countries which have medium hdi and medium-low population.

x: Gold

x: Silver

x: Bronze

X-Axis: Population

Y-Axis: HDI

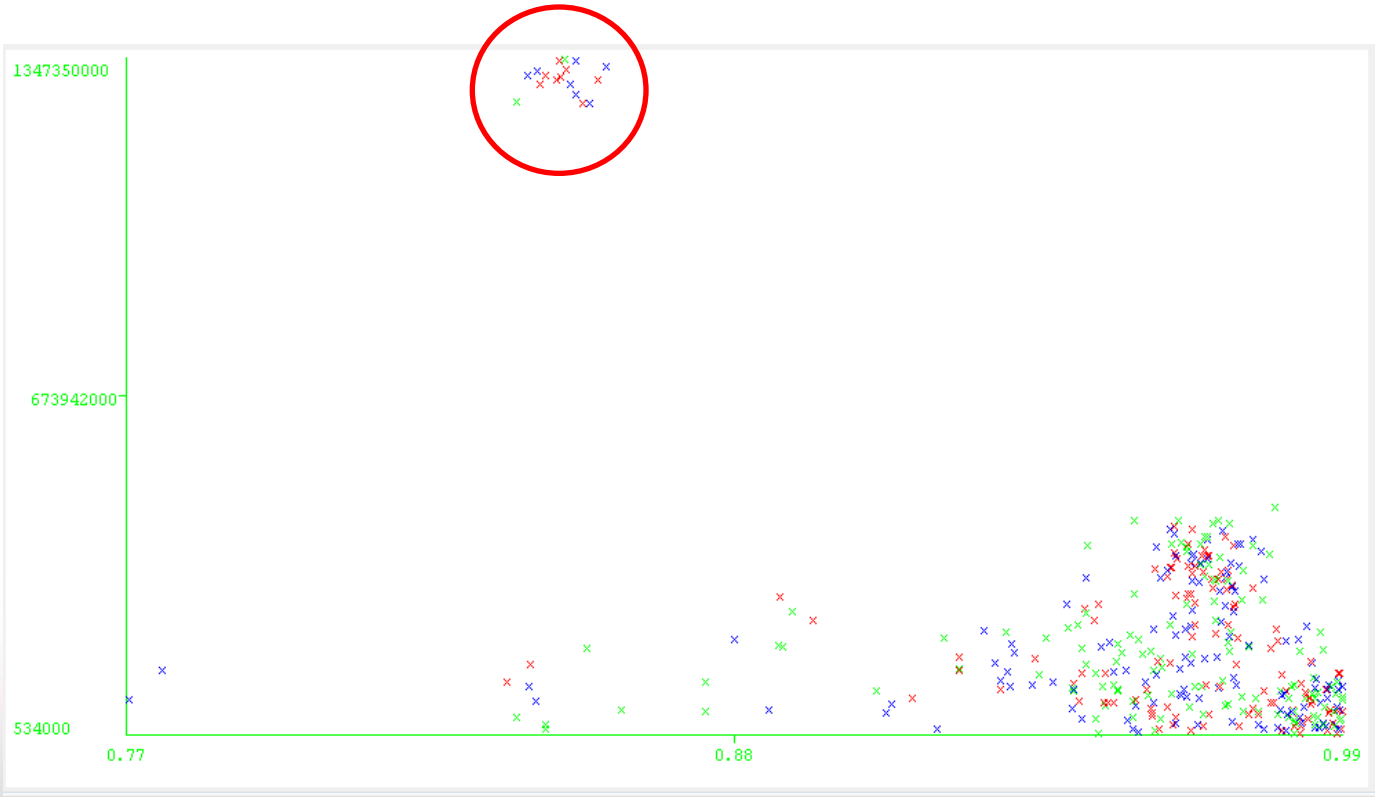
Simple Visualization

- Probability of getting Bronze Medal is lower than other medals for the countries which have very high population and mid-low education index.

x: Gold

x: Silver

x: Bronze



X-Axis: Education

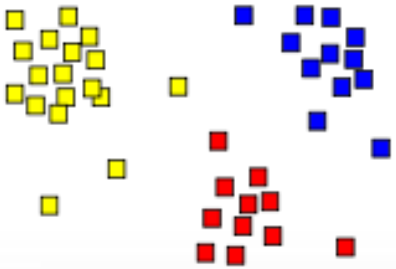
Y-Axis: Population

Simple Visualization

What about more general results?

Clustering

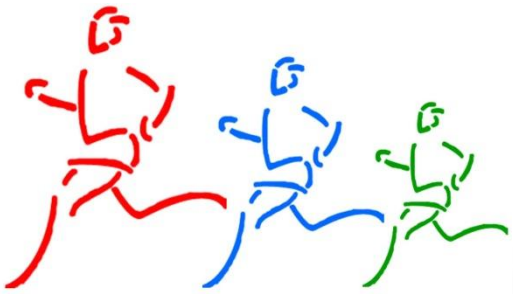
- The goal is finding natural groups, which have similar features and medal results, on the data.
- Some of the attributes are removed;
  - CountryName, RaceTiming, Year
- All data belongs to 1996 and later Olympics



Clustering

- K-Means and EM algorithms were run with different cluster number parameters.
  - With 3, 5 and 10

- K-Means algorithm with 5 clusters exhibited the best performance.



Clustering

- %66 of 304 instances were used for training and leftovers were used for testing.

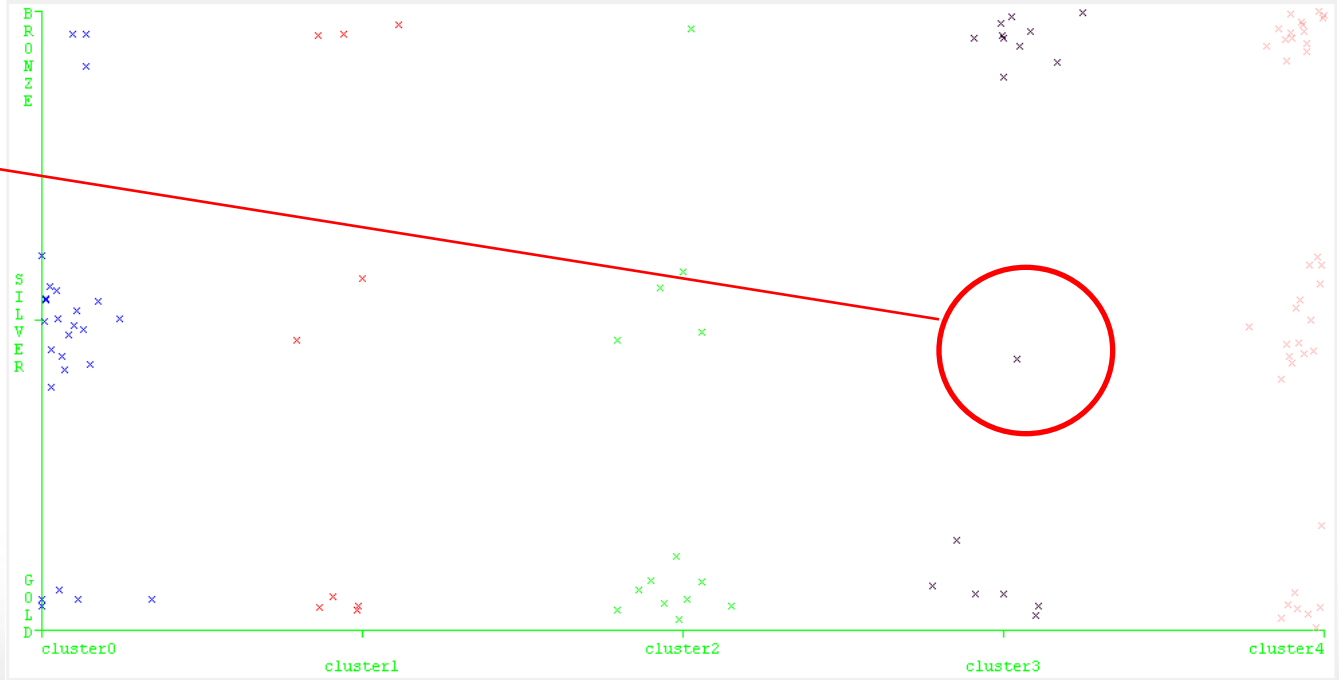
Cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(200)	(48)	(14)	(36)	(40)	(62)
Coastline	66552.555	63901.2708	63334.9286	133312	29955.8	54178.9839
EducationRate	0.9583	0.96	0.9016	0.968	0.9401	0.9761
GNI	35757.45	37713.3333	15185.7143	48890	26199.25	37429.6774
Population	182442969.66	243053645.0625	124374568.6429	314419000	195846751.35	63351950.871
HDI	0.8705	0.8832	0.7217	0.91	0.8226	0.9022
GDP	5664302.195	6377401.2917	1252518.2143	14447100	2923731.425	2776855.8871
ForestArea	1688294.975	1518583.375	4926650.6429	3030890	748929.725	914913.7903
Fishing	2559371.135	4569233.6875	410797.7143	100000	5616683.2	944072.8065
BirthRate	12.508	12.2896	14	13.7	12.4275	11.7
CO2	2248048.815	2668110.0417	1025752.2857	5461014	1519502.15	803279.9839
DeathRate	9.105	8.3646	13.45	8.2	9.9825	8.6565
Gender	Men	Women	Men	Men	Women	Men
RaceCource	267.5	229.1667	221.4286	181.9444	220	387.9032
RaceType	Freestyle	Freestyle	Freestyle	Backstroke	Medley	Freestyle
MedalType	BRONZE	SILVER	BRONZE	GOLD	BRONZE	BRONZE

Clustering

- Are there similar countries, which have not earn any silver medal, on some events?

Almost No Cluster – 3 instance for Silver Medal



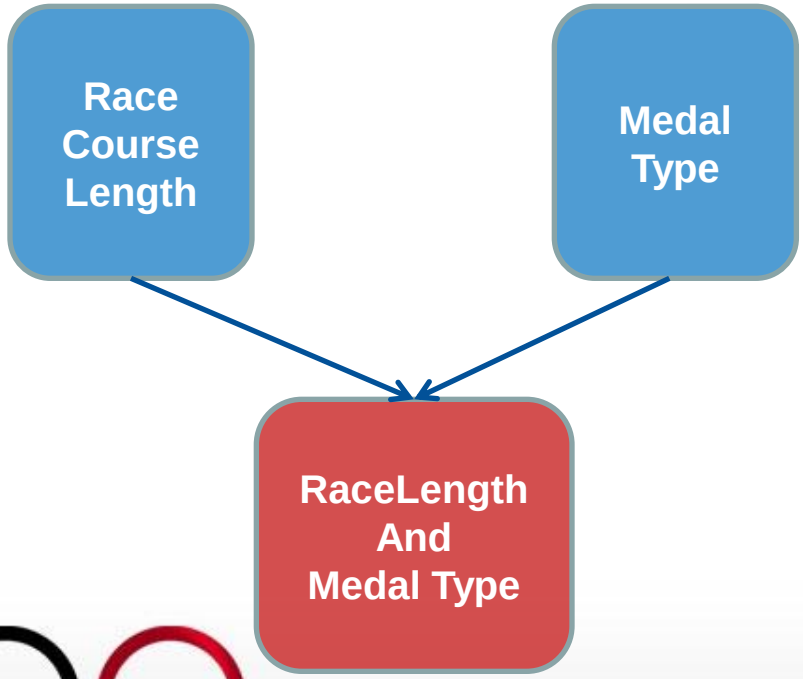
Clustering

- Are there similar countries, which have not earn any silver medal, on some events?
  - 17 instances were found over 200 instance (%16)
    - **Coastline:** [Medium – Low] Scale (Mean: 29995 Km)
    - **Education Rate:** High Scale (Mean: 0.9401)
    - **GNI:** Medium Scale (Mean: 26199 Dolars)
    - **Population:** Low Scale (Mean: 135M)
    - **GDP:** [Medium – Low] Scale (Mean: 2723931 Dolars)
    - **Forest Area:** [Medium – Low] Scale (Mean: 748929 Km2)
    - **CO2:** [Medium – Low] Scale (Mean: 1519502 Tones)
    - **HDI:** High Scale (Mean: 0.8226)
    - **Category:** Women
    - **Race Type:** All types except freestyle

South Africa, Australia Poland ...

Clustering

- Combining some attributes can be useful.



No.	Label
1	1500_GOLD
2	1500_SILVER
3	1500_BRONZE
4	100_GOLD
5	100_SILVER
6	100_BRONZE
7	200_GOLD
8	200_SILVER
9	200_BRONZE
10	400_GOLD
11	400_SILVER

Clustering

- K-Means algorithm result with 5 clusters.

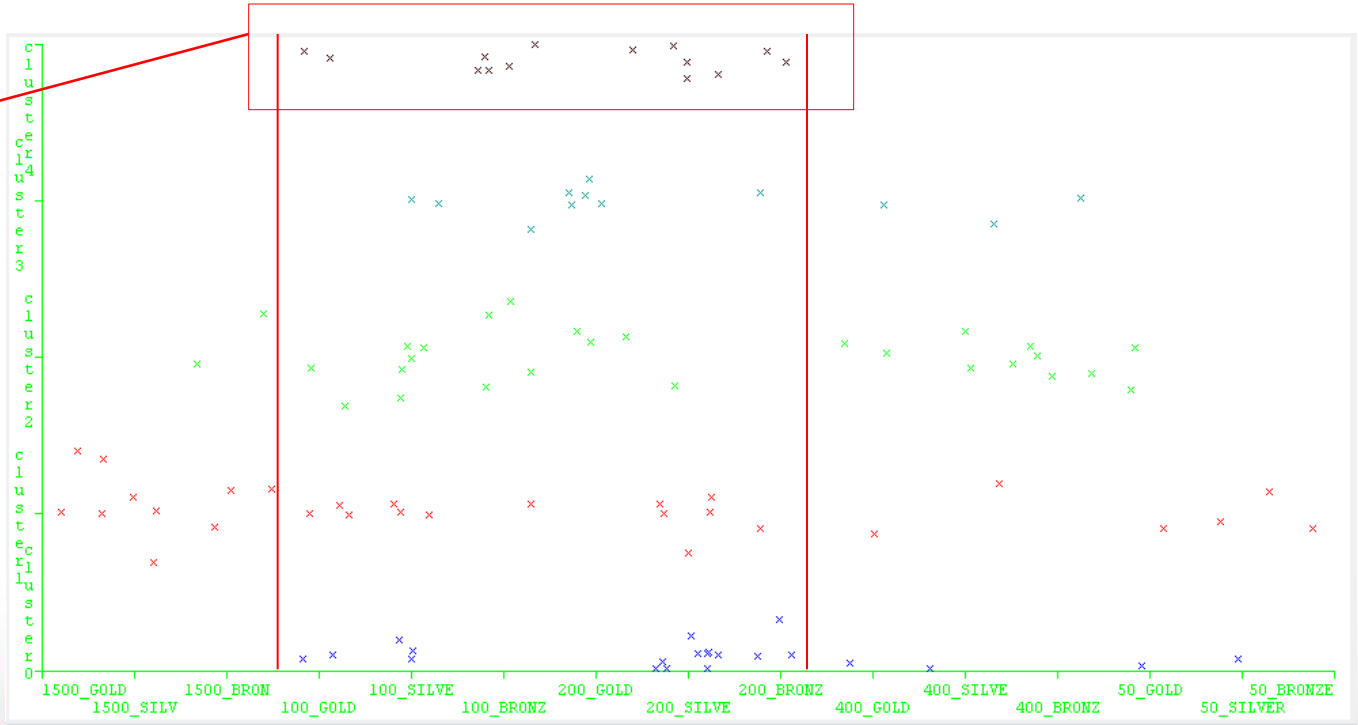
Cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(200)	(54)	(35)	(58)	(24)	(29)
Coastline	66552.555	32609.4259	47267.6	133312	10217.4167	66135.2414
EducationRate	0.9583	0.9625	0.9573	0.968	0.9293	0.9566
GNI	35757.45	35378.3333	30997.1429	48890	18000.4167	30638.9655
Population	182442969.66	154210489.2778	86905868.6857	314419000	193982076.7917	76815526.3448
HDI	0.8705	0.8846	0.8601	0.91	0.7835	0.8497
GDP	5664302.195	2776866.1667	1512588.1143	14447100	1204850.5417	2176581.6897
ForestArea	1688294.975	737830.0926	1433430.8857	3030890	379717.0417	2163491.7241
Fishing	2559371.135	4480266.5185	1858177.3714	100000	6603660.2917	1400544.1379
BirthRate	12.508	11.5537	13.5	13.7	12.0875	11.0517
CO2	2248048.815	1151855.1111	601530.8286	5461014	1075534.2083	820840.5172
DeathRate	9.105	8.4444	8.4286	8.2	11.1167	11.2966
Gender	Men	Women	Men	Men	Women	Men
RaceType	Freestyle	Freestyle	Freestyle	Freestyle	Medley	Breaststroke
RaceAndMedalType	200_BRONZE	200_SILVER	1500_GOLD	200_GOLD	200_GOLD	100_BRONZE

Clustering

- Are there similar countries, which have earned medal, only on 100m and 200m events?

All instances of Cluster – 4 With in range



Clustering

- Are there similar countries, which have earn medal, only on 100m and 200m events?
  - 14 instances were found over 200 instance (%13)
    - **Coastline:** Medium Scale (Mean: 66135 Km)
    - **Education Rate:** High Scale (Mean: 0.9566)
    - **GNI:** [Medium – High] - Scale (Mean: 30638 Dolars)
    - **Population:** Low Scale (Mean: 76M)
    - **GDP:** [Medium – Low] Scale (Mean: 2176581 Dolars)
    - **Forest Area:** [Medium – Low] Scale (Mean: 2163491 Km2)
    - **CO2:** [Low] Scale (Mean: 820840Tones)
    - **HDI:** High Scale (Mean: 0.8497)
    - **Category:** Men
    - **Race Type:** All types except freestyle and medley

Clustering

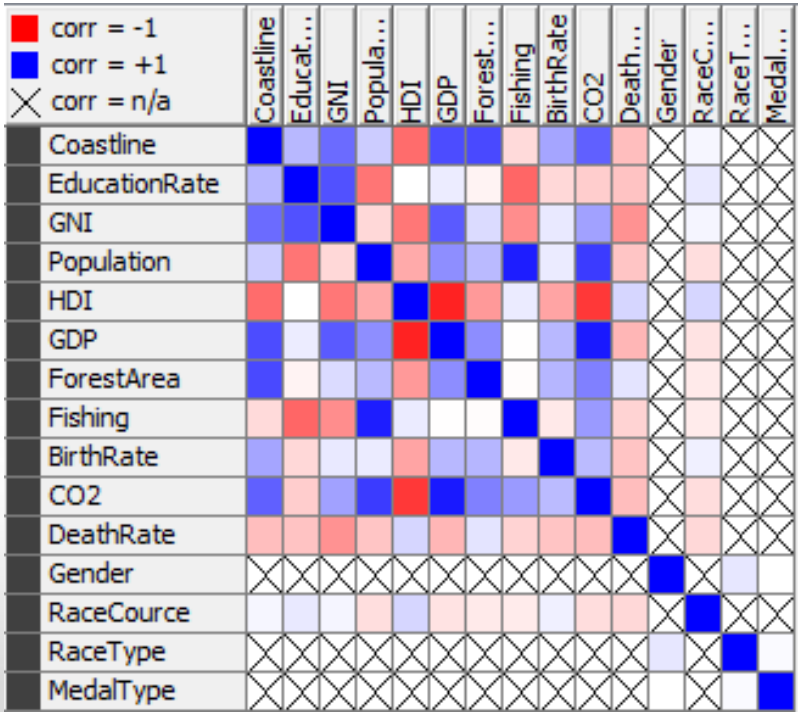
Can be a model built for medal type prediction?

Preparation for Classification

- There were not any correlation with medal type attribute and other features.
- Some preprocessing operations were required.

InfoGainAttrEval With Ranker

```
Ranked attributes:
 0.0006656   14 RaceType
 0.0000259   12 Gender
 0            4 Population
 0            6 GDP
 0            5 HDI
 0            1 Coastline
 0            3 GNI
 0            2 EducationRate
 0           11 DeathRate
 0           15 Year
 0           13 RaceCource
 0            8 Fishing
 0            7 ForestArea
 0           10 CO2
 0            9 BirthRate
```



Preparation for Classification

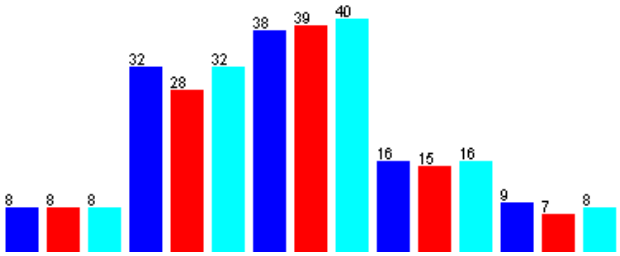
- Z Score Standardization was used to standardize all numeric attributes to have zero mean and unit variance.

Preparation for Classification

- For better classification performance, discretization method was used.
  - Sturges' Rule

$$k = \lceil 1 + \log_2 n \rceil$$

  - 10 Bin is required for 304 instance.
  - 5 Bin discretization showed better performance on classification.

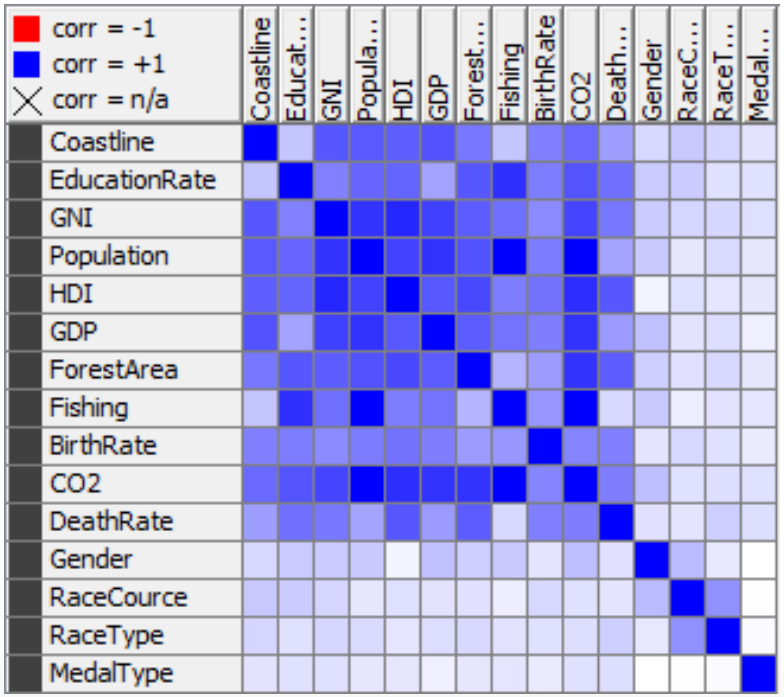


Preparation for Classification

- After the data preparation operations;

InfoGainAttrEval With Ranker

```
Ranked attributes:
 0.0271458    1 Coastline
 0.0242433    2 EducationRate
 0.0218696    9 BirthRate
 0.0136533    3 GNI
 0.0123667   11 DeathRate
 0.0092241   10 CO2
 0.0075056    7 ForestArea
 0.0072773    5 HDI
 0.0071924    4 Population
 0.0071838    6 GDP
 0.002023     8 Fishing
 0.0014055   15 Year
 0.0006656   14 RaceType
 0.0000619   13 RaceCource
 0.0000259   12 Gender
```



Classification

- Different classification algorithms were tested with ten folds cross validation.
  - J48, Decision Stump, Random Forest, RBF Network...

- The most successfull one is J48

Classification

- J48 classification results;

Correctly Classified Instances	120	39.4737 %
Incorrectly Classified Instances	184	60.5263 %
Kappa statistic	0.0916	
Mean absolute error	0.4287	
Root mean squared error	0.5089	
Relative absolute error	96.4949 %	
Root relative squared error	107.9713 %	
Total Number of Instances	304	

Success rate is too low.

Random classification rate would be 33.3333 %

Classification

- J48 classification results;

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.417     0.328      0.394     0.417     0.406      0.534    GOLD
                   0.289     0.324      0.295     0.289     0.292      0.46     SILVER
                   0.471     0.255      0.49      0.471     0.48       0.625    BRONZE
Weighted Avg.      0.395     0.302      0.395     0.395     0.395      0.542

=== Confusion Matrix ===

  a  b  c   <-- classified as
 43 33 27 |  a = GOLD
 45 28 24 |  b = SILVER
 21 34 49 |  c = BRONZE
```

Classification

- Classification by using combined attribute; RaceLengthAndMedalType

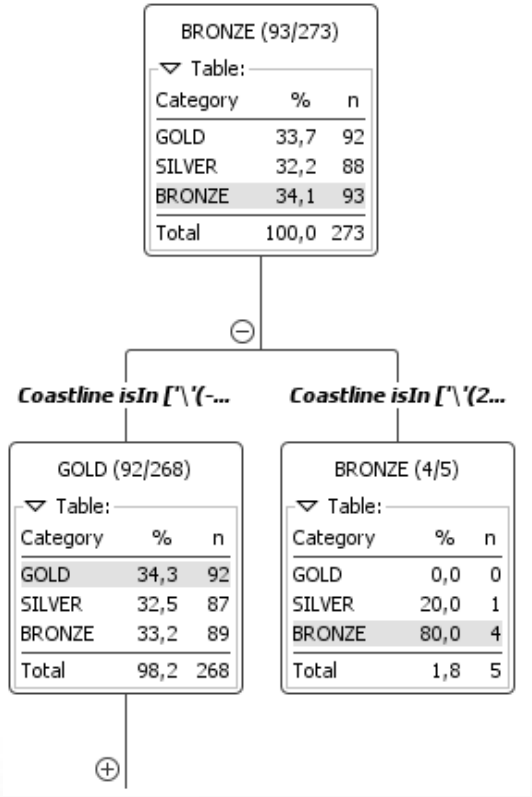
Correctly Classified Instances	67	22.0395 %
Incorrectly Classified Instances	237	77.9605 %
Kappa statistic	0.1548	
Mean absolute error	0.1105	
Root mean squared error	0.2507	
Relative absolute error	90.3079 %	
Root relative squared error	101.3435 %	
Total Number of Instances	304	

Success rate is too low.

Random classification rate would be 6.6667%

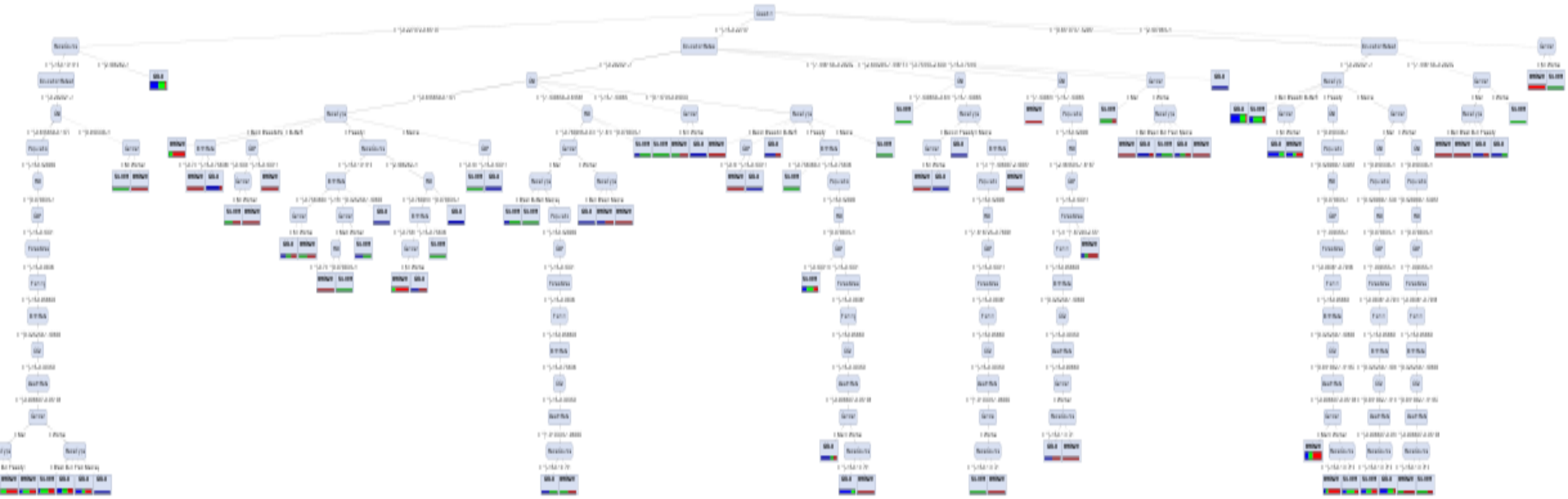
Classification

- Knime's decision tree view;
  (first three node)



Classification

- Rapid Miner's whole tree view (without prunning);



Conclusions

- Coastline Length is the most effective feature to decide medal type but not good enough.
- Some patterns like clusters, exist in the data.
- But they are not sufficient to create a accurate model.

Thanks!

? || !

Questions Comments