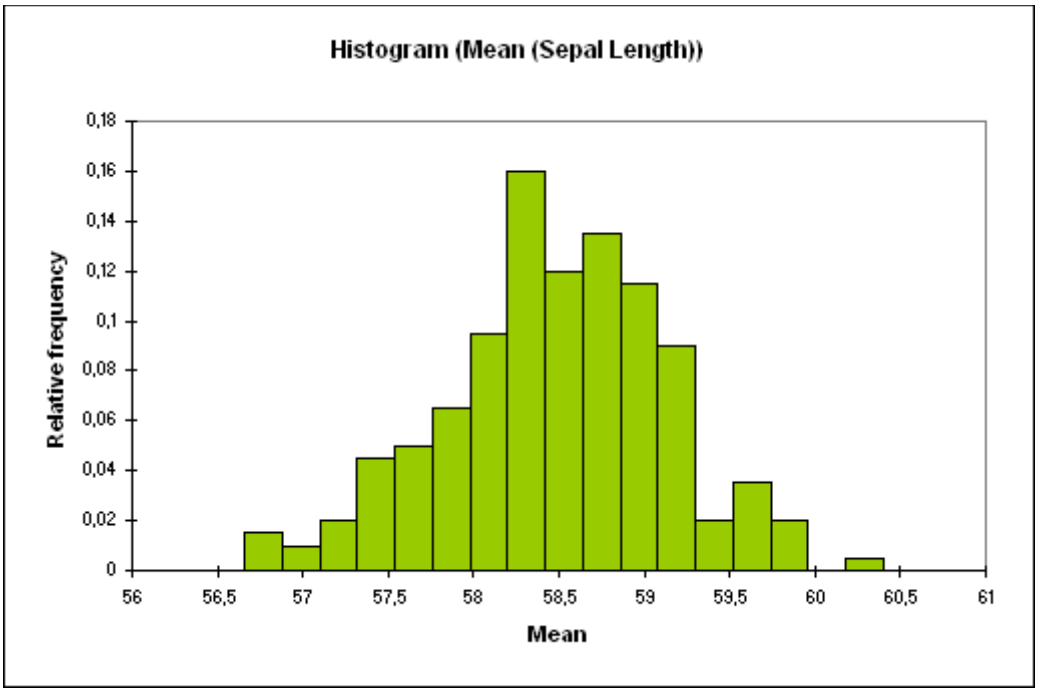
Classification de Data Vizualisation : état de l'art

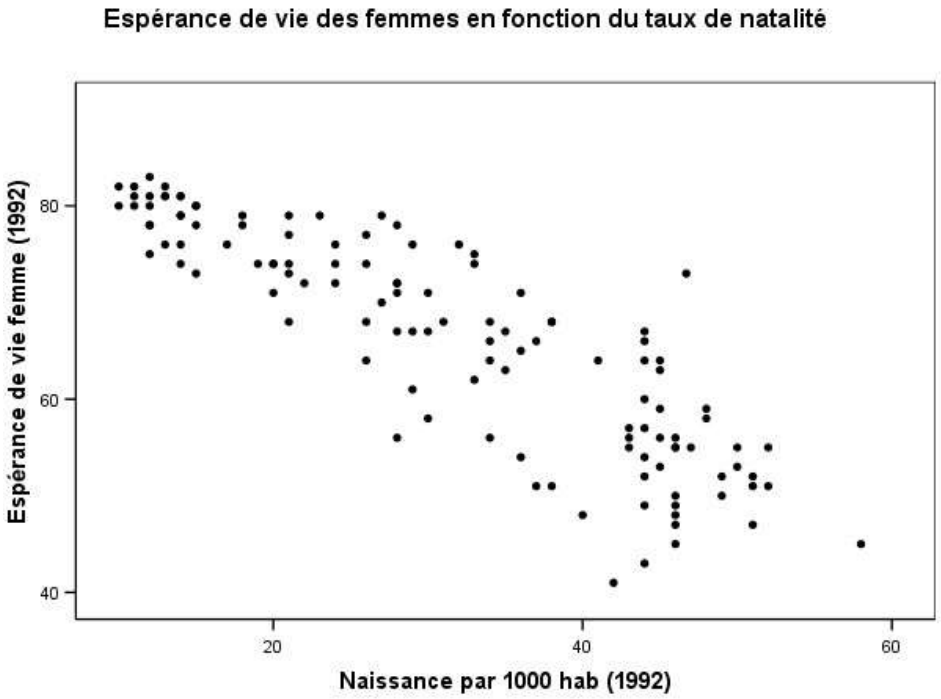
1. Etat de l'art des différents types de Data vizualisation

1.1. Data vizualisation les plus utilisées

Histogramme (bar chart & histogram)



Nuage de point (scatter plot)



Carte proportionnelle (Tree map)

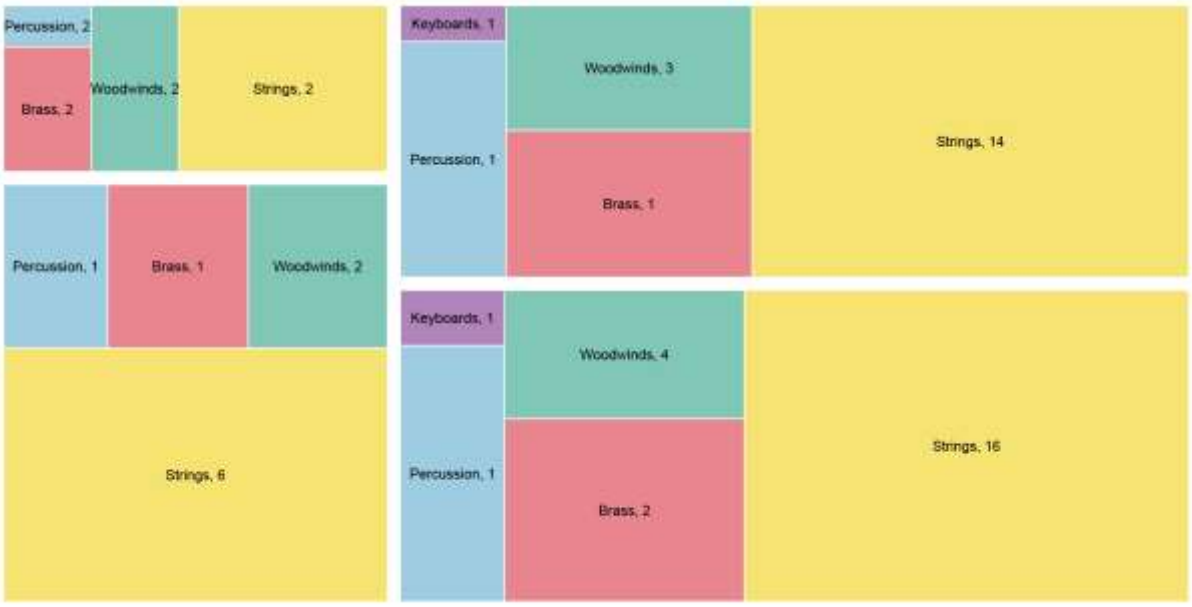
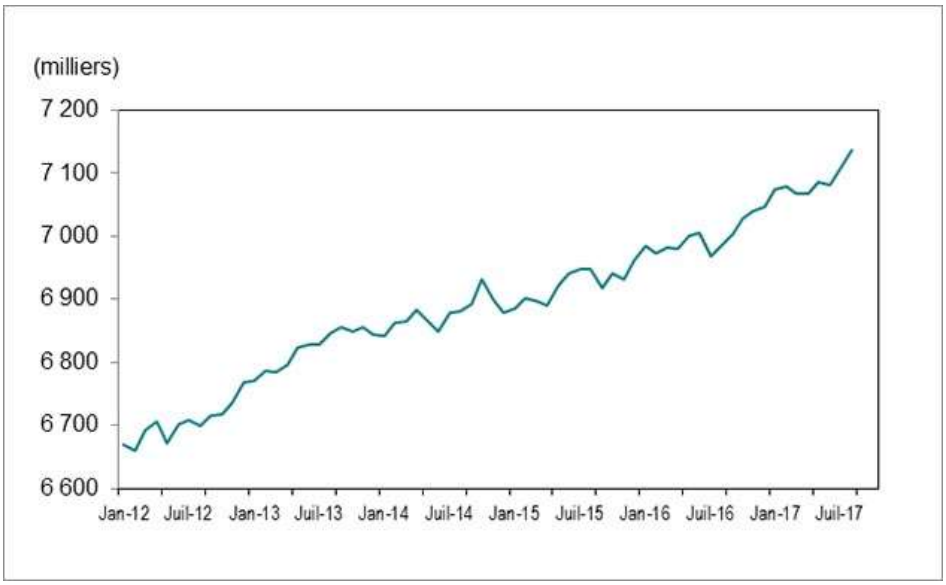


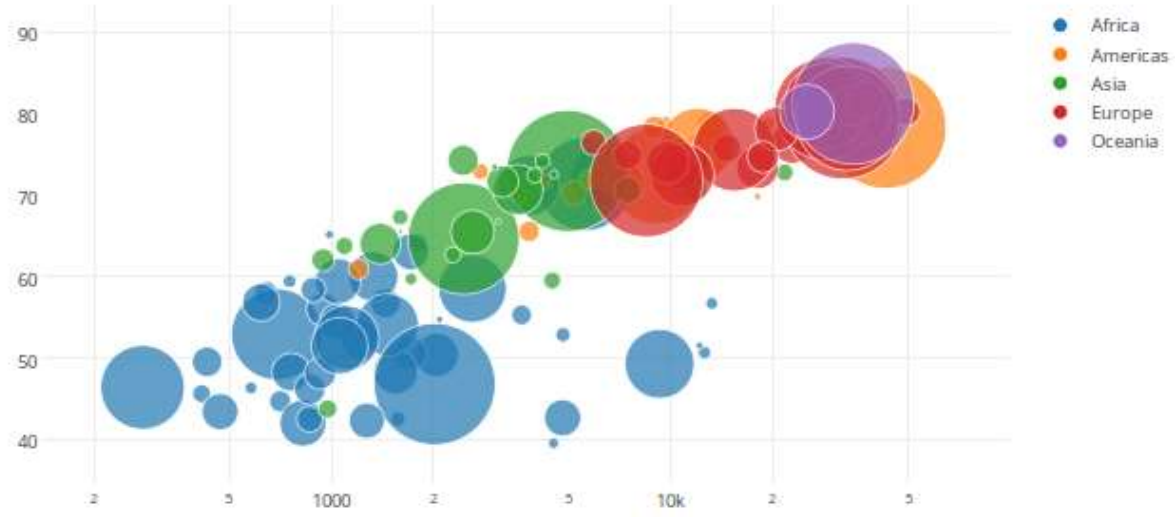
Diagramme de Gantt (Gantt chart)

Cartographie d'activitié (Heat Map)

Diagramme linéaire (line chart)



Il est à noter que chaque grand type de diagramme peut comporter des sous types. Par exemple, les diagrammes à bulles (bubble chart) représenté ci-dessous peuvent être considéré comme une sous-catégorie des nuages de points.



1.2. Etat de l'art (quasi) exhaustif des types de data vizualisation possible

2 projets se sont proposés de faire une ontologie de l'ensemble des data vizualisations existantes.

Ils se nomment « dataviz catalogue » et « dataviz project » dont les liens sont ci-dessous.

https://datavizcatalogue.com/

http://datavizproject.com/

Voici une proposition de compilation de ces 2 sources :

- Arc Diagram
- Area Graph
- Bar Chart
- Box & Whisker Plot
- Brainstorm
- Bubble Chart
- Bubble Map
- Bullet Graph
- Calendar
- Candlestick Chart
- Chord Diagram
- Choropleth Map
- Circle Packing
- Connection Map
- Density Plot
- Donut Chart
- Dot Map
- Dot Matrix Chart
- Error Bars
- Flow Chart
- Flow Map
- Gantt Chart
- Heatmap
- Histogram
- Illustration Diagram
- Kagi Chart
- Line Graph
- Marimekko Chart
- Multi-set Bar Chart
- Network Diagram
- Nightingale Rose Chart
- Non-ribbon Chord Diagram
- Open-high-low-close Chart
- Parallel Coordinates Plot
- Parallel Sets
- Pictogram Chart
- Pie Chart
- Point & Figure Chart
- Population Pyramid
- Proportional Area Chart
- Radar Chart
- Radial Bar Chart
- Radial Column Chart
- Sankey Diagram
- Scatterplot
- Semi Circle Donut Chart
- Slope Chart
- Sociogram
- Solid Gauge Chart
- Sorted Stream Graph
- Span Chart
- Sparkline
- Spiral Heat Map
- Spiral Plot
- Spline Graph
- Stacked Area Chart
- Stacked Area Graph
- Stacked Bar Chart
- Stacked Bar Graph
- Stacked Ordered Area Chart
- Stem & Leaf Plot
- Step by Step Illustration
- Stepped Line Graph
- Stream Graph
- Sunburst Diagram
- Swimlane Flow Chart
- SWOT Analysis
- Table Chart
- Tally Chart
- Target Diagram
- Taylor diagram
- Ternary Contour Plot
- Ternary Plot
- Three-dimensional Stream Graph
- Timeline
- Timetable
- Topographic Map
- Transit Map
- Tree Diagram
- Treemap
- Trendline
- Triangle Bar Chart
- Venn Diagram
- Violin Plot
- Waffle Chart
- Waterfall Chart
- Waterfall Plot
- Win-loss Sparkline
- Word Cloud

2. Etat de l'art de l'existant en matière de classification de Data vizualisation

2.1. Points généraux

Les étapes impliquées par les algorithmes de classification de diagramme sont généralement au nombre de 2 :

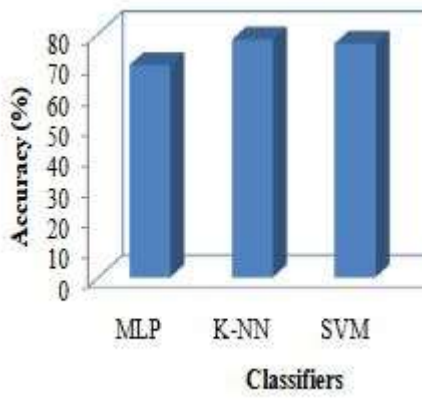
- D'abord l'extraction depuis l'image
- Ensuite la classification des données extraites

Notre but est de nous concentrer sur cette deuxième phase à travers 3 algorithmes différents de classification.

Les détails de fonctionnement sur ces algorithmes sont détaillés dans la spécification des algorithmes de classification. Nous ferons ici le focus sur leur application pour les diagrammes.

2.2. **Résultats expérimentaux**

L'expérience consiste à distinguer des Bar Chart, des line chart, des Doughnutchart, des Pie chart et d'autres diagrammes. Il existe des variations 2D ou 3D de chaque diagramme.



Précision :

Rapidité :

Classifier	Time Taken
MLP	8.38
K-NN	0.26
SVM	0.31

Taux d'erreur :

Classifier	Error Rate (%)
MLP	0.30
K-NN	0.22
SVM	0.23

1.1. Algorithme des k plus proches voisins

Rappelons que cet algorithme consiste à calculer la distance euclidienne entre le digramme entrant et à comparer ces descripteurs avec et ceux des autres diagrammes du dataset.

Malgré sa simplicité apparente, elle est la plus précise pour des diagrammes basiques tels que ceux qui sont en jeu aussi. Cette simplicité explique pourquoi elle est également la plus rapide. Elle garde néanmoins un taux d'erreur qui n'est pas optimal comparé par exemple au SVM.

1.2. Algorithme SVM

Rappelons que le principe est de tracer un hyperplan qui optimise les distances entre plusieurs classes grâce à une classification de type linéaire.

Cette méthode est sans doute celle qui donne les meilleurs résultats en termes de qualité : sa précision rivalise avec celle des algorithmes à k plus proches voisins et le taux d'erreur est le meilleur. La contrepartie de ses performances est la relative lenteur par rapport à un simple algorithme de KNN même si l'ordre de grandeur reste le même.

1.3. Algorithme sous forme de réseau de neurone

Rappelons que le principe est d'utiliser un procédé multicouche qui composé entre elles et avec des méthodes de rétro-propagation des résultats prennent en entrée un diagramme input et par apprentissage automatique sur un grand nombre de donnée sort en output la classe correspondante.

Cette méthode bien que puissante, facilement généralisable et relativement performante (les ordres de grandeurs sont les mêmes que pour SVM et KNN) présente le gros défaut d'être extrêmement lente d'un ordre de 30 par rapport aux autres algorithmes. C'est pourquoi, les réseaux de neurones doivent être utilisé pour des cas de recherches poussées ce qui est notre cas dans le projet.

Source :

https://en.wikipedia.org/wiki/Data_visualization

https://datavizcatalogue.com/

http://datavizproject.com/

https://www.researchgate.net/publication/258650813_Machine_Learning_Classification_Algorithms_to_Recognize_Chart_Types_in_Portable_Document_Format_PDF_Files

https://pdfs.semanticscholar.org/8785/6f2754451d93458fa45b8749ef1e8a55f609.pdf

https://www.yzu.edu.tw/admin/rd/files/rdso/G04/96/26/G04026(1).pdf

http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf

https://link.springer.com/chapter/10.1007/978-3-540-25977-0_8