**Interim Report due on 07/03 (Tuesday) at 3pm** but write on the 05/03 (Saturday)

Must be a single page and include:

1. Paper Title - 25/02 (Saturday) code

* Come up with a title that reflects what you plan to do in terms of analysis of the data
* Keep in mind the central theme of the paper and therefore identify the main issue
* Should be within a few sentences
* Establish the scope, context and importance of the research by summarizing current understanding, background, information about the topic, stating the purpose of the work in the form of research problem supported by hypothesis or a set of questions explaining briefly the methodology used to examine the problem, highlighting the potential outcome the study can reveal
* Begin with attractive keywords that make the paper unique (should not be answered in different papers)
* Title should be accurate, specific, unambiguous and complete
* Eliminate abbreviations unless they are well known by the target audience (the professors!!)

1. Introduction to dataset + statistical description + initial research questions

* Format
  + Holistic essay
  + Answer why the paper is being written and what it is about, its purpose is to examine the research gap and connect it with the subject, gather information and share information with the readers
  + Leading the reader from a general subject area to a particular field of research
  + 3 phases:
    - Underline the importance of the subject of the paper in the field and present mode consideration of current research and subject matter
    - Establish a role – oppose an existing belief, reveal a research gap or prepare a research question or problem – we chose research questions!
    - Outline the key features, results and a brief outline of the structure of the paper
  + Summarize all prior studies about the topic and their gaps
  + Add findings and conclusion summary
* Intro to dataset
  + What is the dataset about
  + When was it created
  + Who created it and how the data was collected => potential issues that may cause
  + Types of data
    - Nominal
    - Ordinal
    - Interval
    - Ratio
  + Data
    - Numerical
      * Continuous
      * Discrete
    - Categorical
      * Binary
      * Ordinal
  + Table
    - Attributes/Features
    - Output/Label/Class
    - Instances
    - Input and Output
* Statistical description
  + Central tendency
    - Mode
    - Median
    - Mean
  + Variability
    - Range
    - IQR
    - Average Absolute Deviation
    - Variance
    - Standard Deviation
  + Shape
    - Skewness
    - Kurtosis
  + Relationships
    - Pearson’s
    - Visualizations
      * TYPE 1
      * TYPE 2
      * TYPE 3
      * TYPE 4
      * TYPE 5
      * TYPE 6
      * TYPE 7
      * TYPE 8
      * TYPE 9
      * TYPE 10
* Initial research questions
  + Can be changed at a later stage but aim for at least 3 now
  + Specific and answerable!
  + 6 types:
    - Descriptive – summarize a set of characteristics of the data
    - Exploratory – find patterns to generate hypothesis
    - Inferential – hypothesis testing on a different dataset (COVID-19)
    - Predictive – can X predict Y (this is simple, make it more complex!)
    - Causal – does X cause Y (this is simple, make it more complex!)
    - Mechanistic – how does X change Y and why

E.g. “What is the interaction of algorithms X and Y and parameters A and B”

1. Summary of the data wrangling and pre-processing methods used – 28/02 (Tuesday) code

* NOT SURE WHAT THE SUMMARY MUST INCLUDE BUT THIS IS HOW YOU WRITE A METHODOLOGY SECTION
* Aim at the restatement of the research problem and the underlying assumptions of the study followed by stating and describing the method applied to undertake the research. Include research publications, surveys, and other research techniques to support that. (remember that the target audience are professors!)
* Identify, select, process and analyze information applied to understand the problem and therefore allow the reader to critically evaluate the study, overall validity and reliability of the study.
* Specify and explain the tools and methods used to process and analyze data and study the hypothesis
* Introduce the overall methodological approach applied for exploring the research problem that specifies whether the study is qualitative, quantitative or a combination of both and how the data was collected
* Explain the process of result analysis that determines the plan to obtain accurate assessment of each part of research with the other
* Provide a background and a rationale for using the method of analysis
* Justify the subject selected and procedure used for sampling
* Describe the limitations on the basis of literature review
* Methods:
  + Selection
    - Determine data types and how it was collected
  + Cleaning
    - Noisy
    - Outliers
    - Duplicates
    - Irrelevant (corr matrix)
    - Typos
    - Type conversions
    - Remove unique values
  + Integration
    - Format conversion
    - Especially important if combining sources
  + Transformations
    - Standardization/Normalization
    - Scaling to a range
    - Clipping
    - Log scaling
    - Z-score
  + Augmentation/Imputation (missing and incomplete data) – Use multiple!
    - Unit
    - Item
    - Hot deck
    - Mean substitution
    - Mode substitution
    - Median substitution
    - Regression
  + Reduction
    - Dimensionality reduction
    - Data compression
    - Discretization
    - Numerosity reduction
    - Attribute subset selection

Literature Review

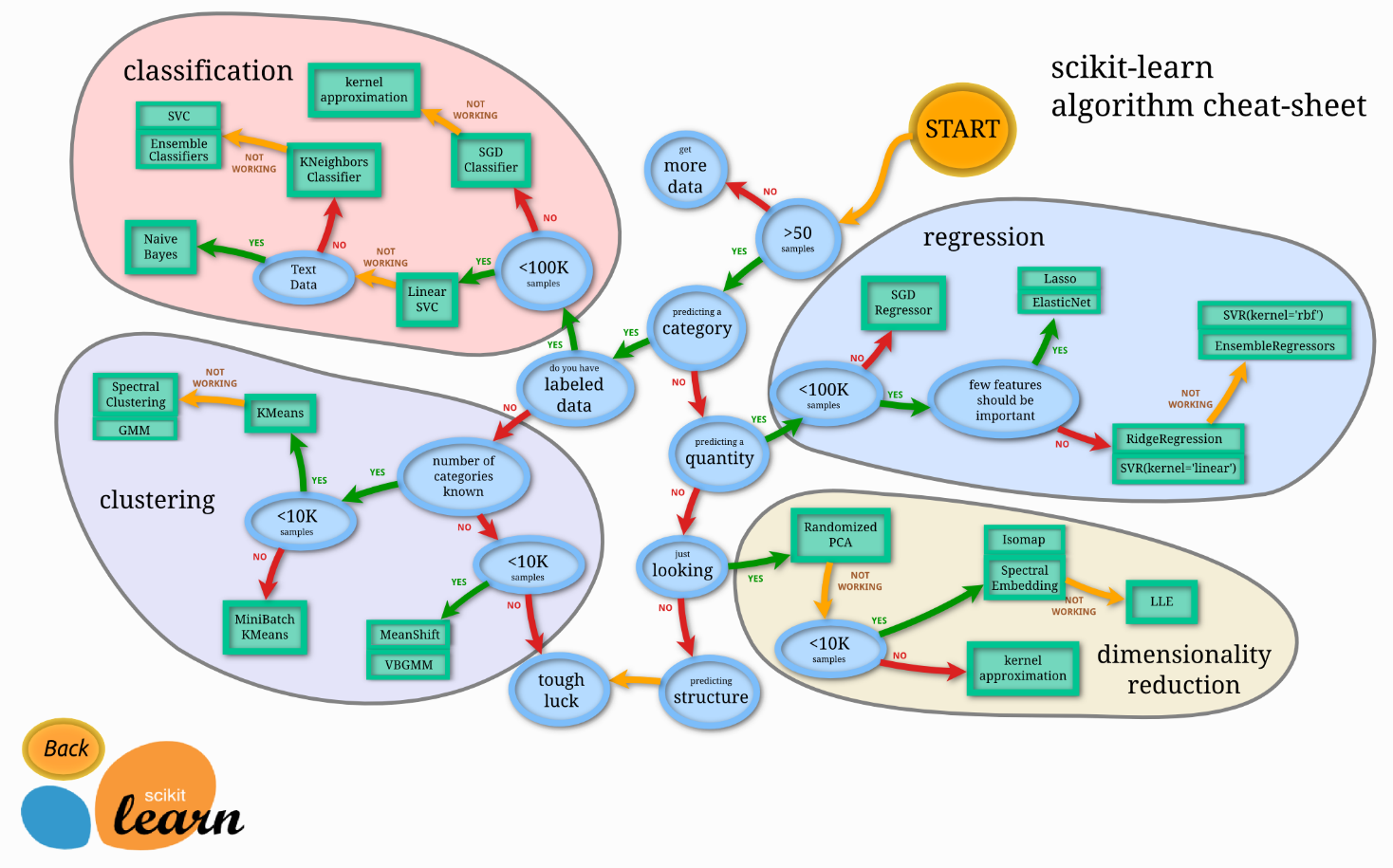
* Already conducted research in the particular field
* Determine the research gap
* How the research fits within a larger field of study
* MAY include

Text

Description automatically generated

* Summary of key sources and synthesis. Review of important information and reorganization in a way that informs the procedure to be followed in investigating the research problem
* Outline of a subject, issue and objectives
* Division on the basis of themes which supports a particular position, those against and those proposing alternative approaches
* Drawing conclusions which are best considered on the basis of argument, convincing in their opinion => making a contribution to the understanding and development of the research area
* Backed by evidence
* Cited properly
* Most important for science points of each work selected and related directly to the problem under study
* Properly systemized and synthesized
* The critical evaluation of each review should be on the basis of credentials of the author, their methodology, persuasiveness, prejudice and their arguments/conclusions
* Ways to organize it:
  + Publication Chronology – materials categorized on the basis of the date of publication
  + Thematic – organized based on themes that the present research focuses upon relating it with the topic/issue
  + Methodological – focuses on the methods utilized by the researchers in which the researcher uses the primary or secondary sources of data collection

## References



* Machine learning algorithm selection with sklearn
* <https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html>

## Next Steps

* What methods are we considering?
  + Reduction – Attribute Subset Selection (to do)
  + Reduction – Dimensionality Reduction (to do)
* Metrics to be used.
* Testing

## Findings

Dropped data

* Naïve Bayes = 60.4%
* Linear SVC using Multi Output Classifier = 67%
* Random Forest = 66.7%

Used Multilabel K-nearest neighbour but there is a bug that stops us from using it

## References

* <https://www.cdc.gov/h1n1flu/vaccination/public/vaccination_qa_pub.htm>
* <https://www.researchgate.net/publication/349186528_Machine_Learning_Based_Prediction_of_H1N1_and_Seasonal_Flu_Vaccination>
* <https://www.researchgate.net/publication/360729997_Fast_COVID-19_versus_H1N1_screening_using_Optimized_Parallel_Inception>
* <https://towardsdatascience.com/converting-data-to-a-numeric-type-in-pandas-db9415caab0b>
* https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data#:~:text=A%20classification%20data%20set%20with,smaller%20proportion%20are%20minority%20classes.