NPSC2001 FINAL REPORT

*Statistical Data Analysis and Predictive Modelling of Genomic Breast Cancer Data*

Name: Jeevanpreet Singh

Student ID: 20153780

Unit Code: NPSC2001

Degree Major: Data Science

Word Count : 3938 (Excludes captions and references)

Introduction

Breast cancer is the most commonly diagnosed cancer type in women, accounting for nearly 23% of all cancer patients and 14% of fatalities . Though this cancer is considerably researched, there appears to be no universal cure for this affliction as it continues to rise, predictably causing 1.2 million diagnoses every year (Sung et al. 2021). Throughout history it was considered untreatable and extensive surgery was applied until the 19th century, in which chemotherapy and radiation treatment currently treat all forms of cancer (Chu & Devita., 2019). Though effective, these techniques are incapable of accurate targeted therapy, relying upon hormonal characteristics to suppress the growth of cancerous cells.

In recent years, gene expression studies have identified hundreds of mRNA expressions associated with breast cancer (Maia et al., 2017). Moreover, the discovery of the PAM50 signature, a set of 50 mRNA gene expressions with classification based on mRNA profiling, has significantly improved prognosis of cancer compared to current tumour characteristic identifiers such as hormone receptors and cancer stages (Minya et al., 2019).

This discovery, coupled with the increasing presence of data science in medicine, paves the way in providing valuable insights on treatment options to health organisations who are concerned with the application of genomic data theory in the workplace (Dolezel, 2021).

Aim

This project aims to increase the relevance of PAM50 classified subtypes to determine the most accurate statistical model to predict cancer prognosis. Different datasets will be used to satisfy the notion that predictive models have different effects on different data while ensuring that the full range of data science concepts learnt throughout the year is utilised.

A further aim is to use data analysis to re-evaluate the current hormonal prognostic techniques applied in the medical workplace and making action-orientated recommendations regarding the insights given by the data. In addition, a personal aim associated with this project is to further associate my technical skills in a medical field capacity to understand insights being gathered, assisting my professional development in genomic analysis as an aspiring data scientist.

Intended Project Outcomes

Listed below are the outcomes I aim to achieve in the project, justifying the importance of this data-driven procedure in the prognostication of breast cancer.

1. I intend to apply data science knowledge to develop a statistical model for disease prediction and utilise health science knowledge and data science skills to develop new data-driven insights on breast cancer genomes. Consultation with supervisors, peers, experts and published work should be accounted for in developing knowledge of genomic analysis, accentuating the involvement of the genomic subtypes in the gene profiling of breast cancer.
2. I hope to demonstrate, through exploratory analysis, the ineffectiveness of current prognostic techniques, in which the data should display inherent flaws within the identification of hormone receptors (specifically progesterone and estrogen). This serves as motivation for predicting cancer subtypes by PAM50 signatures. This should be reinforced with principal component analysis(PCA), a method which reduces dimensionality of the large datasets by accounting for variance and generating clusters (Lever et al., 2017). With this, I aim to determine the clustering nature of PAM50 subtypes to determine any distinctive differences within them.
3. The primary outcome of this project is to identify suitable predictive models which can accurately classify the genomic subtype exhibited within each specimen. Hence, I intend to compare a multitude of predictive models, each with an individual selection purpose, and demonstrate model effectiveness in terms of accuracy and fit.

Intended Personal Outcomes

In reference to my self-assessment completed as part of the independent study contract, a multitude of issues concerning organisation, problem identification and innovative thinking were identified.

1. Improving personal weaknesses identified in the self-assessment (Singh, 2021a).
2. Innovating on previous knowledge to enhance innovative thinking
3. Enhance personal skills such as communication, implementing feedback and time-management.

1.

In reference to my self-assessment completed at the beginning of the year, the most prominent personal outcome would be to position myself in situations in which personal weaknesses can be recognised and eliminated accordingly. The most concerning issues were my organisation and teamwork capabilities as someone who has minimal experience in developing new research insights and working collaboratively with other individuals towards a shared goal. As outlined in my self-assessment, I aim to improve my organisation and time-management by setting achievable deadlines which are pre-approved by my supervisor. Moreover, reading more published work and setting concise goals according to previous genome knowledge will accentuate my personal understanding of the limitations of the project, thus improving the personal skill of understanding how far my research can determine new insights. Below are specific instances in which I aim to enhance weaknesses identified in the self-assessment.

2.

My professional development as a future data scientist is stringent upon the notion that the work I complete is done to a discretionary level of satisfaction, given the objectives laid out by my project supervisor. This is an outcome I hope to achieve in building my integrity as an aspiring data scientist as my work utilises firm, established information to generate more accurate insights.

3.

There will be tasks that I set myself to ensure that my work is done efficiently and consistently receives feedback. Through weekly meetings, I aim to make full use of my supervisor’s knowledge in predictive modelling and disease analysis, which should also satisfy another personal goal of improving my communication use of feedback. The contents of the project will be shared to peers to gauge a reaction of the work from a non-technical perspective to maintain an appropriate balance of information.

Background

Literature Review

Breast cancer is not a singular tumour, but rather a diverse set of diseases with unique biological entities that are characteristically categorised through the hormone receptors that are triggered in each individual, which inherently distinguished five forms of invasive ductal carcinomas. These are Luminal A, Luminal B, HER2-enriched, Basal-like (known as normal-like) and triple negative, all in order of best to worst prognosis (Bernhardt et al, 2016).

.

Diagram

Description automatically generated

**Figure 1** Schematic illustrating various breast cancer subtypes.

Figure 1 illustrates the four distinct cancer subtypes that have been developed through historical genetic profiling and simple receptor analysis. ER and PR refer to estrogen and progesterone respectively, both important reproductive hormones found in women. It is these hormone receptors that cancers are genetically profiled, namely immunohistochemistry (Bae et al., 2015). Luminal A and B are double-positive in both receptors while Basal-like is double-negative, with the additional HER2-receptor positive composition evident in both Luminal B and HER2-enriched subtypes. Many issues occur in this form of classification as single-positive hormone receptors (ER+/PR- or ER-/PR+) were not statistically different for cancers which are HER2 negative (Bae et al., 2015). This inability to appropriately classify Luminal A and basal-like subtypes causes opportunity for false identification, which adversely impacts targeted therapy as many subtypes share the same therapy regardless of its unique genetic composition.

The PAM50 signature, an intrinsic gene test, eliminates the issues associated with the hormone receptor prognosis, as tumours are classified on their messenger RNAs. The mRNA composition is responsible for the behaviour of the protein sets being synthesised (proteomes). It is the sequence of mRNA that identifies unique proteome behaviours relative to the molecular composition of the cancer , thus establishing the unique classification of cancer subtypes which are proven to improve prognostication of breast cancer (Bernhardt et al, 2016).

Moreover, an additional benefit of the PAM50 signature over hormone receptors is its ability to accurately predict cancer in premenopausal women, which is not possible with ER/PR tests as hormonal fluctuations in that period is minimal (Bernhardt et al, 2016).

Diagram

Description automatically generated

**Figure 2** Prognosis level and targeted treatments for each subtype

Figure 2 demonstrates that Luminal A, B and basal-like subtypes share the same hormone modulator Tamoxifen, suppressing the estrogen receptor to prevent progression of the breast cancer, while Herceptin suppresses the HER2-receptor. The issue arises within the basal-like subtype, in which its triple positive (including HER2+) composition causes immunity towards chemotherapy given its unsuppressed HER2+ hormone (Bernhardt et al., 2016). In summary, it is through the introduced gene profiling method that implores greater understanding of cancer subtype classification over current, flawed techniques which are dependent on receptor status.

Dataset/s

The dataset being utilised to analyse trends within cancer prognosis is a culmination of sample cancer patients uniquely characterised by their Cancer Genome Atlas (TCGA) identification number, composing of a dataset including factor variables regarding ‘questionnaire’ type responses regarding prognosis and a second dataset compiling all cancer patients across 12000 unique proteins (Mertins & Mani, 2016). The data was curated by Kaggle user Kajot and makes use of published proteome profiling of proteome profiling of TCGA samples collated by the Clinical Proteomic Consortium (NCI/NIH) (Kajot, 2019). The purpose of the dataset in the original study was to identify optimal clusters for each subtype, whereas the aim of this project is to utilise the same data to demonstrate accuracy of predictive models.

Methodology

Overview

To begin, the clinical dataset containing factor variables regarding stage, diagnosis age, etc. will first undergo an exploratory analysis to identify initial trends in the data. Following this analysis, data wrangling will be implemented to transform the complex dataset to merge both clinical and proteomic datasets for further analysis. Data cleaning will be utilised eliminate the thousands of N/A values and principal components will be generated to reduce the dimensionality of the dataset and identify clustering patterns within subtypes. Finally, predictive modelling will be applied on the merged dataset to determine the most accurate predictor for subtype, using cross-validation to ensure predictions are consistent.

Data Wrangling/Cleaning

Data wrangling is an essential aspect of every data science project, in which largely unstructured datasets are manipulated to conform to the objectives of the project. Initial action is to transform the proteome dataset to ensure it can be merged to the clinical data through its shared TCGA-ID column. When completed, data cleaning will omit missing values by applying a function which removes proteomes with over 5% values missing. This shortens the dataset to a more manageable and usable state for the predictive models.

Exploratory Analysis

Exploratory analysis will identify significant trends within the clinical dataset, prior to predicting the subtypes. The use of comparative visualisation will determine inconsistencies within hormone receptors and other aspects such as diagnosis age and methylation clusters to underline the significant observations between each factor. This should communicate trends within the subtypes as well as general insights regarding cancer patient characteristics. Proposed visualisations are bar-plots, boxplots, histograms and density-cluster plots.

Principal Component Analysis

PCA involves the summarising of multiple variables into features to reduce the wideness of a dataset while maintaining trends and patterns (Lever et al., 2017). This is optimally used with genomic data as it maximises variance by minimising distance, in this case compressing over 8000 protein variables into 81 principal components. By applying PCA, visualisations can be made in which trends and patterns are not lost from the main dataset, and clustering can determine any significant variations between each subtype (Lever et al., 2017).

Predictive Modelling

By utilising the CARET package and native predictive functions (Rakshit, 2021), we can determine the accuracy of many prediction models according to the classification outcome of subtypes. Partitioning training and test sets are crucial in evaluating predictive accuracy as the predictive values generated by the trained model must be compared to the true values, being the test set. An initial 65/35 split will occur, with a 3-fold cross validation to control the algorithm. This places 65% of the data into the predictive models and folds it into 3 groups as an additional splitting parameter that reproduces the accuracy outcomes three times (Brownlee, 2020).

LogitBoost:

Considering the response variable has four outcomes, this multinomial logistics regression model is optimal. It utilises a multi-value dependent variable and estimates its parameters, best-suited for predictors with distinct outcomes rather than numerical values.

Naïve bayes:

Uses Bayes theorem for conditional independence between values, hence for classification problems (Scikit, n.d.).

Random Forests:

Generates roots of a classification tree in which a class of data splits into new synthetic classes, allowing thousands of variables to be inputted for selection (Breiman et al., n.d.).

Rpart:

An extension of the random forests by applying regression and recursive partitioning on the trees, thus also effective for large datasets (Atkinson, 2019).

Neural Networks:

Algorithms which emulate human cognition, determining the best outcome given a set of results. This is efficient for multi-factor outcomes (Nielsen, 2019).

Results

Exploratory Analysis

Below are insights that were discovered by comparatively visualising the variables in the clinical data, which leads to greater interpretation of data

Chart, box and whisker chart

Description automatically generated

**Figure 3** Diagnosis Age by Subtype

Figure 3 demonstrates that both Luminal subtypes appear to be right skewed, indicating that most cases will have a higher diagnosis age on average. Basal-like, however, is left-skewed indicating a lower diagnosis age. This can imply the notion that premenopausal cancers are more likely to be Basal-like, with unpredictable hormonal compositions as mentioned in the background literature.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

**Figure 4** Boxplot of Progesterone Hormone

**Figure 5** Boxplot of Estrogen Hormone

Figure 4 and 5 demonstrate extreme similarity between the progesterone(PR) and estrogen(ER) hormone receptors. A negative status shows lower diagnosis age, indicating that the absence of these hormones may accelerate cancer growth. Moreover, basal-like being negative in both hormones may correspond to the lower age exhibited in Figure 3. This shows inefficiency in hormone receptor classification, in which single-positive receptor patients could not be characteristically diagnosed without the other. Figure 6 accentuates this discovery as fewer double-negative patients are present in the dataset, hinting at imbalance.

Chart, pie chart

Description automatically generated

**Figure 6** Pie Chart: 86% +/+, 14% -/-.

Chart, scatter chart

Description automatically generated

**Figure 7** Scatterplot of Stages classified by Subtypes

Figure 7 shows a weak positive linear correlation of tumour stages within the specific subtypes. It is expected as the tumour stages should align similarly to the AJCC classifications (AJCC has subclasses within each stage) (Kajot, 2019). However, HER2-enriched appears to be more independent with random spread rather than the other three subtypes. This raises concerns and confusion as different classifications appear to read out different stage types; for instance a stage two tumour(y-axis) being characterised as stage 3(III) in the AJCC system.

Graphical user interface

Description automatically generated with medium confidence

**Figure 8** Density plot exhibiting cluster nature of each subtype.

The clusters in Figure 8 numerically indicate abnormal miRNA/methylation compositions which are responsible for emitting tumour-suppressing genes, thus vital in cancer research (Robertson, 2016). Her2 and Luminal B subtypes appear to have similar high densities for higher methylation but intermediate miRNA, though HER2 appears tightly spread horizontally while Luminal B is well-rounded. Luminal A appears distinguished with high levels at lower methylation, and though Basal-like is not reproducible, it’s points indicate a higher methylation tight spread similar to HER2.

Chart, radar chart

Description automatically generated

**Figure 9** Density plots of clustering within progesterone status.

When compared to Figure 9, it appears that subtypes are far easier to distinguish clusters rather than hormone receptors, as progesterone positive and negative patients have extremely similar density clusters, proving hormone-dependent cancer prognosis to be more difficult. Essentially, similarities cause difficulty prognosis while unique compositions are far simpler to classify.

Principal Component Analysis

Chart, histogram

Description automatically generated

**Figure 10** Scree plot of first 15 components

Figure 10 displays the first 15 components of the PCA, in which each percentage highlights the variation covered by that specific components. These reflect portions of the variation within the dataset where a single row covers the variation of multiple proteome variables in the original data.

Diagram

Description automatically generated

**Figure 11** PCA Cluster plot (PC1 vs PC2)

Chart, bubble chart

Description automatically generated

**Figure 12** PCA Cluster plot (PC3 vs PC4)

Figure 11 demonstrates the clustering of subtypes within the first two components. While accounting for 21% variation, it appears that all clusters have unique and distinct positions and shapes which indicate certain variation differences between each subtype. Basal-like is generally at a higher variation, while all other subtypes are spread in different directions. Figure 12, using components 3 and 4 for 12% of the variation, confirms distinct clustering for HER2, though all other subtypes appear similar. This reflects the poorer variation of components being used, though still justify some difference for the HER2+ subtype.

Chart

Description automatically generated

**Figure 13** PCA Cluster plot of Progesterone Status(PR)

Figure 13 demonstrates the clustering of progesterone hormonal status with 21% variation. There is an identifiable difference in cluster pattern between PR+ and PR- patients, however, there is enough overlap to notice that the subtype components were far more differentiated than the hormonal components displayed above.

Predictive Modelling

To interpret the following results, the confusion matrix plots demonstrate model-predicted values against the actual values in the test set to determine whether the model can predict the referenced test values. Values in the top left to bottom right diagonal indicates accurately predicted values (Rakshit, 2021), as matched between the predicted and true reference values. The ROC curve identifies balance of specificity and sensitivity of the model. Lower specificity and higher sensitivity indicate better model fitting, hence a higher AUC (area under curve) value reflects better model fitting (Zou et al., 2007).

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

**Figure 15** ROC curve of NNet model.

**Figure 14** Confusion Matrix of NNet Prediction

Chart

Description automatically generatedChart, line chart

Description automatically generated

**Figure 16** Confusion Matrix of Naïve Bayes Prediction

**Figure 17** ROC curve of Naïve Bayes model.

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

**Figure 19** ROC curve of Random Forest model.

**Figure 18** Confusion Matrix of Random Forest Prediction

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

**Figure 20** Confusion Matrix of logBoost Prediction

**Figure 21** ROC curve of logBoost model.

Square

Description automatically generatedChart, line chart

Description automatically generated

**Figure 23** ROC curve of Rpart model.

**Figure 22** Confusion Matrix of Rpart Prediction

Figure 20 displays 100% accuracy for the LogitBoost model with an AUC value of 0.537. This is a significant discovery as perfect accuracy is generally unattainable with extremely large datasets with thousands of variables. This is followed by the neural network model in Figure 14 with 1 incorrectly predicted value and a higher AUC of 0.56. Though Rpart appears the least accurate (Figure 22), the AUC value of 0.68 demonstrates its reliability as a model for the classification of subtypes. The observer should account for the fact that the best predictive models will not always be the most reliable, therefore hinting to a balance between accuracy and optimal fit (Zou et al., 2007).

Chart, box and whisker chart

Description automatically generated

**Figure 24** Dotplot of Model Accuracies

Figure 24 demonstrates accuracy according to the cross-validation applied, in which the average and spread of accuracy are determined by three separate folded training sets run across the same model. On average, cross-validation confirms that logBoost remains the most accurate predictive model, though the spread of random forest and Rpart models show potential to be more accurate than logBoost.

Personal Success

Throughout this project, I had acknowledged and identified key weaknesses mentioned in my self-assessment, all of which were improved through specific situations encountered throughout the year (Singh, 2021a).

Technical Skills - Solved Problem Research Issues

I enhanced many new technical skills while undertaking this project, specifically relating to predictive modelling and variance estimation of principal components. PCA and the CARET package were both introduced to me by my supervisor and I was able to find published research to efficiently choose the most appropriate models while communicating the results in a non-technical capacity (Rakshit, 2021). Technical skills were not a weakness highlighted in my self-assessment, but the activity of researching enhanced my technical problem solving and I’m grateful to learn new predictive algorithms which are usable for future prediction analyses.

Interpersonal Skills – Solved Feedback and Communication Issues

The most significant skill developed throughout this year was implementing given feedback, evidence of this improvement being a data science peer recommending I facet the plot as shown in Figure 7, in which 4 separate plots can be combined for simpler comparative analysis. Moreover, a feedback received from my supervisor was the requirement for cross validation to ensure that predictions were not coincidental. By applying this, I could further justify the accuracy of logBoost and other models. Communication was another weakness which was improved through weekly meetings with my supervisor. Through these meetings, I gradually found new ways to concisely represent my project outcomes, one which satisfied the expectation of my supervisor and personal intended outcomes of this project. A piece of evidence reflecting this improvement would be the fact that I transformed my aim to encompass more aspects of exploratory and principal component analysis rather than focusing entirely on predictive modelling.

Management Skills – Solved Organisational Issues

The easiest weakness to eliminate was my time-management skills. The self-assessment mentioned poor management of timelines, however this was solved by setting achievable goals which were pre-approved by my supervisor. Deadlines were all presented in an ePortfolio (Singh, 2021b), thus making it simpler to carve out time to complete certain tasks. The portfolio implementation assisted me in completing all my assessments on time while developing the project outcomes in a proficient manner.

Aspirations/Future Inquiries

This project has furthered my aspiration to become personally involved in this field of genomics, one which significantly benefits from the application of data science and complex statistical methods. Medicine is of a high interest to my personal development, thus applying analytical procedures to benefit public health would be significant in my future. I have begun building on these aspirations by signing up BioInfoSummer event in the summer, a workshop discussing new genomic innovations, as well as applying for a genomic data project offered by Curtin School of Sc/Eng. Without this project, I would have never involved myself in these opportunities.

I’ve posed the question of how far I aim to take the results of this project as a form of future inquiry. In terms of my career development, I aim to find new opportunities to enhance my knowledge of cancer genomics and the specific statistical methods tailored to solving medical problems. Moreover, I may utilise this work in future career projects which require knowledge of genomic data or simple predictive modelling.

Conclusion

In summary, exploratory analysis achieved the objective of distinguishing certain flaws within the hormone receptor variables, in which similar spread and inconsistent tumour stages highlight the issues regarding these current prognosis techniques. PCA achieved the aim of reducing the dimensionality of the data to demonstrate distinct clustering differences between each subtype, in which more different clusters indicate easier classification of PAM50 subtypes. Finally, logBoost was determined to be the most accurate predictive model for subtypes, hence imploring further investigation into logistics regression modelling for future breast cancer prediction.

Recommendations

Through the findings of this project, combined with contemporary literature, I unequivocally recommend clinical practices to apply PAM50 genetic subtypes for genetic profiling over current hormone receptor techniques. Cancer research should focalise on genetic profiling of mRNA subtypes to ensure that therapy is not restricted to the presence of hormonal fluctuations, which is in turn far more difficult to identify in premenopausal women. An immediate action would be to further analyse basal-like subtypes to analyse its lower diagnosis age and whether its triple positive hormonal composition contributes to earlier development of cancer. Further action in regards to analysing individual subtypes would be to further extend knowledge on the PAM50 signature which in turn leads to better treatment decisions. Although PAM50 subtyping is currently the only accepted molecular identification system for breast cancer, there is limited literature and research in terms of developing more consistent classifications (Raj-Kumar et al., 2019). By conducting more research, this action will enhance profiling systems currently used in healthcare, providing more accurate analysis of breast cancers.

In regards to the statistical methods applied, medical fields should encourage greater use of data science in the analysis of genome data, posing the knowledge of prior prognosis techniques against new innovative genomic classification, verified by cross-validated predictive modelling. There is a distinct absence of data-proficient individuals within health science, in which the only analytical tools are used for organising data storage and implementing cloud-based software for user experience. (TheAppSolutions.com, n.d ). Health practices should take action to utilise statistical methods in identifying data-driven solutions that will inevitably assist healthcare workers in identifying patient cancer outcomes and the targeted treatments that are recommended (Raj-Kumar et al., 2019).

Future Work/Discussion

In addition to the future inquiry raised within my personal successes, there are multiple pathways to conduct further research with the key findings developed in this project. Most importantly, the use of a larger dataset and computer with higher processing power will give a more reflective prediction of breast cancer afflicted patients. Although 12000 proteomes are quantified, only eighty samples were used in predicting the subtypes. Furthermore, a future project using additional cancer-related time data should assist in producing a survival analysis, using a Kaplan-Meier plot to determine survival chance over a period of time (Etikan et al., 2017). This is a useful further analysis as it justifies differing survival rates depending on cancer type, confirming that the PAM50 subtypes are sufficiently distinct and may produce different survival rates beginning from the onset of the illness. Along with the additional data, an additional future project could include samples of the Triple-Negative subtype (See Figure 2). It is the most fatal form of cancer and is not displayed within the dataset used in this project. The triple negative nature further accentuates the lack of prognosis capability that hormone receptors offer as all three (ER, PR, HER2) hormones are absent. A further project regarding the PAM50 triple negative subtype will develop insight into a hardly-researched and largely incurable form of breast cancer. In summary, these are all future project opportunities that will improve the consistency of current cancer profiling and develop further insight into lesser-known forms of cancer within society.

The key findings of this project should implore discussion regarding a critical re-evaluation of current cancer profiling techniques such as hormone receptors and simplified tumour stage classification. The culmination of discoveries in this project raises further concerns regarding the use of data-driven solutions within health practices. Current computational application in medicine is underutilised and there should be further discussion regarding the involvement of predictive models to determine patient outcomes. In specific, raising discussion in the logistic regression modelling will prove useful considering its perfect accuracy and may be used to identify certain cancer subtypes going forward.

**REFERENCES:**

Atkinson, B. 2019. Package ‘rpart’. *CRAN Packages.* <https://cran.r-project.org/web/packages/rpart/rpart.pdf>

Bae SY, Kim S, Lee JH, Lee HC, Lee SK, Kil WH, Kim SW, Lee JE, Nam SJ. 2015. Poor prognosis of single hormone receptor- positive breast cancer: similar outcome as triple-negative breast cancer. *BMC Cancer.* <https://pubmed.ncbi.nlm.nih.gov/25880075/>

Bernhardt, S., Dasari, P., Walsh, A., Timothy, J., Ingman, VW. 2016. Frontiers in Oncology Volume 6. <https://www.frontiersin.org/article/10.3389/fonc.2016.00241>

Breiman, L., Cutler, A. n.d. Random Forests. *Berkely.edu.* <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>

Brownlee, J. 2020. A Gentle Introduction to k-fold Cross-Validation. *Machine Leaning Mastery*. <https://machinelearningmastery.com/k-fold-cross-validation/>

Chu, E. Devita, VT. 2019. Physicians' cancer chemotherapy drug manual 2019. *Worldcat.org.* <https://www.worldcat.org/title/physicians-cancer-chemotherapy-drug-manual-2019/oclc/1080638311>

Dolezel, D., McLeod, A. (2021). Big Data Skills: Bridging the Data Science Theory-Practice Gap in Healthcare. Webpage: <https://perspectives.ahima.org/big-data-skills-bridging-the-data-science-theory-practice-gap-in-healthcare/>

Etikan I, Abubakar S, Alkassim R. 2017. The kaplan meier estimate in survival analysis. Biom Biostat Int Journal. Accessed through DOI: 10.15406/bbij.2017.05.00128

Kajot, 2019. Breast Cancer Proteomes. *Kaggle.com.* <https://www.kaggle.com/piotrgrabo/breastcancerproteomes/metadata>

Lever, J., Krzywinski, M., Altman, N. 2017. Principal component analysis. *Nature.com.* <https://www.nature.com/articles/nmeth.4346>

Maia, A.T., Sammut, S.J., Jacinta-Fernandes, A., Chin, S.F., (2017). Big data in cancer genomics, Current Opinion in Systems Biology, Volume 4, Pages 78-84, ISSN 2452-3100, <https://doi.org/10.1016/j.coisb.2017.07.007>

Mertins, P., Mani, DR. 2016 Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature.com*. <https://www.nature.com/articles/nature18003>

Minya, P., Messer. K. Davies, SR., Vickery, TL. 2019. Research-based PAM50 signature and long-term breast cancer survival. *NCBI.* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6985186/>

Nielsen, M. 2019. Using neural nets to recognize handwritten digits. *Neural Networks and Deep Learning.* <http://neuralnetworksanddeeplearning.com/chap1.html>

Raj-Kumar, PK., Liu, J., Hooke, J.A. et al. 2019. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Sci Rep 9*. <https://doi.org/10.1038/s41598-019-44339-4>

Rakshit, S. 2021. Predictive Analytics Using CARET. *Blackboard.* Accessed through lms.curtin.edu.au.

Robertson, S. 2016. What is DNA Methylation. *News Medical.* <https://www.news-medical.net/life-sciences/What-is-DNA-Methylation.aspx>

Scikit. N.d. 1.9 Naïve Bayes. <https://scikit-learn.org/stable/modules/naive_bayes.html>

Singh, J. 2021a. Independent Study Contract.

Singh, J. 2021b. NPSC2001 Portfolio. *Microsoft OneNote* <https://studentcurtinedu-my.sharepoint.com/:o:/g/personal/20153780_student_curtin_edu_au/EpvOiFxfy3pFv4JbX2oJ9tYBjmtKYjWciaXgj9nJ2W0ZhA>.

Sung, H., Ferlay, J., Siegel, R., , Laversanne, M. Soerjomataram, I., Jemal, A., Bray, F. 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *A Cancer Journal for Clinicians. Page 209-249.* [*https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21660*](https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21660)

TheAppSolutions.com. n.d. HOW DATA SCIENCE IS CHANGING HEALTHCARE. *TheAppSolutions Blog.* [*https://theappsolutions.com/blog/development/data-science-healthcare/*](https://theappsolutions.com/blog/development/data-science-healthcare/)

Zou, KH. O’Malley, J., Mauri, L. 2007. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation Volume 115.* [*https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.105.594929*](https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.105.594929)

**List of Figures:**

Figure 1. from Oxford University Press. *Lab Med*, Volume 41, Issue 6, June 2010, Pages 364–372, <https://doi.org/10.1309/LMLIK0VIE3CJK0WD>

Figure 2. from BioRender.com. by A. Lazaratos, n.d., BioRender. <https://app.biorender.com/biorender-templates/figures/5f15ea4438c5ef002876ab4b>

Figure 3-24: Reproduced through R-studio. Follow repository.

**Data Repository:**

<https://github.com/jeevandatafreak/NPSC2001> (Follow README comments).