SENG 474: Data Mining - Formal Proposal

Group I - Team Members (Undergraduate)

Cameron Lindsay, Claire Champernowne, Ethan Getty, Elliot Kong, Anton Nikitenko
Note: Anton Nikitenko is no longer taking the course but did contribute to this body of work.

1. The Problem

Head and neck cancers (HNC) are the sixth leading cancer worldwide, with over 600 000 new cases per year [1]. While the overall incidence of HNC has decreased, oropharyngeal squamous cell carcinomas (OPSCC), a subset of HNC, has increased due to human papillomavirus (HPV)‑mediated oncogenesis [2-4]. Fortunately, HPV‑positive OPSCC have shown to respond better to treatment, resulting in significantly higher survival outcomes [5-7]. Overall survival rates have been shown to be significantly lower in patients with lower socioeconomic status, where factors such as income, education, lifestyle, and access to health care services may impact their quality of treatment [8-11]. While Canada's socialized health care system may mitigate some of the economic disparities encountered by other countries [12], later stages of presentation at diagnosis are still associated with lower socioeconomic status which results in lower survival outcomes[13].

Prognostic models of survival are an integral resource for physicians in the clinical strategy and management of disease. They are evidence-based decision tools that have been supported by prior diagnostic and treatment information [14]. The Kapalan-Meier estimate is the current gold standard utilized by health care providers in determining patient survival following clinical intervention (or lack thereof) [15]. However, as computational power increases, prognostic models are becoming increasingly more accurate and complex, resulting in their progression to the forefront of clinical decision making. Artificial intelligence-based strategies including Bayesian networks, neural networks, and support vector machines are rapidly gaining popularity in the creation of these models [14, 16].

Table 1.1: Features and Classifiers of the Dataset

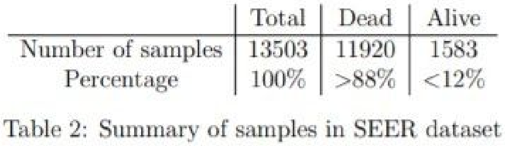
Feature	Definition
Age	Age of the patient (provided in a range).
Sex	Sex of the patient.
Year of Diagnosis	Year of cancer diagnosis.
Race and Origin	Race and genetic lineage of the patient.
Primary Site	Location of the main (primary) tumor's point of origin.
Grade	Description of a tumor based on cell morphology. As they increase in grade, cells are increasingly more abnormal and cancerous. Grade I: cancer cells that resemble nearby tissues and aren't

	growing rapidly, Grade II: cancer cells that look abnormal and are dividing rapidly, Grade III: cancer cells that look abnormal, divide rapidly, and are likely to spread to nearby tissues.
ICD-O-3, Histology, and Behavior	ICD code, histological description of the tumor sample, and its level of aggressivity (behavior). Note: NOS = not otherwise specified.
Median Household Income (Inflation Adjusted to 2018)	Patient's household income adjusted to 2018 level of inflation.
Combined Summary Stage	Categorizes how far a cancer has spread from its point of origin.
Tumor Size	Size of primary tumor (mm).
RX Summary -- Surgical Primary Site	Codes related to specific surgeries to the primary tumor site. 0 = no surgery performed or diagnosed at autopsy, 10-19 tumor destruction with no pathologic specimen, 20-80 resection with pathologic specimen, and 90 surgery performed but not otherwise specified.
RX Summary -- Scope of Regional Lymph Node Surgeries	Surgery to lymph nodes.
RX Summary -- Surgery to Other Regional or Distant tumors	Surgery to any secondary tumors.
Reason for No Cancer-Directed Surgery	Whether a surgery was performed or not recommended.
Sequence Number	Number and sequence of known tumors occurring over the patient's lifetime.
Classifiers	
Death Classification	Whether the patient died due to the cancer diagnosis or lived through the cancer diagnosis.

Using the National Cancer Institute's (NIH) Surveillance, Epidemiology, and End Results Program (SEER) (https://seer.cancer.gov/) databases to train potential algorithms, we propose the formulation of a prognostic model of OPSCC survival according to patient socioeconomic status. As socioeconomic status is a complex factor that cannot be formally detailed by the SEER database, patient income will be used as a surrogate marker. The SEER database is based out of the United States of America, and theoretically should display the same or greater disparities in survival outcomes mentioned above. Please note, this project has been altered from the initial proposal due to the limitations of the data available within the

SEER database. These limitations are detailed in the initial results section below. While a formal model for use has not yet been identified, several models will be created and compared to a Kapalan-Meier estimate using the same dataset. The observed dataset was collected from 18 registries between 2000 and 2017 with malignant tumors (according to ICD-O-3 behavior classifications). While larger datasets taken from a longer timeline are available, cancer treatment protocols have greatly improved and survival outcomes have improved accordingly. By choosing a narrower timeline, survival outcomes will be less dependent on the year of diagnosis as a predictor of survival. As outlined in Table 1.1, 15 features were observed that lead to a binary classifier of either “Dead (attributable to this cancer dx)” or “Alive or dead of other cause”. Due to the dataset’s original ratio of 11920 “Dead” to 1583 “Alive or dead of other cause”, this resulted in a default high prediction quality for selecting “Dead” for all data points (Table 1.2). To resolve this, 1583 samples of each outcome were selected and reshuffled into a modified dataset containing a sample size of 3166 patients.

Table 1.2: Summary of Samples in SEER Dataset



	Total	Dead	Alive
Number of samples	13503	11920	1583
Percentage	100%	>88%	<12%

Table 2: Summary of samples in SEER dataset

A major bottleneck in acquiring useful data for this research was being approved by the SEER program to access such resources. All team members are required to sign a data-use agreement which stipulates that the data must only be used for the purpose of the research and allow adequate time for processing. Furthermore, access to databases containing useful datasets pertaining to the treatment of HNC (head and neck cancers) with HPV status require further signing of data-use agreements due to the nature of data, as identifiable information may be contained in the dataset. Processing this agreement has occasionally taken up to several weeks to become approved, as the maintainers of the dataset must ensure that the researchers are using this data for academic purposes only.

As the data available for our analyses currently stands, it does not include various lifestyle and disease-specific risk factors such as smoking or HPV status. Additionally, detailed patient histories including medications and genetic backgrounds are not available. The absence of these factors remains a glaring limitation in our study and will likely hinder the applicability of any models made, regardless of the cancer observed [1, 3]. However, the sample sizes available to our study from this database are larger than any other that is readily available, allowing for a far higher potential accuracy from the observed factors. The database remains valuable from an educational perspective, but a more thorough database would be required for practical utilization. Therefore, the problem has shifted to account for the possibility of encountering insufficient time for acquiring specific datasets on OPSCC. The problem can be potentially resolved by expanding the scope to consider other datasets that may be more readily available. Unfortunately, access to similar databases will likely rely on an agreement process comparable to that of the SEER program, as the risk of identifiable patient information will always be a factor. Requests to other databases including the National Cancer Database (https://www.facs.org/quality-programs/cancer/ncdb), which provides freely available data for use on similar issues, will be made to attempt obtaining better quality data. However, the time restriction of this project will likely make registration to the database an exercise in futility. It is likely that we will address the problem in such a way that our methods will be geared towards scaling to datasets for OPSCC without having to perform extensive refactoring of our experiments.

2. Goals

Given the Kaplan-Meier is considered a gold standard for predictive measures or survival [15], future experimentation will look at methods of comparing our developed machine learning models to the final Kaplan-Meier estimate. Cross validation techniques such as k-fold methodology will likely be implemented to test the predictive power of our models to the final Kaplan-Meier curve. Due to the imbalance of the testing set, using accuracy as the metric of performance is not very informative. Evaluation is performed with Area Under the Receiver Operating Characteristic Curve (ROC-AUC), F1 score (F1), and weighted F1 score (F1-weighted) . The best model will be chosen based on the maximum F1-weighted. The accuracy (ACC) will be reported as a comparison. Additionally, as income is a focus of this study, separate survival curves observing several income brackets will be developed using the Kaplan-Meier method. The potential of comparing our models to other methods of survival analysis, such as the Cox Proportional Hazards method, may also be implemented if time permits.

3. Plan

3.1 Kaplan-Meier:

The Kaplan-Meier estimate was developed using Scikit-survival analysis libraries developed from a modified version of the dataset described above. Modifications to the dataset were only the inclusion of the "Survival months" parameter, calculated from Equation 3.1.1:

Equation 3.1.1: Survival Months Calculation [17]

$$\text{Survival Months} = \text{floor}((\text{date last contact} - \text{date dx}) / \text{days in a month})$$

Due to the survival months calculation being based on the date of last contact (date follow ups have stopped) versus actual survival times, otherwise known as right-censored data. This type of data would likely impact the predictive accuracy of potential machine learning-based models and therefore was excluded from their developmental process. However, the analytical tools utilized in the generation of the Kaplan-Meier estimate take this into account. A preliminary estimate can be seen in Section 5 of this document. Since we aim to look at income as a potential predictor, future Kaplan-Meier estimates will contrast and compare various income brackets to their probability of survival. However, this will be done through a different library, as Scikit-survival lacks certain features for optimization and proper testing.

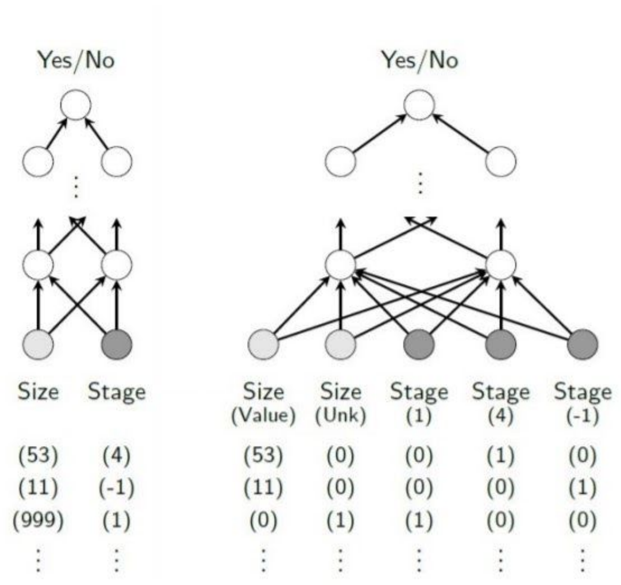
3.2 Decision Tree:

Various parameters for a decision tree model were tested to achieve maximum accuracy. As well, pruned vs unpruned trees were tested at various dataset sizes to determine the optimal model for our dataset. F1 and mean squared error were the metrics used during experimentation to measure success. F1 score was included to ensure accuracy was not simply reflecting the ratio of samples.

3.3 Neural Networks:

Tests were performed using a multilayer perceptron (MLP) with various changes to the number of hidden layers and variable alpha (α). The different values used for the hidden layer were 3, 7, 13, and 21 neurons. For the values of α, the values used were 1, 0.1, 0.01, and 0.001 given 4 neurons in the hidden layer. From the results of the training data, the more neurons that were used in the hidden layer, the higher the accuracy. The lower the value of α, the higher the accuracy was as well. When looking at the test data, however, the highest accuracies were when there were 7 neurons in the hidden layer, and an α of 0.1.

A survival prediction was performed using MLP with back-propagation and log loss function. Artificial neural networks (ANN) are one of the most common approaches in survival analysis and MLPs are often applied as a powerful tool for prediction and classification problems. In particular, we aim to investigate whether MLPs are capable of learning from our SEER data and predicting a dead to alive ratio with an acceptable level of accuracy.
First, two different encoding techniques were applied to categorical variables in the original (raw) data: ordinal encoding and one-hot encoding. Ordinal encoding transforms categorical features to ordinal integers that range from 0 to $n_{categories}$ (number of categories) -1. One-hot encoding transforms each category to a binary column. One-hot encoding can avoid the model from assuming a natural ordering between categories that do not exist. Two MLP models were applied to two datasets that are encoded from the raw data using different encoding techniques (Figure 3.3.1a, b).



(a) MLP

(b) MLP + OneHotEncoder

Figure 1: Illustration of MLP models

As mentioned previously, the original dataset we used is incredibly imbalanced (Table 1.2). To handle the imbalance, an oversampling and an undersampling technique was applied to the training set before being fed to MLP models. Synthetic Minority Oversampling Technique (SMOTE) is a popular oversampling method that increases the minority samples by interpolating between a sample and its k-nearest neighbors. A Random Undersampling technique is applied to shrink the majority samples. The effect of these two techniques was verified on two MLP models (Figure 5.2.4).

Hyperparameters for MPL models were tuned using a 5-fold cross-validation strategy. In particular, for MLP models, the number of hidden layers (1 ~ 3), number of neurons in each layer (20, 50, 100, 200), learning rate (0.0001 ~ 1) and L2 penalty factor α (0.01 ~ 100) were tuned (Table 5.2.1).

As shown in Table 5.2.4, most of the metrics are low for all four models. This could be possible due to the imbalanced dataset, or inappropriate model configurations. In particular, the threshold of logistic regression will be tuned and testified by ROC-AUC to find a better equilibrium of which sample would be concluded as “dead/alive”. Furthermore, one or more feature selection algorithms will be implemented in an attempt to reduce misleading and redundant data and increase the performance.

4. Task Breakdown

Cameron Lindsay:

- Collected and cleaned data, parsed relevant features, experimentation using Kaplan Meier estimator
- Going to continue modifying Kaplan-Meier estimates and research into proper cross-validation techniques, estimate hazard rates using Nelson-Aalen
- Look into an SVM-based approach

Claire Champernowne:

- Cleaned and modified data
- experimentation on decision tree models with/without reduced error pruning
- Will pursue logistic regression based approach

Ethan Getty:

- Experimentation on multi-layer neural networks for accuracy and loss
- Going to continue validating neural network learning methods

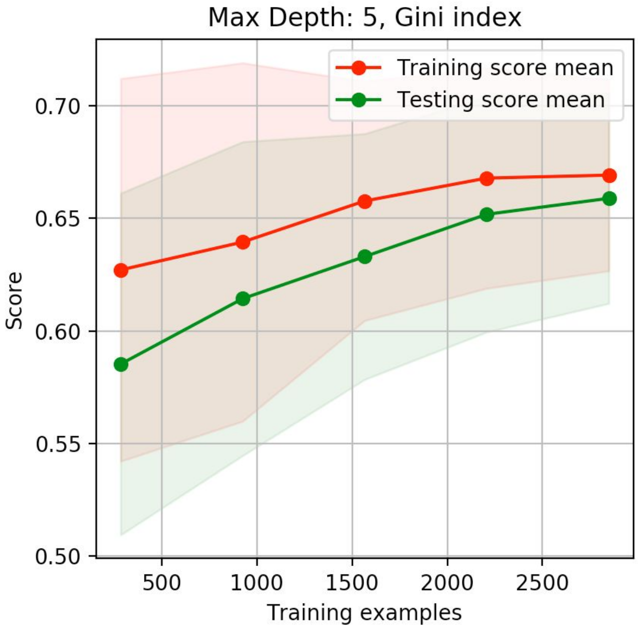
Elliot Kong:

- Experimentation on preprocessing with ordinal and one-hot encoding, resampling with SMOTE and RUS strategies, different neural network structures, tuning learning rate and L2 penalty
- Going to verify different feature selection methods and complicated neural network structures

5. Initial Results

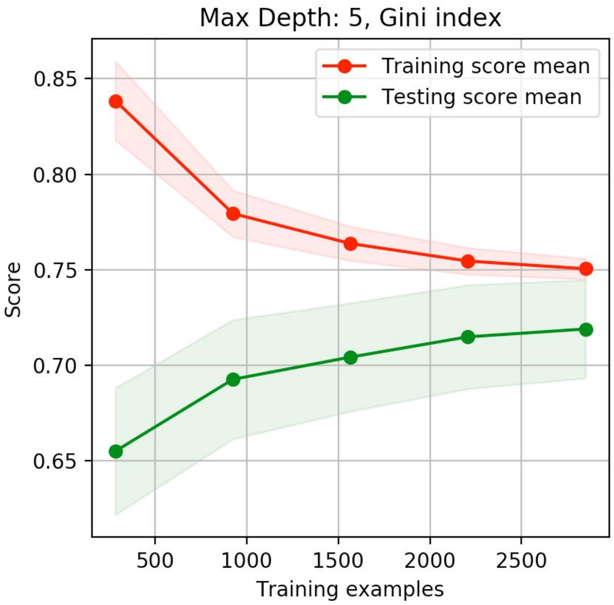
5.1 Decision Tree:

Maximum accuracy for both decision tree models were achieved using Gini index and a maximum tree depth of 5. Over smaller sample sizes, the pruned tree performed better because it is less prone to overfitting; however a higher accuracy of approximately 72% was achieved by the unpruned tree when using our full dataset. Because other methods consistently achieve a higher accuracy than decision tree models, our following tests will focus on Kaplan-Meier estimators and Artificial Neural Networks.



Results using full dataset:
Mean Squared Error: 0.299
F1: 0.654

Figure 5.1.1: Reduced error decision tree model over different dataset sizes



Results using full dataset:
Mean Squared Error: 0.285
F1: 0.714

Figure 5.1.2: Unpruned decision tree model over different dataset sizes

5.2 Neural Networks:

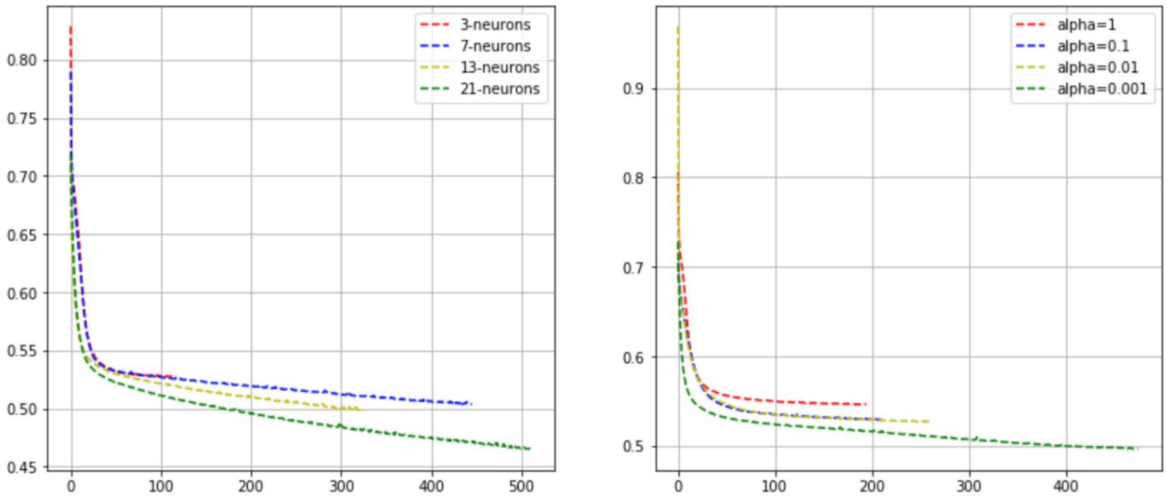
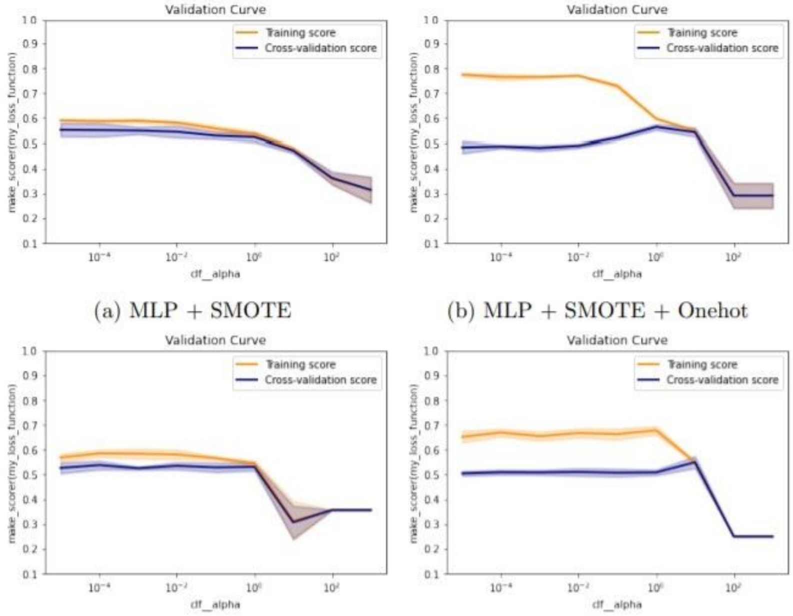


Figure 5.2.1: Loss curves for perceptrons given varying neurons (left) and alphas (right).
From this graph we can see that the loss decreases to a point of stability, therefore our processed data allows the perceptron to learn at an acceptable rate.



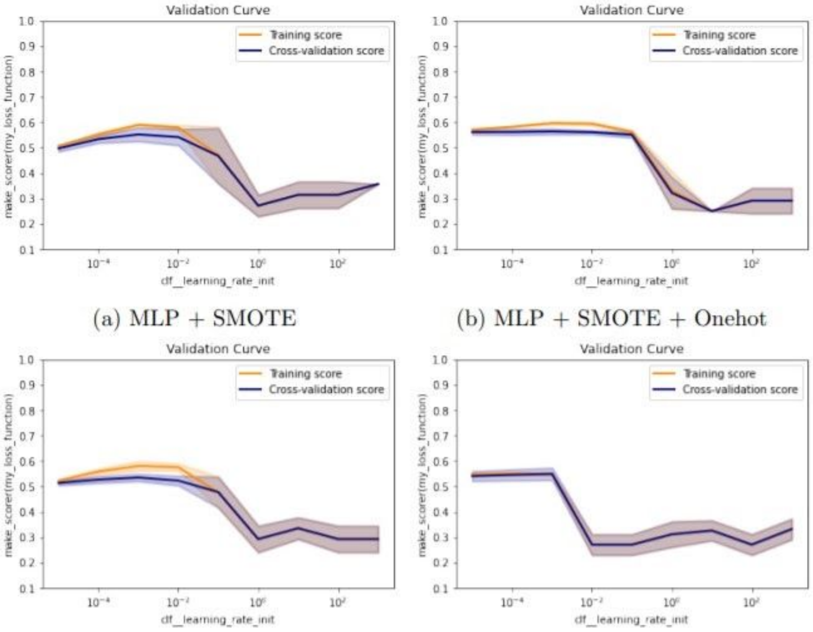
(a) MLP + SMOTE

(b) MLP + SMOTE + Onehot

(c) MLP + RUS

(d) MLP + RUS + Onehot

Figure 2: Validation curves of tuning L2 penalty factors



(a) MLP + SMOTE

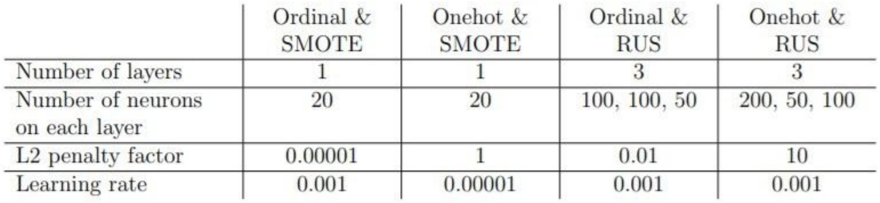
(b) MLP + SMOTE + Onehot

(c) MLP + RUS

(d) MLP + RUS + Onehot

Figure 3: Validation curves of tuning learning rate

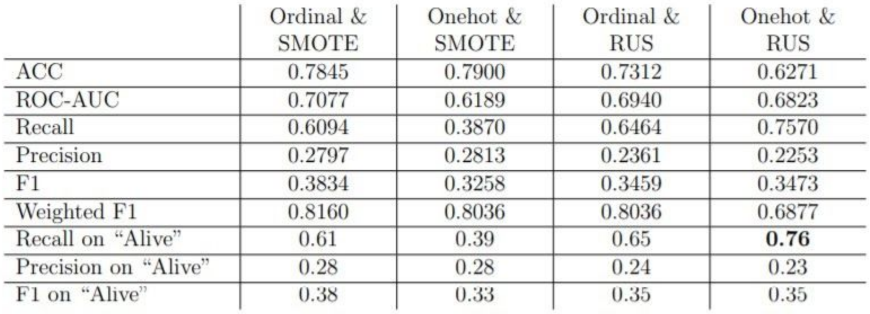
Figure 5.2.2 and 5.2.3. Models with One-hot encoding show clear over-fitting as tuning L2 penalty factor and hardly any as tuning learning rate. The models without one-hot are quite opposite.



	Ordinal & SMOTE	Onehot & SMOTE	Ordinal & RUS	Onehot & RUS
Number of layers	1	1	3	3
Number of neurons on each layer	20	20	100, 100, 50	200, 50, 100
L2 penalty factor	0.00001	1	0.01	10
Learning rate	0.001	0.00001	0.001	0.001

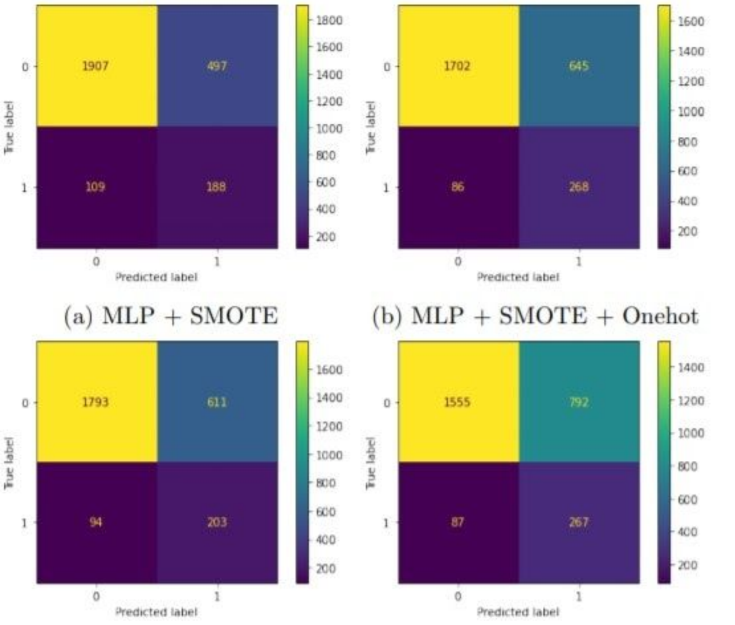
Table 3: Neural network structures of MLP models

Table 5.2.1 shows the optimal hyperparameter configurations. The L2 penalty factors are quite different for each model while the learning rate is almost the same. It is possibly because the L2 penalty is strongly connected to the feature space, which is different for each model as well, while the learning rate is more connected to the learning algorithms, which is the same for four models.



	Ordinal & SMOTE	Onehot & SMOTE	Ordinal & RUS	Onehot & RUS
ACC	0.7845	0.7900	0.7312	0.6271
ROC-AUC	0.7077	0.6189	0.6940	0.6823
Recall	0.6094	0.3870	0.6464	0.7570
Precision	0.2797	0.2813	0.2361	0.2253
F1	0.3834	0.3258	0.3459	0.3473
Weighted F1	0.8160	0.8036	0.8036	0.6877
Recall on “Alive”	0.61	0.39	0.65	**0.76**
Precision on “Alive”	0.28	0.28	0.24	0.23
F1 on “Alive”	0.38	0.33	0.35	0.35

Table 4: Metrics of MLP models



(a) MLP + SMOTE

(b) MLP + SMOTE + Onehot

(c) MLP + RUS

(d) MLP + RUS + Onehot

Figure 4: Validation curves of tuning learning rate

Table 5.2.2 and Figure 5.2.4 shows metrics for each model. The model with ordinal encoding and SMOTE has the highest F1-weighted score. While not been used as the criterion of tuning hyperparameters, it's worth noticing that the model with one-hot and RUS has the highest recall score for label "alive", which is the most "human" model (low recall score for "alive" means telling people they are going to dead while actually they will be alive).

5.3 Kaplan-Meier:

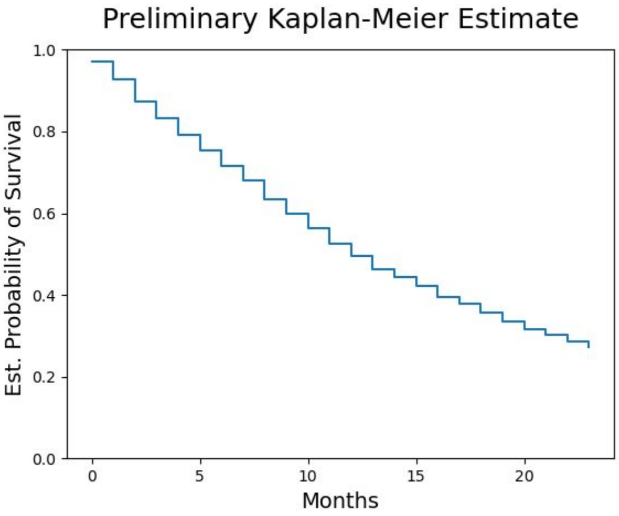


Figure 5.3.1: Kaplan-Meier Estimate
The current Kaplan-Meier curve displays a median value of 0.51, with a minimum value of 0.27 and a maximum value of 0.97. The SE is 0.21. As the SE is quite high, future models will aim at reducing this through the comparison of individual features and alteration of parameters available.

Bibliography

1. Siegel, R. L., Miller, K. D., & Jemal, A. Cancer statistics, 2015 CA Cancer J Clin 2015; 65: 5-29. *External Resources Pubmed/Medline (NLM) Crossref (DOI)*
2. Gillison, M. L., Chaturvedi, A. K., Anderson, W. F., & Fakhry, C. (2015). Epidemiology of human papillomavirus–positive head and neck squamous cell carcinoma. *Journal of Clinical Oncology*, *33*(29), 3235.
3. Forte, T., Niu, J., Lockwood, G. A., & Bryant, H. E. (2012). Incidence trends in head and neck cancers and human papillomavirus (HPV)-associated oropharyngeal cancer in Canada, 1992–2009. *Cancer Causes & Control*, *23*(8), 1343-1348.
4. Chaturvedi, A. K., Anderson, W. F., Lortet-Tieulent, J., Curado, M. P., Ferlay, J., Franceschi, S., ... & Gillison, M. L. (2013). Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers. *Journal of clinical oncology*, *31*(36), 4550.
5. Fakhry, C., Westra, W. H., Li, S., Cmelak, A., Ridge, J. A., Pinto, H., ... & Gillison, M. L. (2008). Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial. *Journal of the National Cancer Institute*, *100*(4), 261-269.
6. Ang, K. K., & Sturgis, E. M. (2012, April). Human papillomavirus as a marker of the natural history and response to therapy of head and neck squamous cell carcinoma. In *Seminars in radiation oncology* (Vol. 22, No. 2, pp. 128-142). WB Saunders.
7. Xu, C. C., Biron, V. L., Puttagunta, L., & Seikaly, H. (2013). HPV status and second primary tumours in oropharyngeal squamous cell carcinoma. *Journal of Otolaryngology-Head & Neck Surgery*, *42*(1), 36.
8. Johnson, S., Corsten, M. J., McDonald, J. T., & Gupta, M. (2010). Cancer prevalence and education by cancer site: logistic regression analysis. *Journal of Otolaryngology--Head & Neck Surgery*, *39*(5).
9. Groome, P. A., Schulze, K. M., Keller, S., Mackillop, W. J., O'Sullivan, B., Irish, J. C., ... & Hammond, J. A. (2006). Explaining socioeconomic status effects in laryngeal cancer. *Clinical Oncology*, *18*(4), 283-292.
10. Johnson, S., McDonald, J. T., & Corsten, M. (2012). Oral cancer screening and socioeconomic status. *Journal of Otolaryngology--Head & Neck Surgery*, *41*(2).
11. Walker, B. B., Schuurman, N., Auluck, A., Lear, S. A., & Rosin, M. (2017). Socioeconomic disparities in head and neck cancer patients' access to cancer treatment centers. *Rural and Remote Health*, *17*(3), 41.
12. Kim, J. D., Firouzbakht, A., Ruan, J. Y., Kornelsen, E., Moghaddamjou, A., Javaheri, K. R., ... & Cheung, W. Y. (2018). Urban and rural differences in outcomes of head and neck cancer. *The Laryngoscope*, *128*(4), 852-858.
13. Khalil, D., Corsten, M. J., Holland, M., Balram, A., McDonald, J. T., & Johnson-Obaseki, S. (2019). Does Socioeconomic Status Affect Stage at Presentation for Larynx Cancer in Canada's Universal Health Care System?. *Otolaryngology–Head and Neck Surgery*, *160*(3), 488-493.
14. Abu-Hanna, A., & Lucas, P. J. (2001). Prognostic models in medicine. *Methods of information in medicine*, *40*(01), 1-5.
15. Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, *1*(4), 274.
16. Lucas, P. J., & Abu-Hanna, A. (1998). *Prognostic methods in medicine* (Vol. 1998). Utrecht University: Information and Computing Sciences.

17. SEER. “Months Survived Based on Complete Dates.” *SEER*, Accessed July 9, 2020 from: https://seer.cancer.gov/survivaltime/SurvivalTimeCalculation.pdf.