SENG 474: Data Mining - Final Report

Group I - Team Members (Undergraduate)

Cameron Lindsay, Claire Champernowne, Ethan Getty, Elliot Kong

1. Introduction

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].

2. The Problem

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]. 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 using patient socioeconomic status as a defining feature for classification.

3. Background Information

As mentioned above, overall survival rates have been shown to be significantly lower in patients with lower socioeconomic status [18-21]. Travel time and by extension, access to specialized health care services, may be a leading factor to the later stages of clinical presentation seen by HNC patients living in rural and at-risk communities [22]. Additionally, access to otolaryngology - head and neck surgery care providers outside of urban centers may only decrease as there remains little-to-no representation in Aboriginal centers and decreased growth in rural areas within Canada; instead, growth in the specialty remains allocated to urban centers [23].
Machine learning methods have a history of being used for the prediction of survivability outcomes. Recent studies have shown logistic regression, SVMs, and ANNs to produce very effective results when predicting survivability outcomes of trauma patients [25]. given these particular models have shown effective outcomes in comparable studies, they have been chosen for observation in our current study.

4. Methodology

The observed dataset was collected from 18 registries between 2000 and 2017 with malignant tumors (according to ICD-O-3 behavior classifications). A narrower timeline was chosen to ensure survival outcomes were less dependent on the year of diagnosis as a predictor of survival. 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". As socioeconomic status is a complex factor that cannot be formally detailed by the SEER database, patient income was used as a surrogate marker. Dataset proportions are outlined in further detail in Table A1.

4.1 Principal Component Analysis

In order to reduce dimensionality and eliminate any irrelevant or redundant data, Principal Component Analysis (PCA) was used to remove insignificant features from the observed dataset. PCA was implemented using SciKit Learn's 'PCA' function with all 15 of our observed features.

4.2 Synthetic Minority Oversampling Technique

Due to the imbalance of our dataset and the issues of bias toward the majority class, we chose to use the oversampling technique Synthetic Minority Oversampling Technique (SMOTE) to mitigate bias through the creation of synthetic examples in our minority class ('Alive'). SMOTE was implemented using SciKit Learn's 'SMOTE' function with default parameters except for a k_neighbors = 5, n_jobs = -1, and m_neighbors = 3.

4.3 ONE-HOT Encoding

Given ONE-HOT encoding's prevalence in addressing multi-class classification tasks and the unknown influence of certain categorical variables within our dataset, ONE-HOT was chosen for observation relative to the preprocessing strategies mentioned earlier. ONE-HOT encoding encodes categorical variables to a new binary variable for each unique integer value. ONE-HOT was implemented only on features without known ordinal values.

4.4 Kaplan Meier Survival Analysis

Kaplan Meier survival curve analysis was implemented on the naive dataset using the Lifelines library function 'KaplanMeierFitter'. Median Income was compared to Survival months. Standard deviation of values surrounds individual curves as a measure of variance between data. Multivariate Log Rank Tests were performed with the Lifelines library as well (lifelines.statistics 'multivariate_logrank_test' function). Multivariate Log Rank was performed on all income brackets observed (Table A2). The confidence interval used $\alpha = 0.005$, degrees of freedom (dof) = 9, Test Statistic = 46.29346630906083, and a null_distribution = chi squared (H0:h1(t)=h2(t)=h3(t)=...=hn(t); HA:there exist at least one group that differs from the other).

4.5 Logistic Regression Model

Logistic regression was done using the sklearn library. Parameters were default, with the exception of multi_class="multinomial", solver="newton-cg", penalty="l2", and C value. C values were different for each preprocessing combination tested, and were determined via validation curve. C=10e-11 was calculated to be optimal for ONE-HOT+PCA+SMOTE, and C=10 for PCA+SMOTE.

4.6 Decision Trees

Decision trees were done using the sklearn library. Parameters were default, with the exception of criterion="gini", and max_depth. Optimal maximum depth was different for each preprocessing combination tested, and was determined via validation curve. Max_depth=22 was calculated to be optimal for ONE-HOT+PCA+SMOTE, and max_depth=20 for PCA+SMOTE.

4.7 Neural Networks

Artificial neural network model was done using *tensorflow2.0* and *keras* library. Parameters include number of hidden layers, number of neurons, dropout rate (changed according to different preprocessing procedures), adam optimizer, binary cross entropy loss function and ReLu activation function.

4.8 Support Vector Machines

Support Vector Machines were done using sklearn's C-Support Vector Classification (sklearn.svm.SVC). All parameters used were default, with the exception of max_iter=5000, and with 5-fold cross validation for both validation curves and learning curves. The only difference in parameters between models was the kernel type; linear and Gaussian. The curves were utilized to observe accuracy, surrounded by standard deviation as a measure of variance between them.

5. Results

5.1 Kaplan Meier and PCA

A Kaplan Meier curve was performed using three values of median income (Figure 1) in conjunction with a multivariate log rank test on all values of median income (Table A2). As demonstrated by the Kaplan Meier survival curve, patients with an overall higher income have better survival outcomes relative to those in a lower income bracket. When comparing the highest income bracket ($75,000+) and the lowest (<$35,000), difference in overall survival at later months (15+) appears to be significantly different.

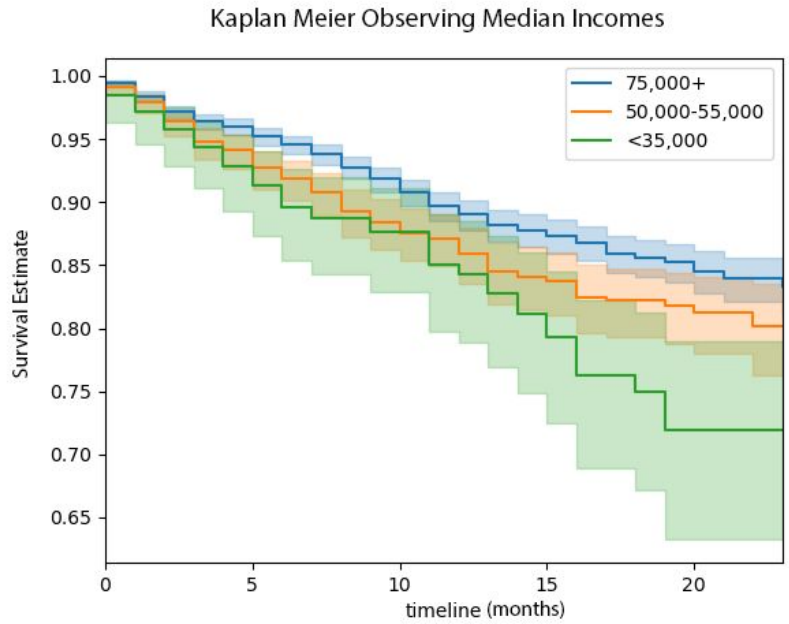

Figure 1. Kaplan Meier Survival Curve Observing Median Income

Results of the Kaplan Meier curve are further supported by the log rank test, as at least one income bracket's mean survival time frame differs from another ($p<0.005$). These results are reflective of the current literature, and suggest that income is an integral factor in the determination of a patient's overall survival.

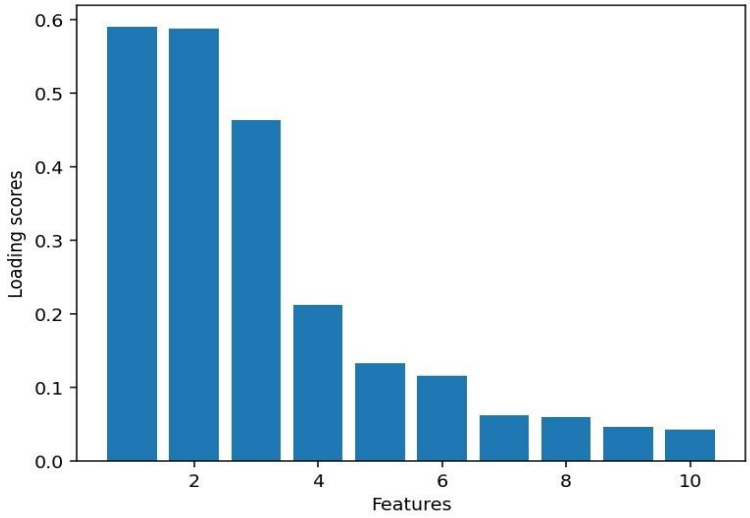

Figure 2. Loading Scores of Features following PCA

Contrary to the results of the Kaplan Meier, PCA analysis demonstrated Median Income to have very little influence on the classification results, ranking 7 out of the selected 10 features with a loading score of 0.062. As this result was obtained early on in our analysis, it was decided that an alternative experimental variable was required in the analysis of our models. The inclusion of the ONE-HOT encoding was utilized in its place, as its influence on the potential models were shown to be observably significant in later experimentation.

5.2 Decision Trees

Optimal maximum depth of decision tree was determined for both SMOTE+PCA and SMOTE+PCA+ONE-HOT using validation curves.

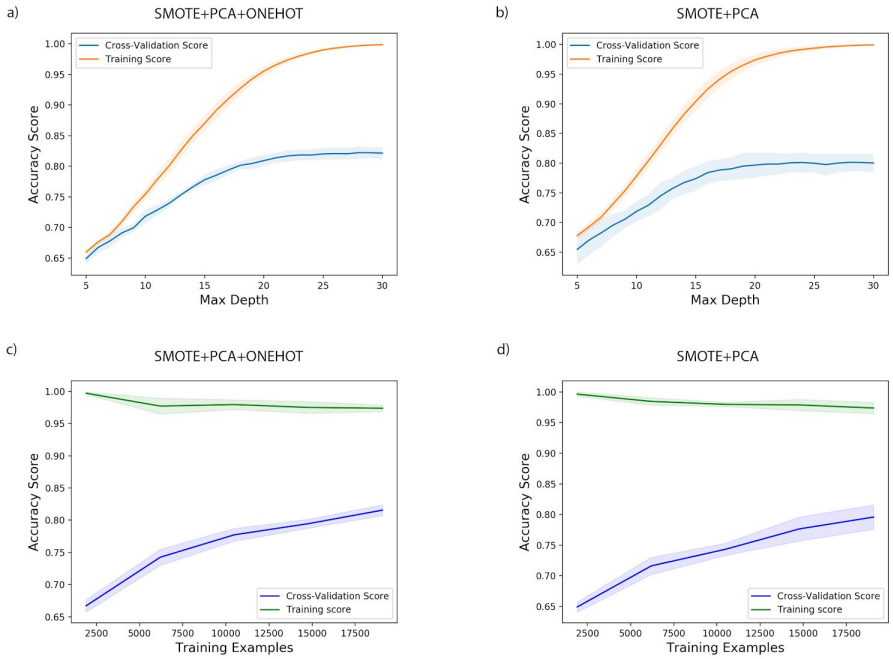

Figure 3. Validation and Learning Curves for Decision Tree Models

Decision trees proved the most effective at predicting outcomes. There were significant differences in metrics from decision trees vs other models; this discrepancy should be explored in further research, given that it is unclear why such a difference in metrics occured. Preprocessing with SMOTE+PCA+ONE-HOT data produced the best results. Specifically both recall and precision increase, showing that the number of false negatives and positives decreases using ONE-HOT encoding.

Table 1. Testing Results of Decision Tree Models

Metrics	SMOTE + PCA	SMOTE + PCA + ONE-HOT
accuracy	0.796	0.819
precision	0.781	0.787
recall	0.827	0.862
roc-auc	0.796	0.821
f1	0.804	0.823
weighted f1	0.796	0.819

5.3 Logistic Regression

Optimal C value for logistic regression was determined for both SMOTE+PCA and SMOTE+PCA+ONE-HOT using validation curves.

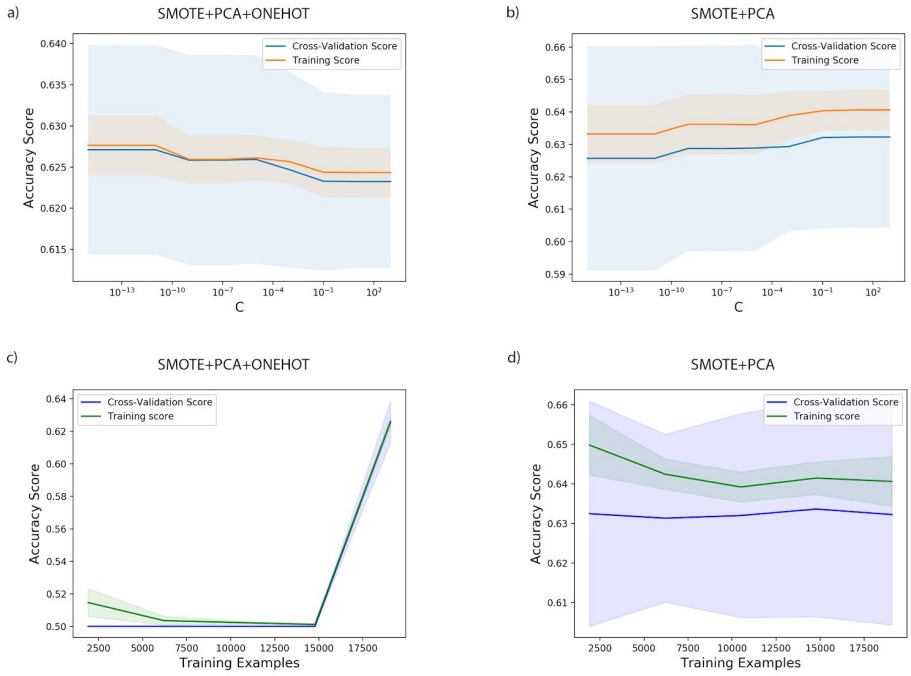

Figure 4. Validation and Learning Curves of Logistic Regression Models
The most successful outcome using logistic regression was achieved using a combination of SMOTE+PCA.

ONE-HOT encoding was ineffective to the point that it reduced the results to less that would be achieved with random guesses. Notably, SMOTE+PCA+ONE-HOT achieved a recall of 1.00, meaning there were 0 false negatives. However, the precision was 0.496, suggesting a high number of false positives. This suggests that the model was always guessing one way and made it ineffective for our purposes.

Table 2. Testing Results of Logistic Regression Models

Metrics	SMOTE + PCA	SMOTE + PCA + ONE-HOT
accuracy	0.651	0.496
roc-auc	0.652	0.5
precision	0.670	0.496
recall	0.627	1.0
f1 score	0.648	0.663
weighted f1 score	0.651	0.329

5.4 Neural Networks

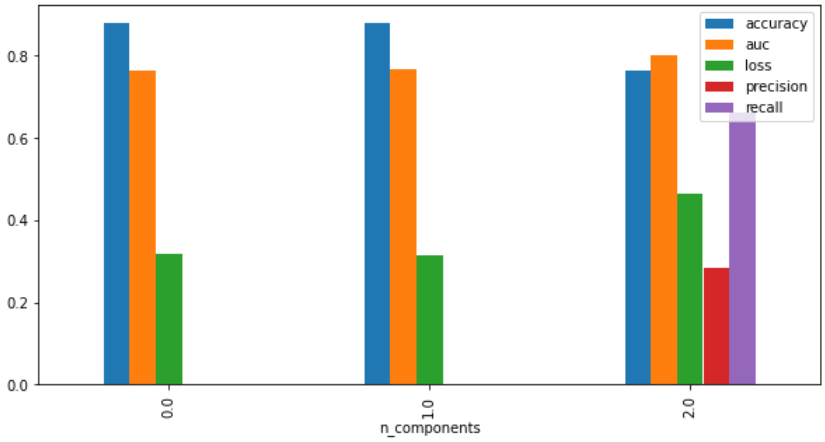

Figure 5. Raw, PCA, PCA+SMOTE

In Figure 5, four metrics are calculated for different methods. When all features are used for classification, the accuracy is above 86 percent. This is because at this stage, the data is still imbalanced and thus the model is blindly classifying all samples as 0. Consequently, precision and recall at this stage are 0. After the application of PCA, we observe that all the metrics haven't been significantly changed. In this case, we keep all 15 principal components included, thus all the information of original features are captured by principal components. After running PCA, SMOTE resampling method is applied and the recall in this condition has been increased to about 64 percent, precision is also increased to near 25 percent. This increase is due to some samples being labeled as 1 correctly.

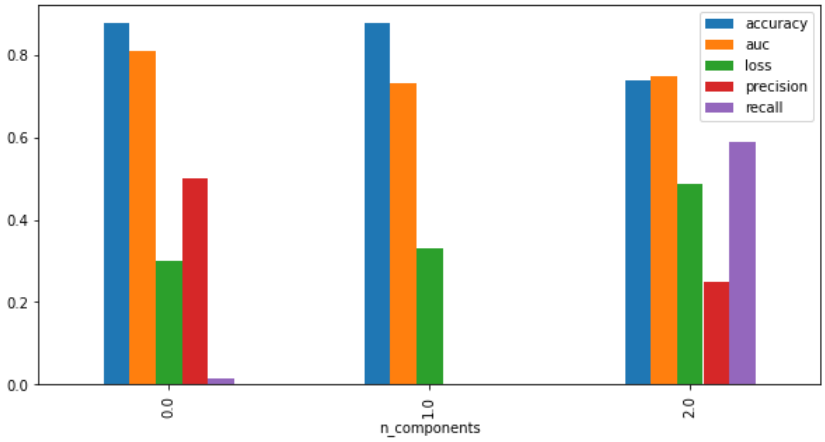

Figure 6. ONE-HOT, ONE-HOT+PCA, ONE-HOT+PCA+SMOTE

In Figure 6, ONE-HOT was applied to the raw data before the application of PCA and SMOTE. 10 features that are not supposed to be ordinal are selected and transformed by ONE-HOT encoding. The number of features is expanded to 135. When all features are used for classification, the recall is just right above 0 percent and precision is about 50 percent. After the application of PCA, we observe that both the recall and precision have been reduced to 0. This reduction shows that reducing the feature space from 135 features to 15 features has led to loss of some information during the application of PCA because for comparison we only keep 15 principal components included. After the running of SMOTE, we observe that this recall has been increased to 0.62 and precision to 0.23, which shows that using SMOTE after PCA helps to increase the performance of classification.

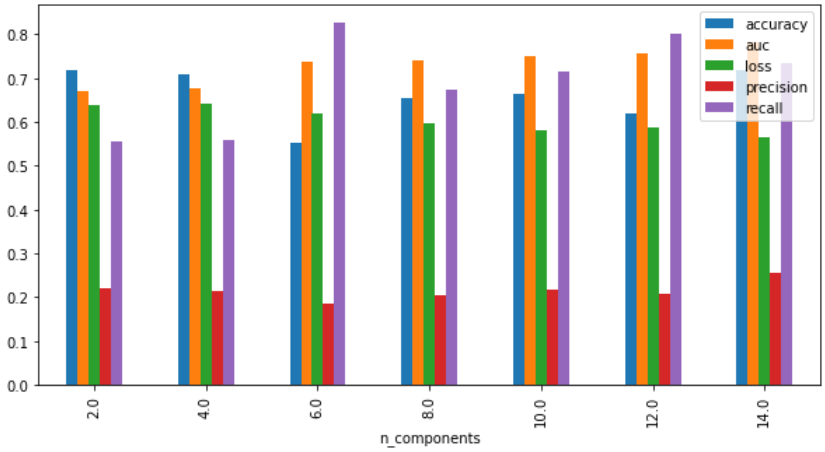

Figure 7. N principal components with Raw data.

The results shown above indicate the number of principal components has a significant influence on the performance of the model. In Figure 7, we gradually increased the number of principal components used to feed the model and calculated metrics for each combination of principal components. As shown in the figure, the model with only the 1st two principal components has the lowest recall and precision, 0.56 and 0.23 respectively. As 14 principal components included, the recall and precision have been increased to 0.74 and 0.25. AUC score was constantly

increased as the number of principal components went up. This result indicates that the last few principal components still provide useful information to help classify the data.

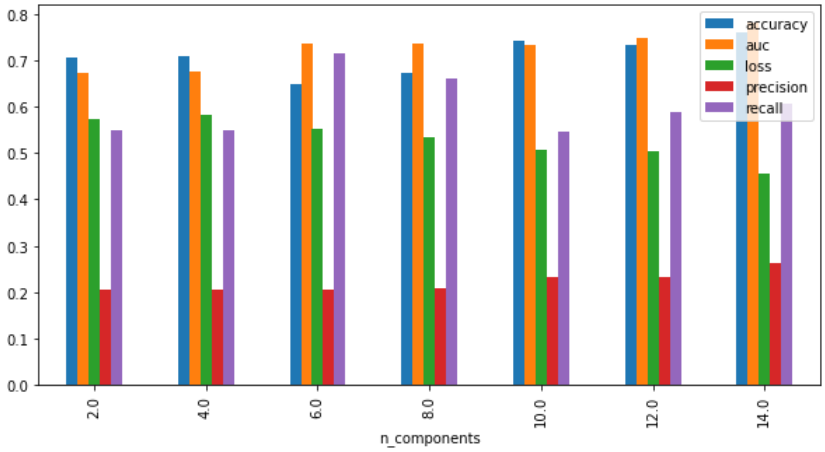

Figure 8. N principal components with ONE-HOT encoded data.

In Figure 8, the result with ONE-HOT encoded data is similar to the results above. But note that the precision was gradually increased with the number of principal components. On the other hand, overall performances of models with ONE-HOT are worse than models with raw data in this case. It is possible due to that PCA significantly reduced the dimension of the data and led to a loss of some useful information.

5.5 SVM

5.5.1. Linear Kernel

Gamma (γ) values were modified using sklearn's default scaling coefficient for γ = 1 / (n_features * X.var), as well as with the modification of the C regularization parameter.

Using validation curves with a 5-fold cross-validation strategy (Figure 9a, b), optimal C values for both the PCA+SMOTE (C = 7.5e-05) and PCA+SMOTE+ONE-HOT (C = 8.0e-03) were obtained and used in the final model comparison. Validation curves used a base range of C-values from 1.0e-30 to 10 for observation and was narrowed accordingly to obtain more accurate values for final models. Learning curves were performed to observe the benefit of training data to the models accuracy. As shown in Figure 9c and d, both models converge completely when the complete dataset is used and training score is maximized at approximately 15000 samples.

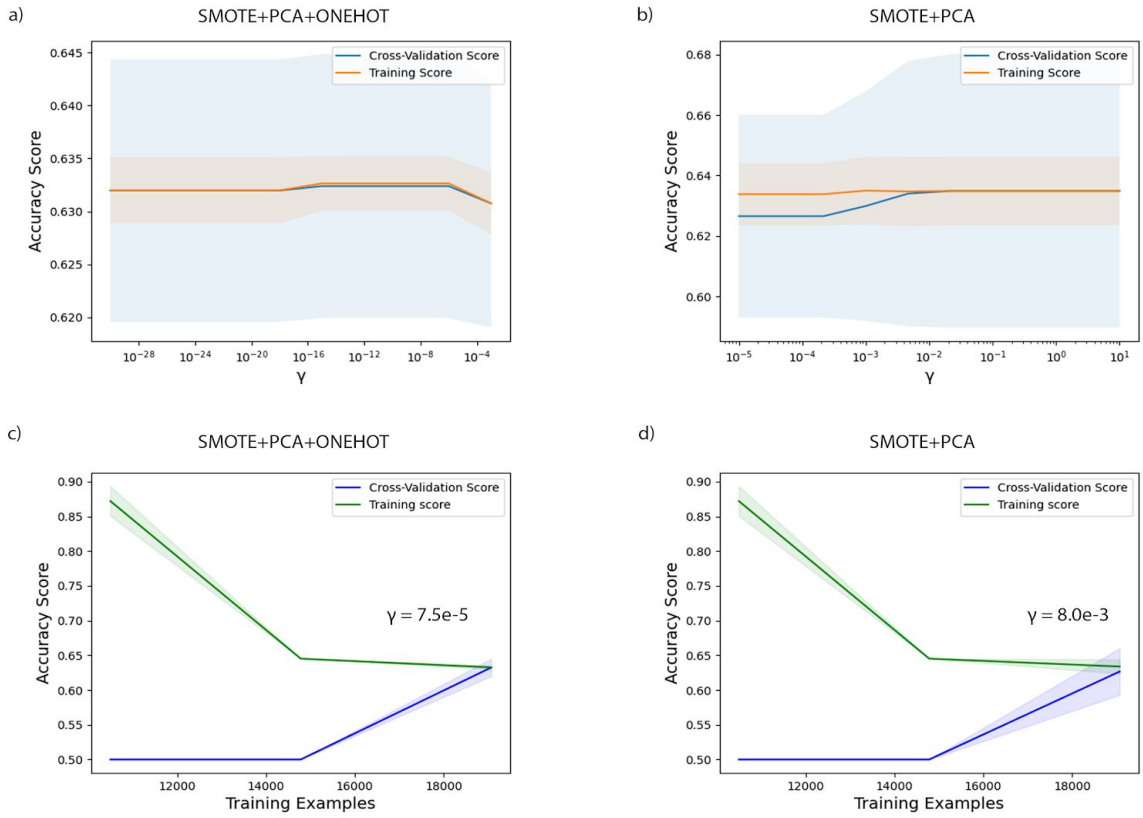

Figure 9. Validation and Learning Curves of SVM Models using a Linear Kernel

Table 3. Testing Results of SVM Models using a Linear Kernel

Metrics	SMOTE + PCA	SMOTE + PCA + ONE-HOT
accuracy	0.636	0.492
roc-auc	0.638	0.5
precision	0.605	0.492
recall	0.749	1.0
f1 score	0.669	0.659
weighted f1 score	0.631	0.324

Testing individual models to the trained dataset shows the disparities between models. Overall scores suggest a combination of SMOTE and PCA to be the more effective. Namely, with the application of ONE-HOT, a small loss to f1 score by 0.01, and much greater losses to weighted f1 score by 0.307, ROC-AUC score by 0.138, and accuracy by 0.144 were observed. With the application of ONE-HOT, recall increases and precision decreases, suggesting that the number of false negatives/positives decreases using ONE-HOT encoding. The PCA+SMOTE+ONE-HOT model accuracy is much worse.

5.5.2. Gaussian Kernel

γ values were modified using sklearn's default scaling coefficient for γ = 1 / (n_features * X.var), as well as with the modification of the C regularization parameter.

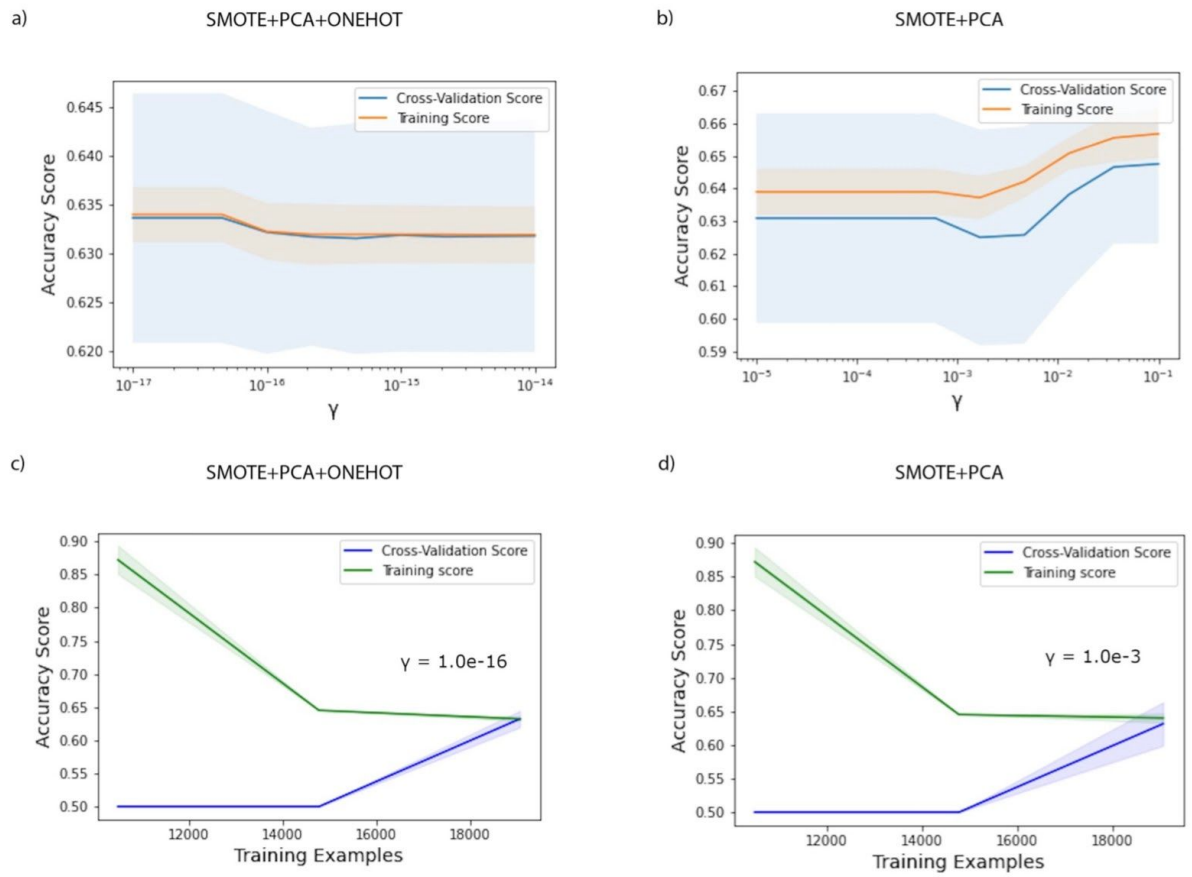

Figure 10. Validation and Learning Curves of SVM Models using a Gaussian Kernel

Using validation curves with a 5-fold cross-validation strategy (Figure 10a, b), optimal C values for both the PCA+SMOTE (C = 1.0e-03) and PCA+SMOTE+ONE-HOT (C = 1.0e-16) were obtained and used in the final model comparison. Validation curves used a base range of C-values from 1.0e-30 to 10 for observation and was narrowed accordingly to obtain more accurate values for final models. Learning curves were performed to observe the benefit of training data to the models accuracy.

Table 4. Testing Results of SVM Models using a Gaussian Kernel

Metrics	SMOTE + PCA	SMOTE + PCA + ONE-HOT
accuracy	0.637	0.492
roc-auc	0.640	0.5
precision	0.592	0.492

recall	0.840	1.0
f1 score	0.695	0.660
weighted f1 score	0.621	0.325

Testing individual models to the trained dataset shows the disparities between models. Overall scores suggest a combination of SMOTE and PCA to be the more effective. Namely, with the application of ONE-HOT, there were great losses to weighted f1 score by 0.296, ROC-AUC score by 0.140, and accuracy by 0.145 were observed. With the application of ONE-HOT, recall increases and precision decreases, suggesting that the number of false negatives/positives decreases using ONE-HOT encoding. The PCA+SMOTE+ONE-HOT model accuracy is much worse.

6. Discussion

Given every model tested was able to achieve >50% accuracy, suggesting that the features in the dataset used are not random. Preprocessing had an effect in every case tested, though the same combination of preprocessing was not the best for every single model. Specifically, decision trees benefited from the use of ONE-HOT encoding. However, every other model tested showed ONE-HOT's implementation to be unhelpful. Notably, in the cases of logistic regression and both SVMs tested, ONE-HOT encoding caused the results to fall below what metrics would be achieved with random guesses. Despite the Kaplan Meier curve demonstrating that there is a correlation between income level and survivability, income level was shown to be a relatively unimportant feature when it came to the machine learning models tested. This resulted in a shift in focus from observing differences following income inclusion/exclusion in model development, to attempting to maximize the viability of a model for OPSCC. ANN, logistic regression, and both SVM models achieved similar results; however, decision trees achieved observably better metrics across the board. Reasons for this large difference in metrics should be explored further in later research.

Several limitations present in the dataset likely impacted the accuracy of our models. In addition to the classifier imbalance, various lifestyle and disease-specific risk factors such as smoking, HPV status, patient histories including medications, and genetic backgrounds were unavailable for view. These factors have all shown demonstrably significant impact on patient mortality [1, 3]. Future models would benefit from the addition of some of these factors mentioned above. Fortunately, SEER has access to specialty datasets with longer approval periods, like the Head and Neck with HPV database, that include some of these features.

Appendix

TableA1: Summary of Samples in SEER Dataset

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

Table A2. Median Household Incomes

Median household income inflation adj to 2018	n samples from dataset
$75,000+	4063
$70,000 - $74,999	1040
$65,000 - $69,999	1157
$60,000 - $64,999	2499
$55,000 - $59,999	898
$50,000 - $54,999	1316
$45,000 - $49,999	1121
$40,000 - $44,999	621
$35,000 - $39,999	466
< $35,000	322

References

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.
18. Johnson, S., et al., *Cancer prevalence and education by cancer site: logistic regression analysis.* J Otolaryngol Head Neck Surg, 2010. **39**(5): p. 555-60.
19. Groome, P.A., et al., *Explaining socioeconomic status effects in laryngeal cancer.* Clin Oncol (R Coll Radiol), 2006. **18**(4): p. 283-92.
20. Johnson, S., J.T. McDonald, and M. Corsten, *Oral cancer screening and socioeconomic status.* J Otolaryngol Head Neck Surg, 2012. **41**(2): p. 102-7.
21. Walker, B.B., et al., *Socioeconomic disparities in head and neck cancer patients' access to cancer treatment centers.* Rural Remote Health, 2017. **17**(3): p. 4210.
22. Khalil, D., et al., *Does Socioeconomic Status Affect Stage at Presentation for Larynx Cancer in Canada's Universal Health Care System?* Otolaryngol Head Neck Surg, 2019. **160**(3): p. 488-493.

23. Crowson, M.G. and V. Lin, *The Canadian Otolaryngology-Head and Neck Surgery Workforce in the Urban-Rural Continuum: Longitudinal Data from 2002 to 2013.* Otolaryngol Head Neck Surg, 2018. **158**(1): p. 127-134.
24. Jaret, P. *Attracting the next generation of physicians to rural medicine*. 2020 [cited 2020 May]; Available from: https://www.aamc.org/news-insights/attracting-next-generation-physicians-rural-medicine.
25. Rau, C. S., Wu, S. C., Chuang, J. F., Huang, C. Y., Liu, H. T., Chien, P. C., & Hsieh, C. H. (2019). Machine Learning Models of Survival Prediction in Trauma Patients. *Journal of clinical medicine*, *8*(6), 799. https://doi.org/10.3390/jcm8060799
26. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences, 374*(2065), 20150202. doi:10.1098/rsta.2015.0202
27. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002;2011;). SMOTE: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research, 16*, 321-357. doi:10.1613/jair.953