IT'S TIME TO COME CLEAN ABOUT COOKING

Constructing Machine Learning Models for Social Welfare Targeting in South Africa to Predict Whether Households Use Clean or Dirty Cooking Fuel

ABSTRACT

Toxins from dirty cooking fuels kill more people than malaria every year, predominantly affecting women and children (Whiting, 2021). Governments could intervene to provide households with clean, safe alternatives, but typically lack the ability to identify which households need alternatives. I use South Africa's DHS survey data to build machine learning models that predict which households use dirty and clean fuels. I begin by restructuring the data and analyzing the relationships between variables with principal component analysis and multiple components analysis. Next, I construct four models: a KNN model, a support vector machine model, a logistic regression model built with forward selection, and an ensemble model that aggregates the predictions of the previous three models. The logistic regression model performs best on test data with an error rate of 17.72%. These models serve as a foundation for future analysis and provide proof of concept for cooking fuel welfare applications. This novel application of machine learning draws attention to the power of predictive modeling in combating persistent health threats and provides resource-limited governments with a means of targeting their welfare interventions.

1 Introduction

1.1 Background

One of the most dangerous threats women face lives in the kitchen: cooking stoves. One-third of all people globally depend upon dirty cooking fuels like biomass and coal (Whiting, 2021). This problem is most acute in low and middle-income countries (LMICs), where two million people burn wood, crop byproducts, dung, and other biomass as their primary energy source (Mudway et al., 2005). The use of dirty cooking fuel has significant implications for women's and children's health: "worldwide, indoor air pollution is the single largest environmental risk factor for female mortality and the leading killer under the age of five" (Perez, 2019, pg. 152). When a traditional stove is used in an unventilated room, it releases toxic fumes that have the equivalent impact on the cook as smoking over 100 cigarettes (Perez, 2019, pg. 152). These fumes are associated with 3.8 million deaths caused by cancer, pneumonia, and other diseases each year, killing more people than Malaria (Whiting, 2021). 753 million people in Africa, which is 80% of the population, still predominantly rely upon biomass to fuel their cooking (Whiting, 2021; Perez, 2019, pg. 152).

South Africa exemplifies how economic development has left millions of people behind and perpetuated the legacy of dirty cooking fuels. South Africa's stark economic divides are borne from a dual economy. One part of the economy is globalized, modern, highly productive, and wealth-producing, while the other contains the poorest urban and rural South Africans, has not significantly benefited from technological advancements, and is unable to produce its own growth (Bojabotseha, 2011). The country's Gini Coefficient was 0.67 in 2018 and 25% of South Africans were still living in food poverty in 2020 (WorldBank, 2023; EquityBrief, 2023). Despite its historic progress in poverty reduction, South Africa has experienced increases in poverty in recent years (EquityBrief, 2023). Although South Africa has many resources as a country, it still needs targeted interventions to reduce poverty and merge the dual economy into a single, functioning whole.

Impoverished households in the marginalized economy are likely to cook using traditional stoves with biomass fuel. The World Bank estimates that 13% of the South African population lacks access to clean cooking fuels and technologies (WorldBankData). This mixed composition of fuel use makes South Africa a viable focus for cooking fuel analysis. Moreover, its relative abundance of financial resources indicates that it may have the capacity to implement a widespread solution if it could identify which households cook with dirty fuel. This paper implements machine learning methods to develop a comprehensive model for the identification of households burning clean and dirty cooking fuels in South Africa.

1.2 Machine Learning in Welfare Targeting

Governments of LMICs often have the most severe resource limitations and the weakest bureaucratic infrastructure. They lack the administrative data on income and wealth that is needed to target who should receive welfare goods (Aiken et al., 2023). Government aid is often distributed with little accuracy and overextends to those who do not need it. Poor targeting of social programs limits the scale of poverty-reduction efforts; countries with the most need have the least capacity to ameliorate poverty.

Machine Learning is a necessary tool in economic development because it compensates for deficiencies in administrative data. Machine learning enables researchers to build models that predict households' need based on their observed characteristics. Aiken et al. (2023) developed a model for targeting households in need of financial support in Afghanistan. They used data on cell phone ownership, call frequency, call duration, and texting to build a model that identifies which households are "ultra poor." The main model applied gradient boosting with 10-fold cross-validation. This model was almost as accurate as targeting using data from household surveys on consumption and wealth. It performs faster and at a lower cost than an administrative data collection program. Machine learning enables dynamic, reactive government interventions.

Sansone and Zhu (2020) demonstrate how machine learning methods can predict which households will continue needing income support. They use Australian social security data and produce a model that is at least 22% more accurate than existing heuristics and early warning systems.

1.3 Goal

This paper's approach is similar to Aiken et al. (2023), as it relies on survey data to avoid potential gaps in administrative data. It uses known household characteristics to predict which households use dirty cooking fuel. South Africa's government could use a similar machine learning model to distribute clean cooking fuel and stoves to these households, reducing the morbidity and mortality of women and children.

2 Data Collection & Description

2.1 Data Source

This analysis uses data from The Demographic and Health Surveys (DHS) Program's 2016 survey in South Africa. DHS surveys are nationally representative and survey between 3,000 to 30,000 households (DHS). Please reference Appendix A for additional details on the selection of survey year 2016.

This analysis utilizes the DHS's Household Member Recode (HMR) file. DHS HMR files use standardized data definitions across countries. The HMR is organized by household and includes an observation for each household member. The 2016 South Africa HMR has 38,850 observations and 11,083 unique households. It includes 441 individual and household-level variables covering topics such as health, housing, assets, household demographics, fertility, and language.

2.2 Data Restructuring

I restructured the DHS survey data so that each observation is a household instead of an individual. I began by extracting a list of the unique household IDs. Each household ID is composed of two numbers that, when combined, may match another household ID in the sample. To avoid losing unique observations with the same numeric ID, I collapsed the households across a set of 15 continuous, binomial, and multinomial categorical variables. The chance of two households having the same numeric ID and combination of all 15 variables is near zero. I dropped households with "NA" responses, resulting in a 16-variable data frame of 10,722 households.

The data included 12 categories for the fuel type variable. I recoded these response categories in two ways, both of which are presented in Appendix B. First, I split the 12 categories into 7 broader categories and dropped the category for "Other" because it is neither "clean" nor "dirty." This facilitates the creation of a model that predicts each household's primary fuel type. Second, I produced a new response variable that codes each fuel as "clean" or "dirty." This binary response variable provides less information but results in a more precise, simpler model. It also distills to the core of what a government would want to know about each household: does this household need a welfare cooking intervention? Through this process, I determined that 2,061 households in the dataset use dirty cooking fuel. Please reference Appendix

C for a discussion of why this observed frequency is greater than the World Bank's estimate.

Next, I cleaned the 14 independent variables[1] by relabeling rare survey responses as "other." My threshold for counting as "rare" was being in 10 or fewer recorded responses. Collapsing the data to the household level and consolidating the response variable resulted in a DHS-derived dataset of 10,722 households and 17 variables. Please reference Appendix D for a list of the 17 variables and their descriptions.

3 Exploratory Data Analysis

This section explores associations between the models' 14 independent variables. Continuous and categorical variables are considered separately due to differences in variance structure.

3.1 Principal Component Analysis (PCA)

There are three continuous variables in this report's models. They are the number of eligible women, number of eligible men, and number of people in each household. I scaled and centered all three variables and then conducted a PCA analysis. Figure 1 in Appendix E provides the biplot and screeplot.

The biplot in Panel A indicates there is a positive correlation between variables hv009 (number of household members) and hv010 (number of eligible women in the household). Surprisingly, this is not the case for hv011 (number of eligible men in the household). The vector hv011 forms an approximately 90°angle with the other vectors, indicating that the number of eligible men is not closely associated with the number of eligible women or total number of people in each household. This result is informative because it implies that the number of eligible men may provide unique information and confirms that it may be valuable to include the eligibility data for both sexes, but not necessarily informative to have a variable for both the number of eligible women and household size.

The screeplot in Panel B of Figure 1 displays the proportion of total variance explained by each principal component. There is not a clear spectral gap. The y-axis begins at 0.5; although the amount of variance explained decreases with each additional principal component, all three components maintain explanatory power. This reflects a core limitation of this PCA analysis. The use of three continuous variables means the analysis has fewer relationships to discern and is less informative. The following analysis of the categorical variables provides insight on the binomial variables.

3.2 Multiple Correspondence Analysis (MCA)

I use MCA to asses the binomial variables. Similar to PCA, MCA finds the associations between variables. It also groups observations by their similarities. This analysis focuses on the associations between variables.

Figure 2 in Appendix F illustrates the relationships between the binomial variables. The variables that are 180°apart are negatively correlated, while those that are proximate to each other have a similar profile. hv225 0 and hv243a 1 are near each other which means that households that share a toilet with at least one other household are similar to those that lack a mobile phone. Moreover, households with a male household head (hv219) have a similar profile to households in urban areas (hv025). This may reflect how cultural norms vary on different sides of the South African dual economy.

hv237 0 and hv243c 0 are close to each other and the origin. This indicates that households without an animal-drawn cart have similar profiles to those that do not do anything to clean or purify their water prior to drinking. Both variables may proxy for wealth, which could explain their association and why they are negatively associated with having a mobile telephone (hv243a 1). Distance to the origin describes how well-represented a variable is by the factor map. Variables near the origin, such as the three discussed in this paragraph, are well-represented by the MCA so there is more confidence in their associations than if they were distant.

4 Statistical Models

This report produces and compares three predictive models and an ensemble model with majority voting. All models are supervised and predict categorical fuel types.

4.1 K Nearest Neighbors (KNN)

KNN analysis plots each household in n-dimensional space and predicts labels based on the most common label among each new point's k closest "neighbors." Changing k alters the prediction accuracy. When k is high, the decision boundary is smooth and variance is low. When k is small, bias falls and variance increases. This paper's KNN models have high dimensionality because there are 14 independent variables.

I constructed two KNN models. The first model has seven response variables and predicts the specific type of cooking fuel that each household uses, while the second simply predicts if the fuel is clean or dirty. I began by tuning the value of k for each model. I used ten-fold cross-validation to assess which value of k produced the smallest average 0-1 loss.

The optimal neighborhood size for the seven-category response variable was 19. Training error was 22.81% and test error was 23.82%. In comparison, the optimal neighborhood size was 15 for the two-category response variable. Training error was 17.79% and test error was 18.55%. The model with two response categories outperformed the more complex model. This is expected because there are fewer concerns with data scarcity and only one way to mislabel each observation, rather than six. See Figure 3 in Appendix G for a visualization of how training error changed with neighborhood size in the tuning process and a comparison of the testing and training error at each model's optimal neighborhood size.

The seven-category KNN model is limited because the large neighborhood size k. Labels that appear close to the number of times as the size of k will not be assigned unless they are clustered closely in n-dimensional space and can form a majority. As a result, the model only predicted three fuel types in the test data: biomass (129), electricity (2,007), and kerosene (9). The true frequencies in the test data were: biomass (309), coal (19), electricity (1,602), gas (97), kerosene (109), no cooking (7), and solar (2). This model underestimates the number of households in need of a welfare intervention due to the relative scarcity of dirty fuels in the data compared to the dominant presence of electricity. Sparsity is a limitation of KNN that prevents complex models from converging to a test error of 0 when the sample size is fixed. The model over-predicts the most common responses and is unable to predict the more "rare" responses.

4.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) find a hyperplane in n-dimensional space that classifies the data points using hinge loss and support vectors, which are the data points near the decision boundary that determine its positioning.

I built a non-linear SVM model. I selected a non-linear boundary because of the high dimensionality of my data

[1]This excludes the household ID variable and both new variables for fuel type.

and because a non-linear boundary can produce a linear-like boundary if such a shape minimizes loss. All multinomial categorical variables were translated into binomial variables, which resulted in 41 independent variables being input into the SVM model. I tuned the radial kernel by testing a series of cost and γ values, applying 10-fold cross-validation to evaluate model performance at each combination of values, and selecting the combination with the lowest loss.

I tested 112 combinations of *cost* and γ in the tuning process; the optimal $\gamma = 0.05$ and the optimal $cost = 0.5$ with 3,846 support vectors. Figure 4 in Appendix H illustrates the test error misclassification rate as a function of γ and cost.

Figure 4's darkest regions correspond to the combinations of cost and γ that minimize test error. The top left of the graph is the region with the optimal combination of cost and γ in the training data. This region is roughly the third darkest color and not an optimal combination of cost and γ for the test data. This is reflected in the test data's relatively large decrease in performance compared to the training data. Relative to KNN, logistic regression, and the ensemble, SVM had the largest gap between its training and test performance. The training error was 16.15% and the test error was 18.69%. This reflects the bias-variance tradeoff. The model with the lowest bias has the lowest performance on outside data.

4.3 Logistic Regression

I constructed two logistic models to predict which households use clean and dirty cooking fuel.

First, I built the independence model, which assumes that each variable's effect does not vary by the level or value of any other. View the model specification in Appendix I. I evaluate the independence model by conducting a likelihood ratio test. I regressed the dependent variable on 1 and calculated the test statistic $F = 2(L_1 - L_0) = G^2(M_0|M_1)$. H_0 is that the explanatory variables are independent of each observation's dirty/clean classification. H_A is that they are not independent. The F statistic, which describes the difference in deviance between the null and alternate models (joint significance of the explanatory variables), is 2,177.3 with a p-value of 2.2e-16. Thus, there is nearly a 0% chance of observing this F statistic when the null is true. We can reject the null of independence and explore more complex models. The independence model has 16.57% training error and 17.76% test error. A more complex logistic model may perform better.

I use forward selection to construct a more complex model. Forward selection starts with the null model regressed on 1 and adds terms until adding new terms no longer improves the model. I use AIC to adjudicate which terms to add. View Appendix J for the forward selection model.

The forward selection model utilizes interaction terms, suggesting there are homogenous associations between some variables. It omitted some variables that the exploratory data analysis suggested had high correlations with other variables. For example, the variable hv011 (number of eligible women in the household) was omitted, which is unsurprising given its high correlation with household size in the PCA analysis.

Foward selection was limited in its application due to the large volume of variables and their many categories, preventing R from completing the process. Please reference Appendix K for analysis on the significance of this limitation. Additionally, the high performance of this model led some predicted probabilities to near 0 and 1. This forced forward selection to terminate before converging, which results in higher bias and underfitting; future research may benefit from exploring specifications that permit more complex specifications to produce a more externally valid model.

Despite its limitations, the forward selection outperformed the independence model. A likelihood ratio test between the models produced an F statistic of 87.25 with a p-value of 5.792e-15, allowing the rejection of H_0 that the independence model is an adequate fit compared to the more complex model. Training and test error are slightly reduced. Training error is 16.49% and test error is 17.72%. Thus, the second, more complex logistic model will be used in the ensemble.

4.4 Ensemble Model

Ensemble models pool predictions from multiple models. They follow the "wisdom of the crowd," the idea that averaging many guesses leads to an accurate response (Prelec et al., 2017). In the context of machine learning, when models in the ensemble individually perform slightly better than a random guess, their aggregated predictions can converge on the truth. In this section, I combine the three previous models for predicting the binary categories "clean" and "dirty." I use majority voting and compare the ensemble model's performance to that of each individual model in Section 5.

Ensemble models aggregate the decisions of independent (and often weak) learners. The models in this report utilize the same subset of variables to generate their predictions. This may interfere with the ensemble's ability to produce more accurate predictions than the individual models. Each model should produce errors in a unique way, which allows them to converge on the true value through majority voting. If all models systematically under or overpredict the likelihood of using dirty cooking fuel for the same types of households, aggregating their predictions is of less value. The three models I combine in the ensemble all have comparable rates of test error. Test error ranges from 17.72% to 18.69%, which implies there is a possibility that they are predicting labels in a similar fashion, and may be misclassifying observations of the same profiles. Future analysis can develop this ensemble by using different variables for each model and including more models in the ensemble to maximize its performance.

5 Conclusion and Discussion

Table 3 in Appendix L displays each model's test and training misclassification rate. The logistic model performed best on the test data, with the ensemble performing second best. Although ensemble models are generally expected to outperform their component pieces, this result is unsurprising due to the limited number of models in the ensemble and the likelihood of systemic/similar error between the three models.

Despite the KNN model's lower performance, its nonparametric nature allows its performance to improve indefinitely as data is added. Thus, the KNN method can grow more valuable as new data is collected and incorporated into the models, which would likely be the case if these models were used for welfare targeting.

Although these models have high test errors for their intended purpose of welfare targeting, they serve as a foundation for future analysis and provide proof of concept for cooking fuel welfare applications. A main limitation of these prediction models is in variable selection and the dimensionality of using multinomial categorical variables. Extensions of this analysis could include additional work on shrinkage. The forward selection process demonstrated that not all of the selected variables were significantly beneficial in reducing AIC, which indicates removing some variables could improve performance; effective shrinkage could decrease variance. Extensions could also build a series of weak learners using methods like random forests to construct a stronger ensemble with greater independence between the composite models. These efforts can enhance governments' ability to spread life-saving cooking innovations, such as clean fuel and stoves, and save lives with predictive modeling.

References

Emily L Aiken, Guadalupe Bedoya, Joshua E Blumenstock, and Aidan Coville. Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in afghanistan. *Journal of Development Economics*, 161:103016, 2023.

Teboho Pankratius Bojabotseha. Dualism and the social formation of south africa. *African Journal of Hospitality, Tourism and Leisure*, 1(3):1–8, 2011.

DHS. Dhs overview. URL `https://dhsprogram.com/Methodology/Survey-Types/DHS.cfm`.

EquityBrief. *Equity Brief South Africa*. 2023.

Ian S Mudway, Sean T Duggan, Chandra Venkataraman, Gazala Habib, Frank J Kelly, and Jonathan Grigg. Combustion of dried animal dung as biofuel results in the generation of highly redox active fine particulates. *Particle and fibre toxicology*, 2:1–11, 2005.

Caroline Criado Perez. *Invisible women: Data bias in a world designed for men*. Abrams, 2019.

Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541 (7638):532–535, 2017.

Dario Sansone and Anna Zhu. Using machine learning to create an early warning system for welfare recipients. *arXiv preprint arXiv:2011.12057*, 2020.

Kate Whiting. Cooking with polluting fuels is a silent killer - here's what can be done. Available at `www.weforum.org/agenda/2021/10/polluting-cooking-fuels-deaths-women-climate/` (2021/10/27), 2021.

WorldBank, Mar 2023. URL `http://www.worldbank.org/en/country/southafrica/overview`.

WorldBankData. URL `https://data.worldbank.org/indicator/EG.CFT.ACCS.ZS?locations=ZG`.

6 Thank You

Thank you [redacted] for a wonderful semester! Applied Machine Learning was an excellent statistics class that I thoroughly enjoyed due to the broad range of topics we surveyed and its nurturing learning environment. Thank you to [redacted] for helping me refine my categorical data analysis skills, which I used to complement the machine learning methods in this final report, and for always having an open door and enthusiastically answering all of my statistics questions. I am excited to apply machine learning to more development economics issues in my senior thesis next year. Thank you for giving me the tools to rigorously explore issues I am passionate about.

Appendex

A Justifications for Selecting DHS Survey Data from 2016

2016 was selected because it is the most recent DHS South Africa survey year. DHS surveys were also administered in South Africa in the years 2003 and 1998. These survey years were not included in this analysis due to the extent that technology and demographics have changed. An effective targeting program should be based on the most recent data. Older DHS surveys also include fewer variables and categories, which diminishes the potential utility of the models that use their data. In the absence of high-quality, national-level administrative data, targeting could instead be applied at the state/district level or variables could be approximated using phone data and additional machine learning techniques.

B Dependent Variables

Table 1: Consolidation of the Cooking Fuel Type Response Variable into Seven and Two Categories

DHS Category	DHS Description	Reclassification One	Reclassification Two
1	electricity	electricity	0 clean
2	lpg	gas	0 clean
3	natural gas	gas	0 clean
4	biogas	gas	0 clean
5	kerosene/paraffin	kerosene	1 dirty
6	coal, lignite	coal	1 dirty
7	charcoal	coal	1 dirty
8	wood	biomass	1 dirty
9	straw/shrubs/grass	biomass	1 dirty
10	agricultural crop	biomass	1 dirty
11	animal dung	biomass	1 dirty
12	electricity from generator	electricity	0 clean
13	electricity from other source	electricity	0 clean
14	solar energy	solar	0 clean
95	no food cooked in house	no_cooking	0 clean
96	other	-	-

C Discussion of Difference Between the Share of Households Using Dirty Cooking Fuel in South Africa According to DHS and World Bank

2,061 South African households in the DHS HMR data use dirty cooking fuel. This is approximately 19.19% of households in the DHS sample, which is a 6 percentage point greater share of households using dirty cooking fuel than the 13% of households that lack access to clean cooking fuels and technology according to the World Bank (WorldBankData). There are many possible explanations for this difference. It is possible that households with access to clean fuel are not using it due to barriers like cost, that the DHS survey data are less representative than expected or are now outdated, that the World Bank data is not representative, or that this report's definition of "dirty" cooking fuel is broader than the definition used by the World Bank.

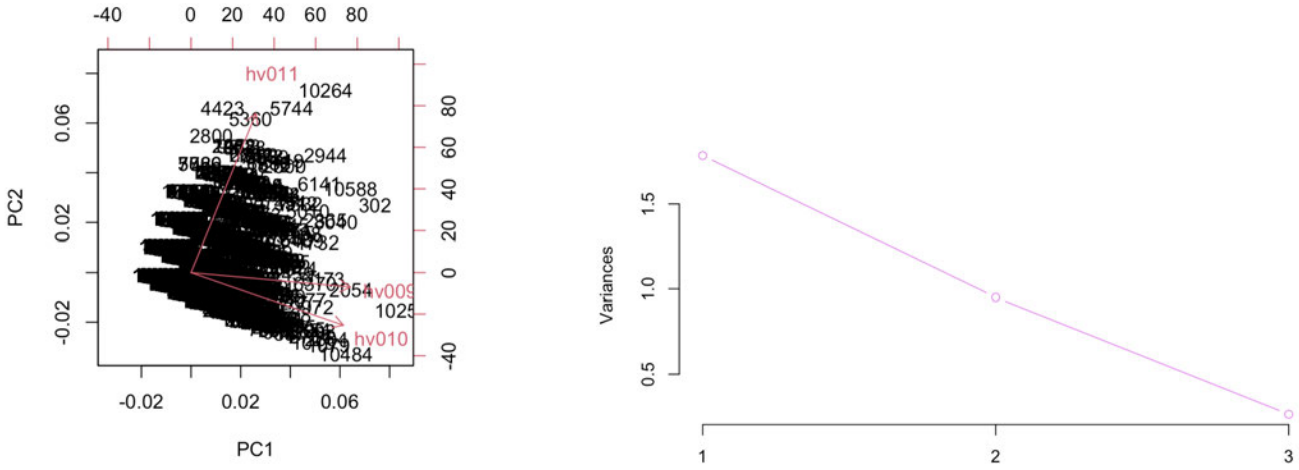
D All Variables

Table 2: Variables in Clean, Household-Level Dataset

	Variable	Description
1	hhid	case identification
2	hv009	number of household members
3	hv010	number of eligible women in household
4	hv011	number of eligible men in household
5	hv025	type of place of residence
6	hv212	has car/truck
7	hv213	main floor material
8	hv215	main roof material
9	hv219	sex of head of household
10	hv221	has telephone (land-line)
11	hv225	share toilet with other households
12	hv226	type of cooking fuel
13	hv237	anything done to water to make safe to drink
14	hv243a	has mobile telephone
15	hv243c	has animal-drawn cart
16	sh141a	type of dwelling

E PCA Biplot and Screeplot

Figure 1: PCA Analysis of Feature Correlations and Variance

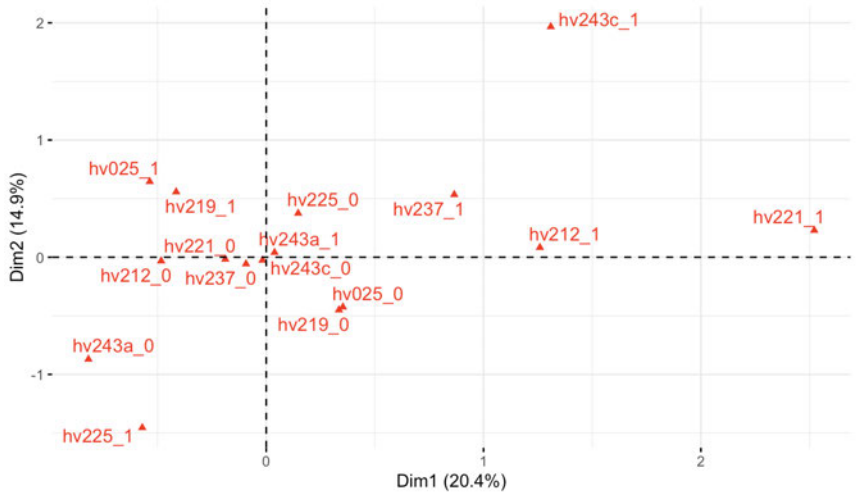


(a) Biplot

(b) Screeplot

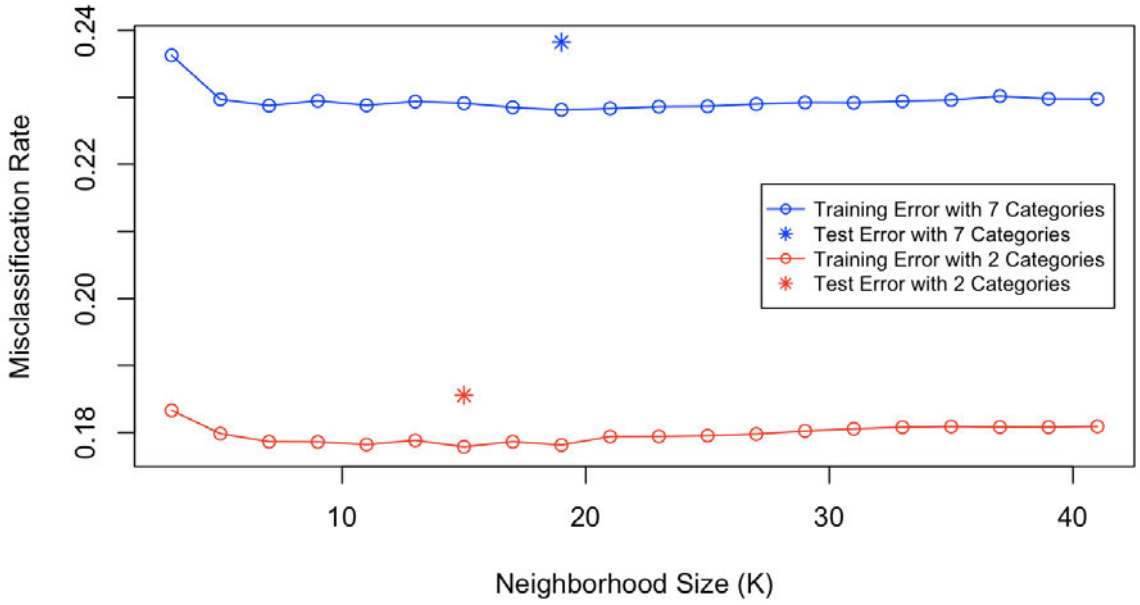
F MCA

Figure 2: MCA Variable Categories



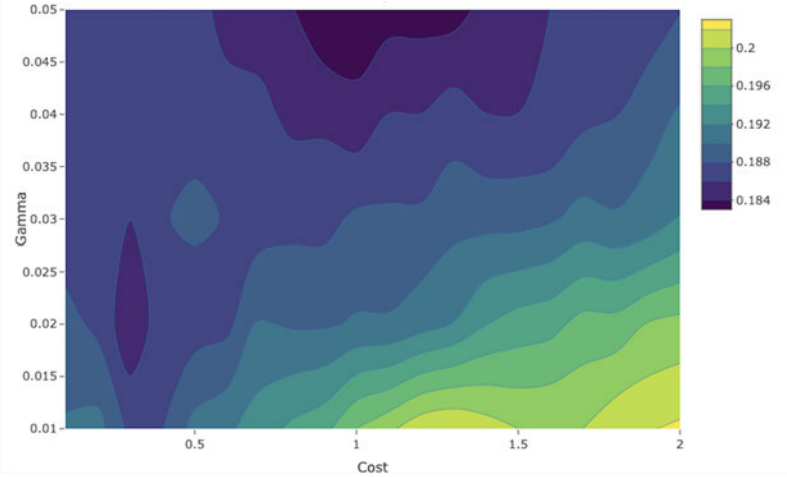
G KNN Error

Figure 3: KNN Misclassification Rates



H SVM Error

Figure 4: Test Data Misclassification Rate by Gamma and Cost Values



I Independence Model

$$logit[P(Y = dirty|X = x)] = hv009 + hv010 + hv011 + hv025 + hv219 + hv221 + hv225 + hv237 + hv243a + hv212 + hv243c + factor(hv213) + factor(hv215) + factor(sh141a) \quad (1)$$

J Foward Selection Model

$$logit[P(Y = dirty|X = x)] = factor(hv213) + hv025 + factor(sh141a) + hv009 + hv212 + factor(hv215) + hv243a + hv243c + hv221 + hv237 + hv219 + hv225 + hv025 * hv009 + hv009 * hv212 + hv025 * hv219 + hv025 * hv212 + hv025 * hv221 + hv025 * hv243a + hv025 * hv243c + hv009 * hv243a + hv025 * hv237 + hv212 * hv221 + hv025 * hv225 \quad (2)$$

K Forward Selection Limitation

R could not run forward selection with a fully saturated model that has 14 multinomial and binomial variables. To solve this problem, I permitted the forward selection model to build up to four-way interactions between the binomial variables instead of a fully saturated model. The final model produced by forward selection did not have any interaction terms of the fourth order, suggesting that AIC is minimized by a model with lower-order terms. This result suggests that not using the fully saturated model was not of significant consequence. Forward selection would not include a fifth or higher order term unless the lower order terms were present, which they are not.

L Model Performance

Table 3: Model Performance & Ranking by Test Error

Ranking	Model	Training Error Rate	Testing Error Rate
		Binomial Categorical	
1	Logistic	16.49%	17.72%
2	Ensemble	16.29%	18.41%
3	KNN	17.79%	18.55%
4	SVM	16.15%	18.69%
		Multinomial Categorical	
–	KNN	22.81%	23.82%