Morgan Durow
Salma Siddiqui
Prabhjot Singh
SI 699
Dr. Ceren Budak

Disease Analysis and Prediction for Isabel Healthcare

1. Executive Summary

Isabel Healthcare provides an online symptom checker that allows users to input their demographic information and symptoms, then predicts potential diseases. Partnering with Isabel Healthcare, our team was tasked with analyzing the demographics of users that were predicted to have rare diseases or diseases that take years to diagnose. Isabel Healthcare's goal is to better understand this subset of users and demonstrate the ability of their tool to recognize rare diseases. They hope to better market the symptom checker tool and partner with hospitals and disease support foundations. These partnerships would increase access, awareness, and usage of Isabel Healthcare's tool, giving users more well-rounded ideas of potential diagnoses based on their conditions.

For our client, we created a dashboard that allows users to determine which demographics are most heavily affected by the various rare or long to diagnose diseases. Input data from the symptom checker tool was cleaned to only include the diseases of interest given to us by our contact. The visualizations are interactive and can be filtered, such as clicking a disease to view what demographics in the user base are affected. This accomplishes the clients' goal of analyzing the characteristics of users and symptoms that make up their rare disease predictions.

As an additional task, our team also attempted to build a classifier to predict a user's disease given their demographic and symptomatic inputs from Isabel Healthcare data. We explored various methods of multi-class classification to build our own predictor. Most of the client's data was unlabeled, therefore we relied on techniques such as pseudo-labelings to increase the dataset size. We were unable to successfully rebuild a classifier for diseases due to the large number of classes in a very small number of records in the labeled data set. Predicting the specialty of the disease rather than the disease itself yielded better performance, with models such as a Ridge Classifier with cross-validation reaching 74% accuracy in predicting disease specialty. This was an easier classification task as there were only about 20 disease specialties to predict rather than over 300 unique diseases. While not ideal results, the research and work on the classifier portion of this project in addition to the rare disease analysis requested by the client helped us to experiment with real world data and explore how to work around limitations. This experience demonstrates the importance of sufficient labeled data for training models, especially for complex multi-class predictions.

2. Introduction and Motivation

Isabel Healthcare provides an online symptom checker that allows patients to input their demographic information and symptoms. The system then returns a list of potential diagnoses, ranked by likelihood. Isabel Healthcare's solution is meant to present a range of potential answers to prevent misdiagnosis and to create a foundation for conversation between doctors and patients. Just as importantly, Isabel Healthcare is designed to detect diseases that have fairly common symptoms but can take years to finally diagnose, such as Celiac Disease or Crohn's Disease.

Isabel Healthcare wants to better understand trends in demographics to identify clusters of patients that experience rare diseases. Rare diseases are more difficult for doctors to traditionally diagnose because they are not seen as often and the presenting symptoms may be mistaken for something more common. According to Isabel Healthcare, diagnostic error occurs in fifteen to twenty percent of all hospital admittances. This high rate of reported error adds stress to the patient and misdiagnosis may even be a life-threatening mistake. Error and prolonged treatment also adds cost to the healthcare system. The effects on patients and providers alike make achieving an accurate diagnosis a central issue for care providers. Data mining from healthcare data presents an opportunity to increase accuracy of diagnosis. Isabel Healthcare's mission is to use their tool to present a range of possibilities so patients and doctors alike think beyond bias and consider all potential causes of symptoms. Their goal is to use the results of the rare disease demographic analysis to partner with foundations and hospitals that provide support for people that suffer from these diseases. Presenting their tool to a wider audience would help more patients achieve a diagnosis and receive the care that they need.

A secondary goal for this project was to create a classifier that would accurately predict diseases based on the symptoms and demographics input. Isabel Healthcare's current tool scans a database of medical descriptions for matches in the user input, returning likely matches. We wanted to explore using machine learning for this task in order to better understand classification tasks.

3. Relevant Work

Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Isabel Healthcare's current system relies on an extensive curated medical database. Isabel Healthcare employees that have medical experience select records to include in their database and these records include symptoms that a disease may present and at-risk or common demographics that are affected. When a user inputs their information and symptoms into the online tool, the backend database is scanned for matches and the diseases that have more matches to the input are returned as the most likely diagnoses. Internal validation by Isabel Heatlhcare reports that a patient's actual diagnosis is in the top ten predicted diseases over ninety percent of the time. We wanted to explore machine

learning methods of disease classification and attempt to create a classifier to predict disease like Isabel's system does. To do this, we researched the current literature on supervised and semi-supervised learning to provide a baseline of understanding for our project.

Dr. Shahadat Uddin of the University of Sydney conducted a literature review on the performance of different variants of supervised machine learning algorithms for disease prediction. The compiled results analyzed the success of models such as Support Vector Machine, Naïve Bayes, and Random Forest in various publications. It was found that Random Forest models performed the best in nine of the seventeen studies (Uddin et al., 2019). This review provides a broad look at the relative performance of different variants of supervised machine learning algorithms for disease prediction and was used to aid our selection of appropriate supervised machine learning algorithms. However, we did not have sufficient labeled data to use traditional supervised methods.

4. Datasets and Data Collection

As a client project, all of our data was provided to us by Isabel Healthcare. Isabel Healthcare provided our team with the input and output data of Isabel Healthcare's online symptom checker for the years of 2018 and 2019. These datasets include the demographic information about the users such as age, gender, pregnancy status, and country of residence. Their input for symptoms is also included as well as the diseases that Isabel Healthcare's tool predicts for the user. The problem with this data is that it contains the predictions from Isabel's tool, but no actual truth to evaluate them on. Users typically do not return to the site to verify their diagnosis, leading to a shortage of labeled truth data to train any supervised machine learning models.

In addition to the extensive unlabeled data, Isabel provided us with a subsidiary ground truth data set of 562 medical records, including patient demographics and symptoms as well as a confirmed diagnosis. This dataset also details the position in which the final diagnosis was ranked in Isabel Healthcare's system, the basis for the client's internal evaluation of accuracy. There are 327 unique diseases present in this dataset, which means that many diseases may only have a single record upon which to base any model. This dataset was the foundation of baseline methods and research that our group conducted since it has explicit independent and dependent variables that are essential for evaluation, but it was also limiting due to its size, scope, and formatting.

Data cleaning was an important part of our process. The symptom information for the user was all present under one column where each of their inputs were recorded. In order to create a classifier that used individual symptoms as predictive features, we were faced with the task of splitting this single column into multiple dummy variables. For every symptom present in the 'Query Text' of the labeled dataset, we created a corresponding dummy column. The columns were binary with '1' indicating the presence of that symptom in the record and a '0' indicating its absence (Figure 1). Finally, we filtered out any columns that were not in the list of

predetermined symptoms given to us by Isabel Healthcare. This included symptom input of longer free text sentences such as “my child has difficulty being scolded by teachers” and “feeling very down and unable to make sense of anything”. These were real user inputs but were unique to individual entries rather than generalizable symptoms such as “fever” or “cough”. Ideally, we would have converted these longer free text inputs into the symptom that was most semantically similar; this is a limitation of our project. In addition to creating dummy variables for each symptom present, we also created dummy variables for the other categorical text columns such as age, gender, location, etc.

Query Text	vomiting	fever	mottled skin	headache
severe leg pain,vomiting,fever,mottled skin	1	1	1	0
headache,word finding difficulties,parietial a...	0	0	0	1
breathing problem,migraine,weight loss,arm and...	0	0	0	0
leg cut,intense pain,vomiting	1	0	0	0
headache,blurred vision	0	0	0	1

Figure 1. Example of converting original text column into symptom dummy variables (not all dummy variables shown).

5. Methods

a. Rare Disease Analysis

The datasets used for the rare disease analysis were the 2018, 2019, and rare disease datasets. The “Results” column of the 2018 and 2019 datasets detail the afflictions that the Isabel Healthcare software predicts. This list is sorted based on probability and can list up to 20 diseases/illnesses. For the sake of our analysis, we used only the top five listed diseases predicted rather than the full list, as these were the most likely. These results were filtered to return only records with the diseases that were listed in the rare disease dataset. Finally, rare diseases that appeared less than ten times in the dataset were not included in the deep analysis since there were not enough records to provide meaningful results about the populations that are affected by the diseases. The resulting data frame is the basis upon which our analysis has been conducted.

The cleaned data was analyzed to find the demographics that are most affected by certain afflictions/diseases. We wanted to create visualizations to make the results more interesting and easy to understand. Interactive visualizations for the data were created with Tableau software. We chose Tableau because it allows filtering and is easy to share and publish on the web if the client desires.

b. Classifier

Ideally, the classifier would have been built upon a large labeled data set that would allow for the use of traditional supervised machine learning algorithms. However, we only had 562 labeled records, in which there were 327 disease classes. This meant that we would have the task of accurately predicting 327 unique classes. Rather than predict the unique disease itself, we switched to predicting the specialty of the disease such as cardiovascular, infectious disease, etc. There were 24 classes that diseases were grouped into, greatly reducing the number of classes for our model to predict and therefore increasing performance. For our baseline classifier, we chose three models for the pseudo-labeling approach: K-Nearest-Neighbors, Random Forest, and Ridge Classifier. These models were chosen as they are able to handle multi-class classification as seen in literature review. KNN was our most basic, easy to interpret model. Random Forest was chosen for its performance in our literature search; it was found to be the best performing model in several classification studies (Uddin, 2019). Finally, Ridge classifier was chosen as it performs well with large numbers of classes to predict. Further analysis of algorithm performance is included in the discussion section.

While we did not have sufficient labeled data, Isabel Healthcare did provide us with unlabeled data that was input to the symptom checker tool. These records did not have a ground truth diagnosis, but they did contain demographic and symptomatic features. Our team explored several methods to work around the lack of labeled data and increase model performance.

i. Pseudo-labeling

Pseudo-labeling is a method to generate labels for unlabeled data. In this technique, instead of manually labeling the unknown points, approximate labels are generated on the basis of the labeled data. These pseudo-labels can then be treated like true value data for the purpose of training and evaluating models. This method allows for the utilization and learning from unlabeled data to increase the overall size of your dataset. First, a model is trained on the labeled data only. This model is then applied to the unlabeled set to predict outputs, or pseudo-labels. True labeled data is then concatenated with a fraction of pseudo-labeled data and retrains the model (Figure 2). In our case, we used ridge regression as our first classifier to train the model and generate pseudo labels (Hyun Lee, 2013).

We defined a function create_augmented_train() to create the "augmented training set" that consists of pseudo-labeled and labeled data. The arguments of the function are the model, training and test set information (data and features), and the parameter sample_rate. Sample_rate allows us to control the percent of pseudo-labeled data that we will concatenate with the true labeled data. For instance, setting sample rate to 0.0 means that the model will use only the labeled data, while sample_rate of 0.7 means that the model will use all the labeled data and 70% of the pseudo-labeled data. However, in both cases, the model will use all the true labeled data.

def *create_augmented_train(X, y, model, test, features, target, sample_rate): Create and return the augmented_train set that consists of pseudo-labeled and labeled data*

1. *Train the ridge regression model and create pseudo labels*
2. *Add the pseudo-labels to the test set*
3. *Take a subset of this test set with pseudo labels and append onto the original training set(true labeled data)*

return *augemented_train*

However, the pseudo-labeling method relies on the assumption that the initial model trained off of the true labels yields a high accuracy and therefore reliable pseudo-labels. Because our dataset contained 327 disease classes in 562 records, we were unable to achieve high accuracy in the initial disease prediction models. If the initial labeled data is too small in size or contains outliers, pseudo-labeling will likely assign incorrect labels to the unlabeled points. Since each class effectively has just one or two points, the model is incapable of learning the underlying structure for any class (Shenoy, 2019). Switching to predicting disease specialty rather than unique disease as the class mitigated some of this problem, though the initial model was still not perfect.

The assumption that the "messy" pseudo-labels are correct is a limitation of the project, but is the underlying assumption that must be made in order to use pseudo-labeling as a method. While the pseudo-generated labels may not be perfect, it increased the training dataset size to where we had sufficient representation of the disease specialties as classes and significantly and improved model performances.

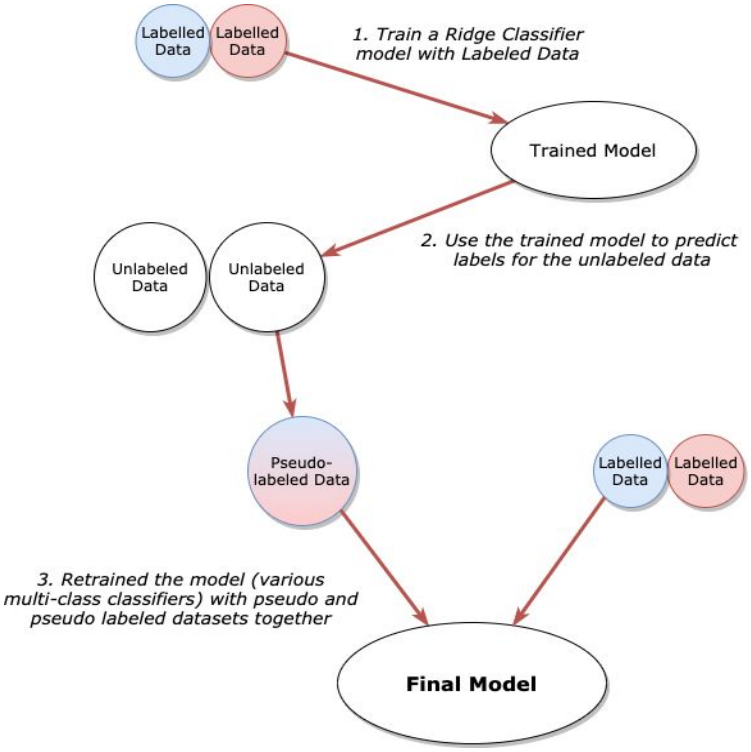

Figure 2. Pseudo-labeling method.

ii. Principal Component Analysis (PCA)

The final method we employed to improve our models' performance was Principal Component Analysis (PCA). Dimensionality reduction was important for our dataset as there were almost 1,000 unique symptoms present in our labeled data. We were treating each symptom and demographic category as its own column, leading to a dataset with a large number of features. The goal of Principal Component Analysis is to retain as much information as possible from the fewest number of principal components. This method combines highly correlated variables together to form a smaller number of variables called "principal components" that account for most variance in the data (Nagpal, 2017). For example, "fever" and "chills" may be highly correlated variables as they are typically co-occurring symptoms, and these would be combined into a single principal component. Principal Component Analysis was used on our pseudo-labeled dataset to reduce the number of dimensions and retain most information in the form of uncorrelated linear combinations of the 828 original features. Through PCA, we reduced our features from 828 to 490, explaining ninety-five percent of the variance in our data (Figure 3). The principal components extracted from the dataset were then used as the features to train our final models.

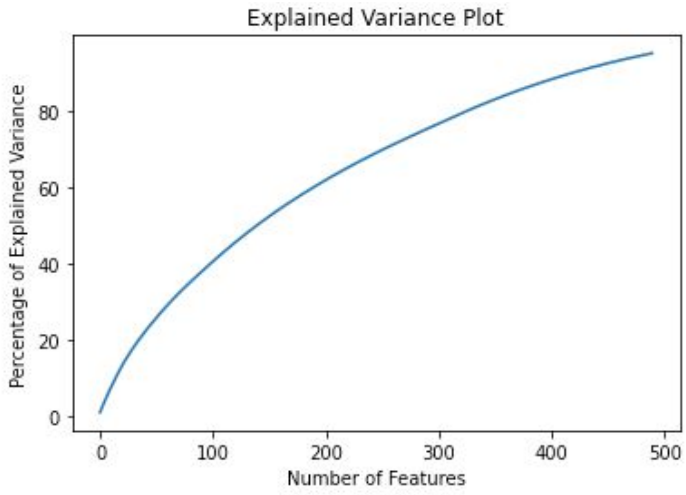

Figure 3. Variance explained by number of features.

6. Results

a. Rare Disease Analysis

Interactive visualizations for rare and hard to diagnose diseases were created in Tableau dashboards. The dashboard also allows for multiple elements to be selected, so that groups of diseases can be examined together and can be further specified to look at the effect of a disease within a certain region. Analysis shows that most users of the Isabel Healthcare tool are women, young adults 17-29 years old, and in North America (Figure 4). Overall, the user base is heavily skewed towards North American residents. This means that the proportions that emerge from the dataset require stratifying or comparison to actual population size of those regions to be useful. Based on the popularity of the Isabel Healthcare tool in North America, it is difficult to draw meaningful conclusions about areas outside of North America since the dataset for other regions is much smaller.

The most common of the rare diseases include endometriosis, depression, and Lyme disease (Figure 5). These are illnesses that are typically difficult to diagnose. For example, endometriosis is actually fairly common among women, occurring in about ten percent of the American female population (Illinois Department of Public Health, 2010). However, it is difficult to diagnose because there is a wide range of presenting symptoms and many of the symptoms are common, such as fatigue, stomach pain, and pelvic pain. Some women may even be asymptomatic but still have the disease. Social aspects can also contribute to endometriosis being a difficult diagnosis. Many complaints of pain from women are not taken seriously as menstrual pain is considered normal and doctors may not understand the extent of their

discomfort. Endometriosis was the most common of the rare diseases of interest, but this may be due to the majority of Isabel users being young adult females. Depression also had a high number of occurrences in the dataset. Individuals experiencing depression may be more open to searching for a diagnosis or treatment and start with Isabel. However, mental health issues continue to have more stigma and lack of access compared to treatment for traditional physical health issues.

Isabel Healthcare's tool is meant to present users with a range of possible diagnoses and start conversations between patients and doctors. This analysis will help Isabel to understand their user base, especially those receiving predictions for rare diseases. Our findings of young adult females being the majority of users was consistent with what Isabel has seen over the years in their own analyses. It also demonstrates to potential partners such as hospitals or disease support foundations that the tool is built to recognize rare diseases that similar online symptom checker tools may frequently miss.

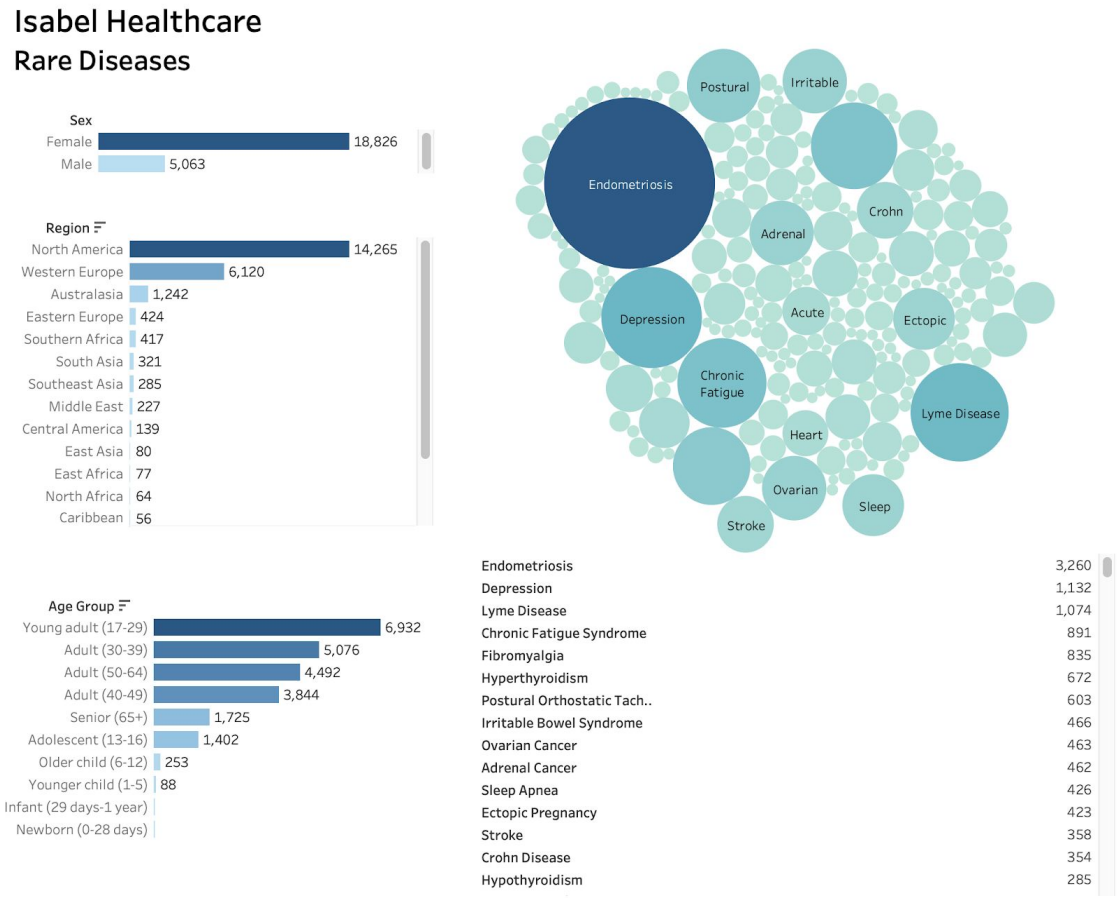

Figure 4. Final dashboard for rare and difficult to diagnose diseases. Displays demographic breakdown by gender, region, age. Also includes a bubble chart representing the most prevalent of the diseases in the list of interest.

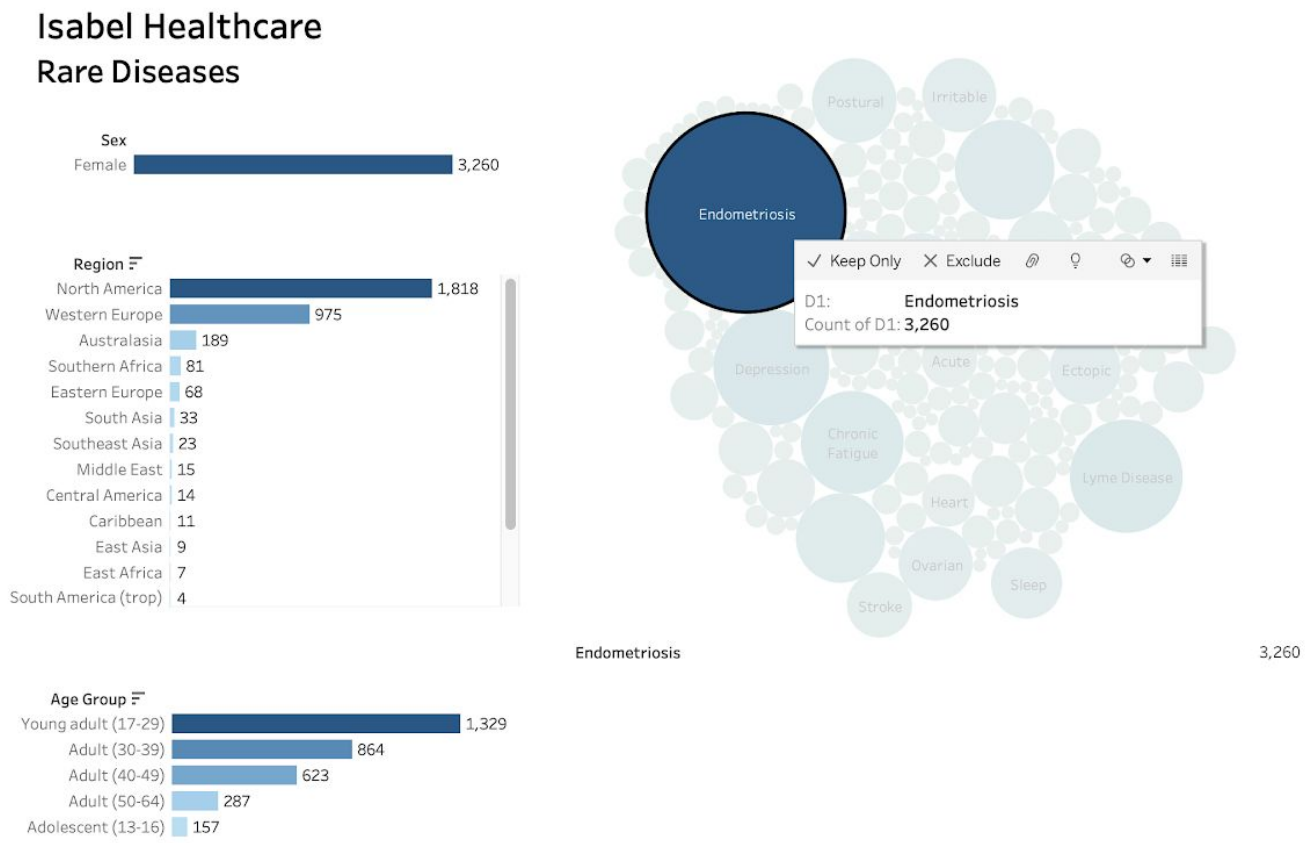

Figure 5. Example of drill down for endometriosis, the most common of the diseases analyzed. Selecting a disease filters the demographic counts. Multiple diseases can be selected for comparison.

b. Classifier

i. Baseline

Our baseline models were three machine learning algorithms: KNN, Random Forest, and Ridge. We trained the algorithms on the labeled dataset. Initially, we tried predicting disease as the class. However, the accuracy scores were very low. Instead, we switched to training the baseline models to predict disease specialty instead. Each record in the labeled dataset had information about the specialty grouping in addition to the final diagnosis. This information was valuable in producing a lower number of classes for the classifiers to predict. The results of the baseline models trained on only the labeled dataset are shown in Table 1. To improve the baseline scores, we implemented pseudo-labeling and principal component analysis, the results of which are discussed in the next section.

Table 1. Baseline Models to Predict Disease and Specialty

Model	Predict Disease: Accuracy Score	Predict Specialty: Accuracy Score
KNN	5.6%	9.8%
Random Forest	7.1%	26.2%
RidgeClassifierCV	15.6%	30.1%

ii. Pseudo-labeling + Classifier

With the labeled data and the training model (Ridge Regression) as input to the pseudo-labeling model, the function '_create_augmented_train()' the final pseudo-labeled data was generated. This dataset was then fed to the classifiers to predict disease speciality for which the performances are as shown in the following Figure 6 ('Pseudo-labeling + Classifier' row). Random Forest was the worst with F-1 score of 0.28 while Ridge Regression (L2 regularization) performed best with an F-1 score of 0.73. Random Forest seems to overfit the data as it manages to predict only two of the sampled classes (out of 24) which suggests it predicts multiple classes with a low recall score (refer to F-1 score summary in the appendix). On the other hand, Ridge regression with the L2 regularization balances the goodness of precision, recall and accuracy scores and is more reliable for a high-degree multiclass classification task such as this one.

iii. Pseudo-labeling + Classifier + PCA

The top row (for each classifier) in the following figure is the final and the most optimized solution amongst all the others mentioned in preceding sub-sections. Implementation of PCA reduced the number of required dimensions for the pseudo-labeling model from 828 to 494. These new dimensions are actually the transformation of the input data into a new space with uncorrelated dimensions (494) - by linearly combining the original dimensions (828). Therefore, as shown in the graph, KNN and Random Forest show significant improvements (as the number of dimensions to train the model on decrease), whereas Ridge regression had no significant difference in its performance.

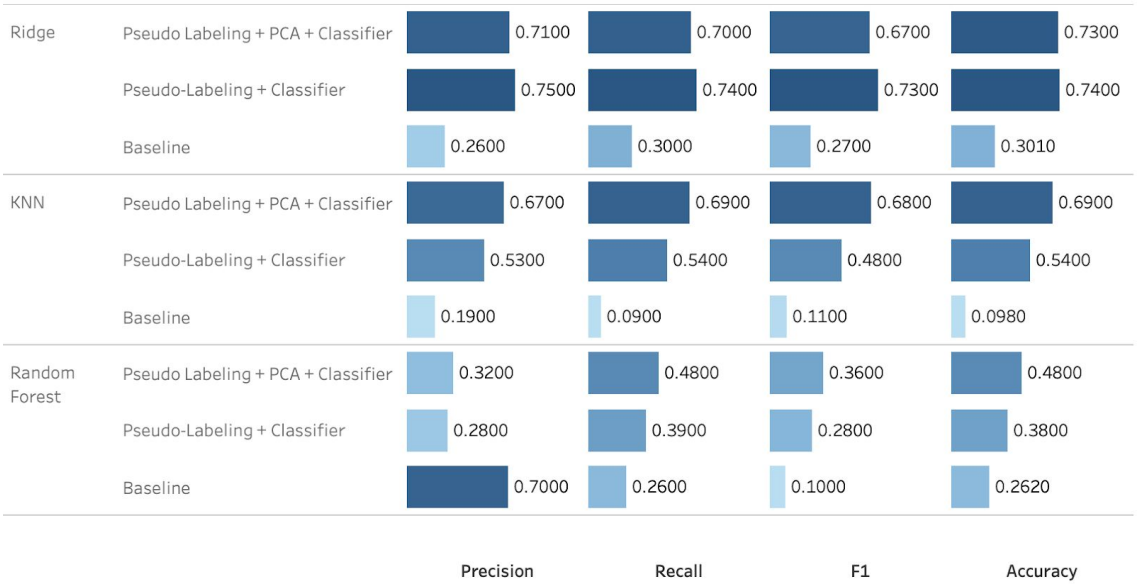

Figure 6. Model performance comparison graph.

7. Discussion

a. Algorithms

Initially, we chose the three algorithms Random Forest, Ridge, and KNN for their various properties in multiclass classification.

Random Forest was chosen as part of our project because of the success of several medical diagnosis classification studies in using Random Forest for a similar healthcare data problem (Uddin, 2019). However, Random Forest performed poorly in terms of both F-1 and accuracy scores. The results for both scores improved after applying PCA, though it was still the lowest performing of the three models. This is suspected to be because there are fewer features that have to be taken into account by the many decision trees that make up the Random Forest (Prinzie et al. 2008). Before consolidating the features, the model may have been applying the most common label to the specialty column, which would account for the poor accuracy and F-1 scores in the baseline models.

K-Nearest Neighbors is an unsupervised machine learning algorithm that does not depend on training data for generalization and considers the closest neighbors of a datapoint in creating predictions (Kim et al., 2020). It performed best when PCA and pseudo-labeling were applied to the model, since it is not inherently optimized for multiclass classification (Guney & Atasoy, 2012). It performed the best after Ridge regression.

Like Random Forest, Ridge Regression also harnesses the power of logistic regression in classification. Ridge is a strong performer in multiclass classification and does not see much improvement between our baseline models and our final models in terms of F-1 and accuracy

scores. This is because the L2 cross-entropy loss function is already optimized for multiclass classification and does not need the additional feature consolidation created by PCA to improve scores (Student & Fujarewicz, 2012).

b. F-1 vs. accuracy

In theory, Accuracy is used when true positives and true negatives are more important or when the class distribution is similar, while F1-score is used when the false negatives and false positives are crucial or when there are imbalanced classes, as in our case (Analytics Vidya). Our client's tool is a real-world self diagnostic tool which is free to use by any user, any number of times. However, some users tend to enter different symptoms each time they run the diagnosis (sometimes, multiple runs in one sitting!) - our data (extracted from the aforementioned tool) pertains to such iterative usage and this nature of our data is a potential cause of sampling bias amongst disease specialities (some are oversampled as opposed to others). Therefore, in this case, F1-score was sought to be a better metric to evaluate our model on as it accounts for the class imbalance and calculates precision and recall scores for each predicted class as opposed to just the total number of true positives and true negatives for the entire set of predictions the accuracy metric.

8. Conclusion

Analysis of rare disease predictions by the Isabel Healthcare tool helps to better understand the user base and verify that what Isabel predicts is in line with known medical cases. By identifying the demographics of people most likely to be afflicted by rare or hard to diagnose diseases, users can be informed that they are at-risk based on their symptoms/demographics and can seek medical attention immediately. It also helps to inform the users of these potential ailments so that they can discuss it with their medical providers. Isabel Healthcare is in a unique position to empower patients to advocate for their own health and potential risks by informing them of diseases that are uncommon and would otherwise take years to diagnose. With this helpful information, patients can raise these issues with their doctors and end their suffering much earlier than it would normally take for the diseases to be diagnosed.

This is also beneficial information for Isabel Healthcare to use for marketing and advertising campaigns. By identifying the groups of people most at-risk, Isabel Healthcare can appeal to those users on web pages frequented by the individual groups. Hospital websites and disease foundations websites may link to the symptom checker tool as a first step for a user to see if they may be affected by a disease or illness. By presenting their tool to a wider audience, more users can get answers about diagnoses to help relieve ailments and make a difference in their lives. It is clear from our analysis that diseases such as endometriosis, depression, and Lyme disease afflict a large group of users. It is important to consider that there are many others who suffer from the same diseases but are not using the Isabel Healthcare tool. Isabel Healthcare

has an opportunity to use demographic analysis as a means of finding these people and marketing itself to them so that they may receive treatment for their ailments in the future.

From our research on classifiers and applications to this dataset, we can also conclude that the Ridge regression algorithm works well for multi-class classification problems. Data mining in the medical field is emerging as a way to better treat patients and their diseases, but issues with data privacy, complex classification tasks, and high permutations or combinations of diseases affect the potential for wide-scale applications.

With more time and data, we would have liked to create a knowledge graph that links symptoms and disease to visualize diseases that have a relationship to each other or are alike in their symptoms and potentially their treatments as well. We also would have liked to consolidate the free text into a subset of medical terms that encapsulate their common text counterparts to stabilize the performance of our models.

9. Appendix

Isabel Healthcare also gave the team two additional years of data for 2017 and 2016. However, their algorithm for ranking the probability of each disease was altered significantly in 2018 and proved to be more accurate than it had previously performed. Therefore, the datasets for 2016 and 2017 were ultimately left out of the analysis and model training to preserve accuracy.

The ground truth dataset only contains 562 records, the majority of which are classified as infectious diseases. We faced a further limitation as only 5 of the 67 rare diseases that the client wants examined are included in the ground truth dataset, making it difficult to evaluate anything related to rare diseases using the ground truth.

Another task that we thought to undertake was creating a knowledge graph that linked rare diseases and symptoms as a means of clustering related diseases to each other. This was ultimately left unfinished and not presented to the client because of technical limitations.

For our first baselines, our team used disease “Specialty” as a target to predict rather than disease itself. We applied basic classifiers such as Naive Bayes, KNN, and SVM on the ground truth dataset as a basic means of identifying how a classifier can be applied to the symptoms and demographic information to predict the general category into which a disease falls (infection, cardiovascular, etc.). This approach was used rather than targeting disease as the final output because of the small data size and concern over not being able to fit the model properly.

The following screenshots show the f1-score characteristics for the three classifiers that we used.

	precision	recall	f1-score
ALLERG	0.00	0.00	0.00
CARDIO	0.00	0.00	0.00
DERM	0.00	0.00	0.00
EAR	0.00	0.00	0.00
ENDO	0.00	0.00	0.00
GASTRO	0.00	0.00	0.00
GENE	0.00	0.00	0.00
HEMAT	0.00	0.00	0.00
HEPATO	0.00	0.00	0.00
IMMUN	0.52	0.99	0.68
INFECT	0.29	0.57	0.39
METAB	0.00	0.00	0.00
NEOPL	0.00	0.00	0.00
NEPHRO	0.00	0.00	0.00
NEURO	0.00	0.00	0.00
OBGYN	0.00	0.00	0.00
OPHTHAL	0.00	0.00	0.00
ORTHO	0.00	0.00	0.00
PSYCH	0.00	0.00	0.00
RESP	0.00	0.00	0.00
RHEUM	0.00	0.00	0.00
SHOCK	0.00	0.00	0.00
SOCIAL	0.00	0.00	0.00
TOXIC	0.00	0.00	0.00
TRAUMA	0.40	0.99	0.57
UROL	0.00	0.00	0.00
VASC	0.00	0.00	0.00
accuracy			0.40
macro avg	0.04	0.09	0.06
weighted avg	0.19	0.40	0.25

	precision	recall	f1-score
ALLERG	1.00	1.00	1.00
CARDIO	0.62	0.29	0.39
DERM	0.50	0.33	0.40
EAR	0.92	0.96	0.94
ENDO	0.89	0.65	0.75
GASTRO	0.79	0.60	0.68
GENE	0.50	0.50	0.50
HEMAT	0.60	0.38	0.46
HEPATO	0.40	0.67	0.50
IMMUN	0.67	0.98	0.80
INFECT	0.82	0.69	0.75
METAB	0.50	0.71	0.59
NEOPL	0.68	0.49	0.57
NEPHRO	0.79	0.47	0.59
NEURO	0.83	0.30	0.44
OBGYN	0.60	0.43	0.50
OPHTHAL	1.00	0.67	0.80
ORTHO	0.73	0.73	0.73
PSYCH	0.83	0.60	0.69
RESP	0.76	0.63	0.69
RHEUM	0.57	0.35	0.43
SHOCK	0.77	0.71	0.74
SOCIAL	1.00	1.00	1.00
TOXIC	1.00	0.29	0.44
TRAUMA	0.67	0.98	0.80
UROL	0.83	0.98	0.90
VASC	0.77	0.50	0.61
accuracy			0.72
macro avg	0.74	0.63	0.66
weighted avg	0.73	0.72	0.70

	precision	recall	f1-score
ALLERG	0.50	1.00	0.67
CARDIO	0.33	0.39	0.36
DERM	0.33	0.33	0.33
EAR	0.67	0.67	0.67
ENDO	0.56	0.48	0.52
GASTRO	0.69	0.52	0.59
GENE	0.50	1.00	0.67
HEMAT	0.40	0.25	0.31
HEPATO	0.25	0.33	0.29
IMMUN	0.69	0.93	0.79
INFECT	0.55	0.40	0.46
METAB	0.55	0.86	0.67
NEOPL	0.57	0.60	0.58
NEPHRO	0.66	0.47	0.55
NEURO	0.59	0.40	0.48
OBGYN	0.43	0.43	0.43
OPHTHAL	0.00	0.00	0.00
ORTHO	0.70	0.47	0.56
PSYCH	0.71	0.52	0.60
RESP	0.59	0.43	0.50
RHEUM	0.58	0.53	0.55
SHOCK	0.57	0.43	0.49
SOCIAL	0.00	0.00	0.00
TOXIC	0.00	0.00	0.00
TRAUMA	0.66	0.90	0.76
UROL	0.93	0.59	0.72
VASC	0.64	0.45	0.53
accuracy			0.63
macro avg	0.51	0.50	0.48
weighted avg	0.62	0.63	0.61

Figure 7. F-1 characteristics for random forest, ridge regression and kNN (left to right)

10. Citations

Dong-Hyun Lee “Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks”. http://deeplearning.net/wp-content/uploads/2013/03/pseudo_label_final.pdf

Facts about Endometriosis. Illinois Department of Health, www.idph.state.il.us/cancer/factsheets/endo.htm.

Güney, Selda, and Atasoy, Ayten. “Multiclass Classification of N-Butanol Concentrations with k-Nearest Neighbor Algorithm and Support Vector Machine in an Electronic Nose.” Sensors & Actuators: B. Chemical 166–167 (20/5/2012).

Pseudo-labeling a simple semi-supervised learning method. https://datawhatnow.com/pseudo-labeling-semi-supervised-learning

Kim, Justin, Xu, Ziyu, and Singh, Shashank. “Multiclass Classification via Class-Weighted Nearest Neighbors,” 9/4/2020.

Prinzie, Anita, and Van den Poel, Dirk. “Random Forests for Multiclass Classification: Random MultiNomial Logit.” Expert Systems With Applications 34, no. 3 (2008): 1721–32.

Nagpal, Anuja. “Principal Component Analysis- Intro.” Medium. Towards Data Science, November 22, 2017. https://towardsdatascience.com/principal-component-analysis-intro-61f236064b38.

Shenoy, Anirudh. “Pseudo-Labeling to Deal with Small Datasets - What, Why & How?” Medium. Towards Data Science, December 3, 2019. https://towardsdatascience.com/pseudo-labeling-to-deal-with-small-datasets-what-why-how-fd6f903213af.

Student, Sebastian, and Fujarewicz, Krzysztof. “Stable Feature Selection and Classification Algorithms for Multiclass Microarray Data.” Evaluation Studies. Biology Direct 7, no. 1 (2/10/2012).

Uddin, Shahadat et al. “Comparing different supervised machine learning algorithms for disease prediction.” BMC medical informatics and decision making vol. 19,1 281. 21 Dec. 2019, doi:10.1186/s12911-019-1004-8