Text classification of UK smallholding communities through Twitter

Samuel Munaf

a.s.munaf@stir.ac.uk

University of Stirling

Kevin Swingler

University of Stirling

Franz Brülisauer

SRUC Veterinary Services, Scotland's Rural College (SRUC)

Anthony O'Hare

University of Stirling

George Gunn

Scotland's Rural College (SRUC)

Aaron Reeves

RTI international

Research Article

Keywords: Smallholdings, backyard keepers, text classification, social media

Posted Date: March 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2670842/v1

Additional Declarations: No competing interests reported.

Abstract

Within the UK, livestock holdings are registered so that livestock can be traced, and animal diseases be controlled. These regulations are enforced irrespective of farm size, however, tend to be better followed on traditional farms, whereas holdings new to keeping livestock are less likely to be aware of their obligations. These smallholdings thereby may evade registration and are less likely to participate in national disease surveillance and ultimately complicate national animal disease control. Less information is known about small-scale livestock keepers, in particular those without a traditional farming background. Smallholders have been known to play a vital role in zoonotic disease outbreaks and more action needs to be taken to improve surveillance systems by incorporating this demographic into current intelligence. Literature indicates that parts of these communities often utilise social media as a means of communication and information sharing. Twitter followers from a prominent smallholder user in the UK were extracted and manually categorized as a smallholder or not, based on profile descriptions. Manual coding of just under 1,000 Twitter profiles was conducted to build a robust training dataset. Text classification algorithms were applied on this annotated data, and the resulting classification algorithms produced accuracies of over 80%. Results indicate that classification can prove to be a highly successful tool, if a sufficient training dataset is curated, and there is enough textual information within the user profiles on social media.

Introduction

A smallholding can be defined as a home with an adjacent piece of land that is used to rear livestock or/and plant crops. The land is usually less than 200,000 square meters, or 50 acres, and yet the distinction between a garden, small-scale farm and smallholding can become ambiguous as there is no standardized definition[1].

Smallholders are often linked with growing their own food, or keeping a small number of livestock either as pets or as a hobby to produce meat, breed pedigree animals or landscape management[2]. Small-scale food production may also be the result of geographic circumstances e.g., where poor quality ground only supports small numbers of livestock as is the case in large parts of Scotland. Another form of smallholding is referred as to crofting, often found in the highlands and Islands of Scotland[3]. Crofting is a small-scale agricultural land use and tenure system practiced in Scotland's Highlands and Islands, as well as sections of Ireland, England, and other regions of the world. It entails subsistence farming or small-scale commercial agriculture on tiny parcels of land known as crofts. They are distinguished by a strong feeling of community, with crofters frequently cooperating to share resources and expertise. An overlap between this community and smallholders exists, creating another potential avenue for disease transmission[4].

The smallholding lifestyle has increased in popularity since the early 2000’s and falls under the concept of self-sustenance[5]. This may be partially driven by shifts in societal concerns regarding animal welfare, quality, and food production. The low barriers to entry results in the market being open for the majority of

the population, if they have adequate land available[6]. Smaller livestock, namely small ruminants and poultry are easier to procure and keep. Furthermore, government regulations are less likely enforced on holdings that don't receive financial support under the rural payment scheme. This is referred to as the Basic Payment Scheme, and provides more of an incentive for larger farms to comply with government regulations as they are monetarily rewarded[7]. Farms receiving payments are subject to cross compliance visits, which entails both their veterinary records and cattle movement records being inspected[8]. Within the UK and western Europe, most households don't have to rely on backyard livestock and crops as the only means of sustenance, given the abundance of commercial farms and global food chains. In contrast, many households in Asia, Africa and South America depend on such means as both food and income[9].

The regulations within the UK state that livestock holdings must be registered so that livestock can be traced, and animal diseases be controlled. This includes all bovine species (cattle, bison, buffalo), sheep, goats, deer, pigs, and poultry. The databases surrounding species type, location and farm size are curated by authorities directly or their contractors[10] (E.g. ScotEID for cattle, sheep and pigs in Scotland)[11]. This system allows for accurate accounting of livestock holdings, their size, and livestock movements. This system is closely linked to the Rural Payments Service, compliance amongst larger, commercial livestock holdings therefore is high. In contrast smallholders, in particular those not relying on the economic output of their livestock, may fail to register their holding. This is exacerbated by the fact that registration of poultry flocks comprising of fifty or fewer is voluntary. This creates challenges when it pertains to disease monitoring and control in these communities. A pertinent example of this relates to the recent story of Geronimo the Alpaca, who tested positive for bovine tuberculosis (bTB), a disease that may be spread to other animals and people, according to the culture findings. But, Geronimo's owner, who has been struggling for his life, questioned the veracity of the test results and requested more testing. The case has prompted a controversy about the usage of bTB testing in alpacas as well as the ethics of euthanizing an animal that may or may not be sick[12].

Disease surveillance becomes difficult when dealing with smallholders as livestock keepers may be ignorant/not incentivized to obey rules surrounding biosecurity, in addition to the difficulty to include them in control efforts e.g., all poultry flocks close to a confirmed bird flu case irrespective of their size need to be visited by international law to keep up international trade[4]. The lack of knowledge surrounding where and how individuals receive their livestock knowledge complicates our understanding of the role of these communities in regard to scanning surveillance[13].

Recent outbreaks in Avian influenza across Europe have reinforced that a holistic approach is needed when it comes to robust disease surveillance, which includes the roles of both commercial and smallholding farms[14]. The part that smallholder poultry keepers play in the spread of such diseases remains underdetermined, as little information is known about this demographic. International research implies that in some nations, they have been observed in being a contributing factor to the spread both Newcastle disease and Avian Influenza[2].

Literature indicates that these communities often use social media as an avenue for communication, biosecurity advice, rules, regulations, and new practices, and tend not to always engage with their vets, livestock organisations or trade bodies[2]. Understanding this demographic through social media provides a method of gathering passive surveillance related to common themes/topics of discussion, locations, how they interact with government bodies and others, and how they respond to outbreaks/guidelines/regulations. Gathering such vital information will then allow for augmenting current knowledge databases[15]. Communities of smallholders exist in farming forums, Facebook groups and Twitter page followers, and act as a repository for information with the potential to be extracted, analyzed and use to inform policies[2].

Social media is a medium of communication exchange on a global scale, and allows for the immediate notification of news, events, and topics[16]. Within this platform, communities of individuals exist through follower networks centering around common subjects and topics[17]. These networks sometimes operate as echo chambers that consist of like-minded individuals sharing their opinions and information. Twitter is amongst the most prominent social media applications used by almost 400 million users worldwide and acts a microblogging site which limits users to 280 characters per message[18].

Text classification has historically centered around spam detection in emails, bot detection in social media and sentiment analysis of messages, as a way of segmenting relevant content from irrelevance[19]. This involves a supervised machine learning approach by creating a sufficient training dataset, consisting of manually coded text, that can then be inputted into a classification algorithm (e.g., Support vector machine). Performance metrics are then calculated, and the accuracy of the algorithm can be determined. As social media has a propensity to contain large amounts of noise and “trolls” who spread misinformation, the first step in any sound public health research of disease outbreaks is to determine the initial source of the information (e.g., academic, health professional, public)[20].

Within public health, studies have employed such approaches in categorizing illnesses amongst the population through text classification methods[21]. This included differentiating between dementia and Alzheimer’s based on tweet content and matching keywords to domain-specific dictionaries. User classification has also been adopted to distinguish the type of entity that the profile or/and tweet originates from. Alsudias & Rayson looked at categorizing users who disseminated content pertaining to Infectious diseases into “Academic”, “Media”, “Government”, “Health professionals” and “public”. Multiclass classification studies such as this require a greater training dataset, as a balanced distribution for each class is preferable[20]. Similar work has not been replicated in the veterinary domain, however similar approaches using with electronic health veterinary records have been applied for the automatic labelling of disease based on clinical signs[22].

This study examines the feasibility of distinguishing Smallholding Twitter user profiles through the descriptions written in the 160-character limit “about me” blurb. A manually classified dataset was curated to build a training dataset with a binary 1 or 0 being coded for a Smallholder and non-

smallholder respectively. Text classification algorithms were applied on this dataset to determine how accurately they could classify the two groups. This would then be used as a predictive tool to allow for the instant classification of a smallholding twitter user, which can then be examined further to determine location, topics discussed and influence within the networks. Most previous studies have focused on the content of the tweet itself, and not on the profile metadata available.

Methodology

Data collection

The Twitter API was applied in Python through the Tweepy package[23], which requires developed-level access to extract information. Tweepy has a maximum limit of 1000 for the extraction of followers. An influential UK based smallholder was selected as the focus of this study, and their user ID was obtained allowing for the scraping of their entire follower network (n = 953). Information pertaining to profile name, location and description were extracted for each of these followers and downloaded in CSV format for annotation.

User categorization

953 distinct users were categorized using the description provided in their profile, and in some cases, the username was also considered if “smallholder” or variations of this term made up part of the name (only performed on one instance). An abridged example of the type of users labelled as “1” (Smallholder) or “0” (Not smallholder) can be seen in Table 1: (Not actual tweets verbatim, but words have been slightly changed to preserve anonymity).

Table 1
Example of coded user profile descriptions

Profile description	Coded label
“We keep chickens and make soap from ….”	1
“Smallholding based in …”	1
“Writer, nature lover…”	0
“Farming equipment and machinery…”	0

As the incorporation of the description column is a vital element to this project, any followers who left this blank were removed from the analysis, reducing the total to 774. Those users coded as Smallholders (“1”) accounted for 26% of the total.

Inter-rater agreement

To assess the agreement percentage between the two manual coders, Cohen’s Kappa statistic was applied[24]. The results show the Kappa statistic as 0.87, indicating a strong agreement amongst the two

manual coders.

Data preprocessing

Upon completion of the annotation, the data was imported into Python and filtered through a text cleaning process. Free text social media data often contains non-characters, such as emojis and symbols therefore needs sufficient cleansing before analysis.

The Natural Language ToolKit (NLTK) was utilised, and the profile descriptions went through the following process:

- Non-characters, punctuation and numbers removed
- Converted to lower case
- URL links removed
- Stopwords removed
- Removal of any white spaces
- Stemmed
- Lemmatized
- Normalized

Data analysis

Descriptive statistics and word clouds were produced for the smallholding group to visualize the common occurring terms in this cohort.

Feature selection

Feature selection was performed by both the TF-IDF and Word2Vec word embedding vectorizers, with the former being chosen as the preferred method.

Dataset balancing

As the Smallholding class only accounted for 26% of the total frequency, two methods of correcting this imbalance were utilised. Firstly, under-sampling of the dominant label (0) using the NearMiss (version 1) method which selects samples from the dominant class with a minimum average distance to the three nearest minority label examples. In contrast, Synthetic minority oversampling (SMOTE) was used for oversampling of the minority class (1), which duplicates the examples to synthesize new examples and augments the dataset.

Text classification

An 85:15 training/test split was performed on the dataset, and five separate classification models were initialized using the Sklearn package in Python;

- K-nearest neighbor (KNN)
- Multinomial Naïve Bayes (MNB)
- Decision tree (DT)
- Logistic regression (LR)
- Random Forest (RF)
- Bagging classifier (BGC)
- Support Vector classifier (SVC)

Hypertuning of parameters was conducted through the grid search method to optimize the input parameters. Performance metrics for each algorithm were assessed through accuracy, precision, recall and the F1-score. An Area under the ROC curve (AUC) plot was visualized to measure how well the models could distinguish between the two classes.

Results

Results from the Word cloud in Fig. 1 depicts that a large proportion of smallholders usually mention this within their profile, in addition to the livestock they keep (i.e., Sheep, chick, pigs, goats). This demonstrates that a dictionary of common terms can be created from the user profile descriptions and mapped against any new data passed through the algorithms via a simple word matching.

Due to the large imbalance between the classes (74% “0” and 26% “1”), results for the Precision, Recall and F1-score were derived from the weighted average metric. This is the preferred method when dealing with an imbalance in class sizes as the metric accounts for the contribution of every class as weighted by the frequency of instances in that class[25].

Table 2
Performance metrics on unbalanced dataset

Model	Accuracy	Precision	Recall	F1-score
KNN	0.83	0.79	0.68	0.71
MNB	0.80	0.72	0.73	0.73
DT	0.81	0.74	0.77	0.75
LR	0.84	0.82	0.69	0.72
RF	0.81	0.76	0.64	0.67
BGC	0.80	0.72	0.70	0.71
SVC	0.85	0.83	0.71	0.74

Table 3
Performance metrics from undersampling

Model	Accuracy	Precision	Recall	F1-score
KNN	0.74	0.66	0.69	0.67
MNB	0.70	0.65	0.70	0.65
DT	0.81	0.74	0.77	0.75
LR	0.83	0.76	0.79	0.77
RF	0.84	0.77	0.78	0.77
BGC	0.81	0.74	0.75	0.74
SVC	0.81	0.74	0.75	0.74

Table 4
Performance metrics from oversampling

Model	Accuracy	Precision	Recall	F1-score
KNN	0.23	0.12	0.50	0.19
MNB	0.73	0.67	0.72	0.67
DT	0.80	0.73	0.76	0.74
LR	0.86	0.83	0.76	0.78
RF	0.84	0.78	0.74	0.75
BGC	0.83	0.76	0.77	0.77
SVC	0.86	0.82	0.77	0.79

Table 2 displays the accuracy, precision, recall and F1-score for all seven algorithms for the imbalanced dataset. The highest accuracy was achieved by SVC with 85%, with a corresponding Recall score of 0.71. This was followed by both LR and KNN, achieving 84% and 81% accuracy respectively. MNB and BGC yielded the lowest accuracy. The model with the highest precision and recall would be classed as the most useful in this study, which is DT (F1 score 0.75). In binary classification, the F1-Score measures the accuracy by considering both precision and recall and provides a weighted average of both. RF performed the worst regarding the F1-score with 0.67.

Table 3 depicts the under-sampled performance, which saw a drastic decrease in accuracy for both KNN and MNB. RF went from having the lowest F1 score in the imbalanced dataset, to the joint highest with 0.77.

Oversampling as presented by Table 4, proved to increase the performance of most algorithms, except KNN, which performed significantly worse with only 23%. Sampling adjustments severely affected KNN

as compared to the imbalanced dataset and could potentially be caused by the parameter turning. SCV possessed the highest accuracy and F1-score amongst all the models and sampling adjustments.

The AUC curve shown in Figs. 2–4 offers a cumulative measure of performance for both classification thresholds. It is the likelihood that the algorithm ranks a random positive (smallholder) higher than a random negative (non-smallholder)[26]. Results from all the curves display that RF achieved the highest AUC scores and DT achieved the lowest.

A heat map confusion matrix was produced for the highest performing model, SVC, depicting the count of accurate and inaccurate predictions for each label. Figure 5 indicates that the model performed substantially better when it pertained to the classifying true negatives (0), rather than the true positives (1), as indicated by the lighter colours in the heatmaps.

Discussion

The results from this analysis, and from relevant literature indicate that social media data can be an effective tool to categorise users into communities, based on their profile descriptions. This study attempts to bridge the knowledge gap of smallholders within the UK by using open-source, publicly available data sources to identify them initially, creating a framework for more granular research to be conducted. High accuracies were achieved with all the algorithms applied on the data, and useful predictive model was created based on a well annotated training dataset.

A large proportion of research has been conducted to account for the entire content of the tweet, or user timeline, as variable for analysis, whereas this study highlights the efficacy of only using the 160-character profile description.

Once this demographic has been identified, topic modelling of user timeline content, in addition to network analysis can be used as avenues of exploration for better insights into what these communities discuss, and how they interact amongst each other. Visualizing these interactions portrays a clearer picture of the influential users in the networks, the dissemination of livestock-based information, the source of this information, and sentiments regarding new government guidelines.

One of the secondary aims of this study is to create a smallholder database large enough to be useful for disease surveillance, as very little is known about this demographic. This can be achieved by inserting new data into the pre-trained model and extracting all the users who were classified as smallholders, along with their locations.

Using the trained models created in this study, new datasets can be extracted through the same method described previously in section 2.1 from various smallholding-related accounts within the UK. The profile descriptions can be passed through the trained models and subsequently categorized as 1 or 0. The finalized dataset would be filtered to retain only those users categorized as smallholders (1). The location field would be a crucial element of this and would rely on users to fill this in from the free-text

option. Precise geolocation on Twitter is almost impossible to achieve as co-ordinates were disabled due to user privacy concerns. A large proportion of users in this study dataset had filled in their location option, and this can be extracted and mapped to potential disease outbreak hotspots.

Conclusion

The dataset does highlight the need to address the imbalance of the classes, which can be achieved through either under-sampling or oversampling techniques. The former works by eliminating portions of data from the larger class (i.e. non-smallholders), while the latter generates additional instances for the smaller class (i.e. smallholders). Literature indicates that these methods don't always improve model performance, and in some instances, cause them to perform poorer (especially under- sampling)[20]. Some results also corroborate this point, as witnessed in the performance by KNN after oversampling. Oversampling is regarded as the preferred option when dealing with smaller datasets, reinforced by the performance of the SVC model.

Further analysis can be performed by attempting to extract species types from the profile description or twitter timeline too, to understand whether they are a single species farm or mixed farm. Additionally, a deeper dive into the content of the user timeline, or retweets, can be performed to gain a greater understanding of the topics discussed.

Declarations

Ethics statement

Developer level access was permitted by Twitter and obtained in October 2019, granting administrative permission to access the raw twitter data. All data was anonymised prior to analysis. No user identifiable data was scraped, and all text was aggregated and analysed together, hence no individual can be identified from the results.

Ethical approval was granted by the University of Stirling's General University Ethics Panel (GUEP) and conformed to the research integrity policies.

All methods were carried out in accordance with relevant guidelines and regulations surrounding social media scraping and analysis provided by UK research and Innovation (UKRI). Further information can be found here: https://www.ukri.org/councils/esrc/guidance-for-applicants/research-ethics-guidance/internet-mediated-research/

Informed consent

The need for informed consent was waived by the University of Stirling's General University Ethics Panel (GUEP).

Consent for publication

Not applicable

Availability of data and materials

All data was scraped from the public domain and can be requested by contacting the corresponding author.

Competing interests

The authors declare that there no competing interests.

Funding

Not applicable.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by SM. Manual coding of the dataset was performed by SM and FB. The first draft of the manuscript was written by SM and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Corresponding author

Correspondence to Samuel Munaf: a.s.munaf@stir.ac.uk

References

1. Addland. Addlands' Guide to Smallholdings. Arbtech. Published July 28, 2021. Accessed January 4, 2023. https://arbtech.co.uk/a-guide-to-smallholdings/
2. Correia-Gomes C, Sparks N. Exploring the attitudes of backyard poultry keepers to health and biosecurity. *Prev Vet Med*. 2020;174:104812. doi:10.1016/j.prevetmed.2019.104812
3. Nature.scot. Crofting. NatureScot. Published January 31, 2023. Accessed March 20, 2023. https://www.nature.scot/professional-advice/land-and-sea-management/managing-land/farming-and-crofting/types-farming/crofting
4. Delabouglise A, Thanh NTL, Xuyen HTA, et al. Poultry farmer response to disease outbreaks in smallholder farming systems in southern Vietnam. Davenport MP, Schiffer JT, Borremans B, Rist C, Garchitorena A, eds. *eLife*. 2020;9:e59212. doi:10.7554/eLife.59212
5. Farming UK team. Half of Brits want to quit the rat race and own a smallholding. Published 2019. Accessed March 20, 2023. https://www.farminguk.com/news/half-of-brits-want-to-quit-the-rat-race-

and-own-a-smallholding_51118.html

6. Fan S, Rue C. The Role of Smallholder Farms in a Changing World. In: Gomez y Paloma S, Riesgo L, Louhichi K, eds. *The Role of Smallholder Farms in Food and Nutrition Security*. Springer International Publishing; 2020:13-28. doi:10.1007/978-3-030-42148-9_2
7. gov.uk. Basic Payment Scheme. GOV.UK. Published March 16, 2023. Accessed March 20, 2023. https://www.gov.uk/guidance/basic-payment-scheme
8. gov.uk. Cross compliance 2022. GOV.UK. Published 2022. Accessed March 20, 2023. https://www.gov.uk/guidance/cross-compliance-2022
9. Smallholders produce one-third of the world's food, less than half of what many headlines claim. Our World in Data. Accessed January 4, 2023. https://ourworldindata.org/smallholder-food-production
10. Poultry (including game birds): registration rules and forms. GOV.UK. Accessed December 27, 2022. https://www.gov.uk/government/publications/poultry-including-game-birds-registration-rules-and-forms
11. SCOT EID. Livestock traceability. Published 2023. Accessed March 20, 2023. https://www.scoteid.com/
12. GOV.UK. Culture results for Geronimo the alpaca. GOV.UK. Published 2021. Accessed March 20, 2023. https://www.gov.uk/government/news/culture-results-for-geronimo-the-alpaca
13. Twomey F. Protecting animal health - the role of scanning surveillance - APHA Science Blog. Published January 17, 2020. Accessed March 20, 2023. https://aphascience.blog.gov.uk/2020/01/17/one-health-scanning-surveillance/
14. Amirgazin A, Shevtsov A, Karibayev T, et al. Highly pathogenic avian influenza virus of the A/H5N8 subtype, clade 2.3.4.4b, caused outbreaks in Kazakhstan in 2020. *PeerJ*. 2022;10:e13038. doi:10.7717/peerj.13038
15. Anholt RM, Berezowski J, Jamal I, Ribble C, Stephen C. Mining free-text medical records for companion animal enteric syndrome surveillance. *Prev Vet Med*. 2014;113(4):417-422. doi:10.1016/j.prevetmed.2014.01.017
16. Aiello AE, Renson A, Zivich PN. Social Media– and Internet-Based Disease Surveillance for Public Health. *Annu Rev Public Health*. 2020;41(1):101-118. doi:10.1146/annurev-publhealth-040119-094402
17. Chilakamarri S. Online Community Detection Using Twitter Data. Published online 2020:95.
18. Alsudias L, Rayson P. Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study. *JMIR Med Inform*. 2021;9(9):e27670. doi:10.2196/27670
19. Braker C, Shiaeles S, Bendiab G, Savage N, Limniotis K. BotSpot: Deep Learning Classification of Bot Accounts Within Twitter. In: Galinina O, Andreev S, Balandin S, Koucheryavy Y, eds. *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Vol 12525. Lecture Notes in Computer Science. Springer International Publishing; 2020:165-175. doi:10.1007/978-3-030-65726-0_16

20. Alsudias L, Rayson P. Classifying Information Sources in Arabic Twitter to Support Online Monitoring of Infectious Diseases. :9.
21. Alhayan F, Pennington D, Ayouni S. Twitter use by the dementia community during COVID-19: a user classification and social network analysis. *Online Inf Rev*. Published online April 19, 2022. doi:10.1108/OIR-04-2021-0208
22. Dórea FC, McEwen BJ, McNab WB, Sanchez J, Revie CW. Syndromic Surveillance Using Veterinary Laboratory Data: Algorithm Combination and Customization of Alerts. *PLOS ONE*. 2013;8(12):e82183. doi:10.1371/journal.pone.0082183
23. Tweepy Documentation — tweepy 4.12.1 documentation. Accessed January 4, 2023. https://docs.tweepy.org/en/stable/
24. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica*. 2012;22(3):276-282.
25. Leung K. Micro, Macro & Weighted Averages of F1 Score, Clearly Explained. Medium. Published September 13, 2022. Accessed January 4, 2023. https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f
26. Classification: ROC Curve and AUC | Machine Learning. Google Developers. Accessed January 4, 2023. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Figures

Figure 1

Wordcloud for aggregated smallholding profile descriptions

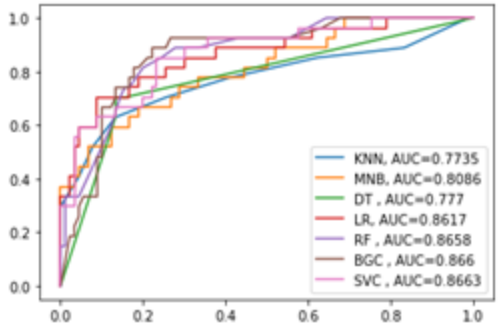

Figure 2

AUC curve imbalanced dataset

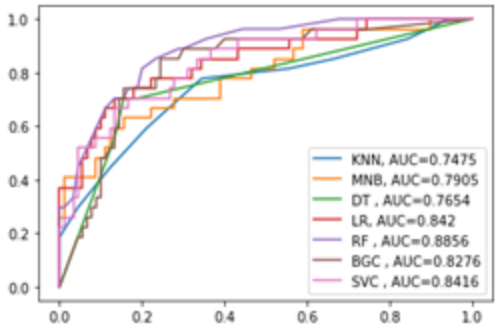

Figure 3

AUC curve undersampling

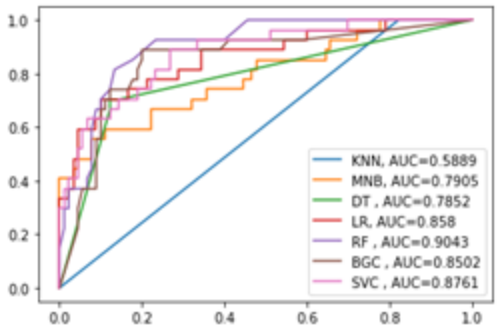

Figure 4

AUC curve oversampling

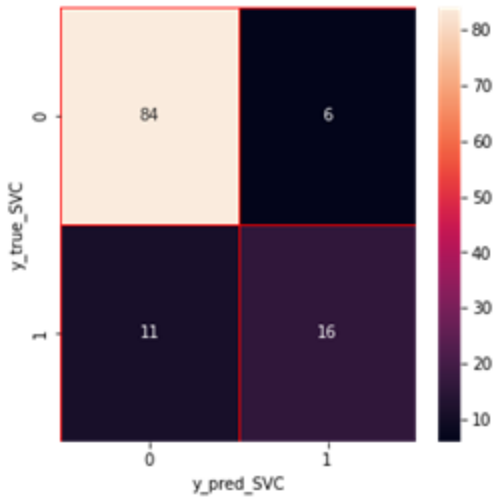

Figure 5

Confusion matrix - SVC