**Predictive Modeling of Undiagnosed Patients of Depression in Korean Elders**

**1. Introduction**

As the Korean population ages, the prevalence of depression in Korean elderly is increasing. According to the 'Survey on Elderly Conditions' published by the Korea Institute for Health and Social Affairs in 2018, 21.1% of 65-year-olds were found to have depressive symptoms, and 6.7% answered that they had considered suicide. Also, 13.2% of the elderly who had contemplated suicide had also attempted suicide. Even though depression in the most common psychiatric disorder in the elderly, it is often underreported, underdiagnosed, and thus undertreated. [12] [13] Prior research highlight that subjective self-rated depression surveys such as the Patient Health Questionnaire-9 (PHQ-9) are not robust measures of the severity of depression in patients. [14] Furthermore, frequently, elders underreport depressive symptoms, possibly due to shame or a lack of understanding of depression, further reducing the validity of these questionnaire as a true indicator of depression. [12]

Instead, we aim to develop a supervised classification model based on objective predictor variables to classify elders into depressed or not. One of the dominant explanation for depression in elderly is morbidity. In addition, loneliness, loss of a valued role and independence, being female, bereavement, alcohol consumption, and smoking are some risk factors for depression in Elders. [15] [16] As the KNAHES 2019 dataset includes a plethora of health and socioeconomic variables, we developed logistic, decision tree, random forest classifiers using these objective predictor variable to classify elders to depressed or not. [15] Such models can be incorporated into the healthcare system to support preliminary identification of patients with a risk of depression. Prior intervention studies demonstrated that education, consultation, increased exposure to communal activity positively impacted depressed elders, whilst case management programs reduced suicidal thoughts. [16] As such, through early identification and treatment of depressed elderly patients based on our supervised learning models, we hope to prevent elderly suicide.

**2. Method**

Firstly, we reviewed prior clinical literature to find which factors that caused based related to depression in elderly and identified relevant variables in the KNHANES 2019 dataset. Afterwards, we performed EDA with a correlation heat map to see the correlated variables. After choosing BP5 as our response variable, we performed stepwise AIC selection to select features that minimize the AIC model. Additionally, for the logistic classifier model, we iteratively checked pairwise second order interaction combinations to see if any interaction terms lowered the cross-validation error.

Based on the final features list, we split our dataset to a training, test, and validation with an 80:10:10 split and trained Logistic Classifier, Decision tree, and Random forest models. On the other hand, KNN, LDA, and QDA were not used as our predictor variables were mainly categorical variables; LDA and QDA require numerical predictor variables, which follow a multivariate Gaussian distribution, whilst KNN uses the numerical variables to compute Euclidean/Manhattan/Minkowski distance to calculate similarity between neighbors. Navies Bayes also has an assumption of independence in predictors. As diseases are often dependent with each other due to genetic pleiotropy and epistasis, the assumptions are not met, and we expected low model performance, thus excluding it. [17]

The trained models were applied to the test dataset and used to plot the Receiver Operating Curve(ROC) and Precision-Recall (PR) curve. Based on these curves, the Area under the Curve (AUC) was used to select the best-performing model. We computed the Balanced Accuracy and F-measure for each threshold in the ROC and PR curve to find the optimal threshold cutoff that maximized the recall of our model using the validation set. Using this threshold, we used it to find the accuracy, specificity, recall, precision, f-measure, balanced accuracy of the model on our test dataset.

**2.1 Variables Used for AIC Selection**

* **BP5:** 2주이상 연속 우울감여부
* **Sex:** 성별: 1: 남자 2: 여자
* **Cfam:** 가구원수: 1: 1명, 2: 2명, 3: 3명, 4: 4명, 5: 5명, 6: 6명 이상, 9: 모름, 무응답
* **Age:** (한국 법과 여러 노인 관련 논문 기준에 따라 만 65새 이상을 노인으로 칭한다.)
* **marri\_2:** 결혼상태 0:유배우자, 동거 1:유배우자, 별거/사별/이혼
* **BE3\_91:** 신체활동 여부:장소이동 1:예 2:아니오
* **BS1\_1:** (성인)평생 일반담여 흡연여부 1: 5갑미만 2:5갑 이상 3:피운 적없음
* **BD1\_11:** (만 12세 이상)1년간 음주빈도 1: 최근 1년간 전혀 마시지 않았다. 2: 월 1회미만 3: 월 1회 정도 4: 월 2~4회 5:주 2~3회 정도 6: 주4회 이상
* **mh\_stress:** 스트레스 인지율 0: 스트레스 적게 느낌 1: 스트레스 많이 느낌
* **X\_dg: ‘**X’는 질병 이름, ‘\_dg’는 의사진단 여부
  + **DI1\_dg:** 고혈압, **DI2\_dg:** 이상지질혈증, **DI3\_dg:** 뇌졸증 , **DI4\_dg:** 심근경색 또는 협심증, **DI5\_dg:** 심근경색증, **DI6\_dg:** 협심증, **DM1\_dg:** 관절염, **DM2\_dg:** 골관절염, **DM3**\_dg: 류마티스성, **DM4\_dg:** 골다공증 , **DJ2\_dg:** 폐결핵 , **DJ4\_dg:** 천식 , **DE2\_dg:** 갑상선, **DE1\_dg:** 당뇨, **DC1\_dg:** 위암, **DC2\_dg:** 간암, **DC3\_dg:** 대장암, **DC4\_dg:** 유방암, **DC5\_dg:** 자궁경부암, **DC6\_dg:** 폐암, **DC7\_dg:**갑상선암, **DC11\_dg:** 기타암, **DL1\_dg:**아토피피부염, **DJ8\_dg:** 알레르기비염, **DJ6\_dg:** 부비동염, **DH4\_dg:** 중이염, **DN1\_dg:** 콩팥병, **DK8\_dg:** B 형간염, **DK9\_dg:** C 형간염, **DK4\_dg:** 간경변증, **DM8\_dg:** 통풍, **BP17\_dg:** 폐쇄성수면무호흡증 …..의 의사진단 여부

**3. Results and Discussion**

**3.1 Exploratory Bivariate Data Analysis**

Figure 1 shows that BP5 is correlated with DM4\_dg, DJ4\_dg, DF2\_dg, DJ8\_dg, DJ6\_dg, DK4\_dg. Similarly, DF2\_dg, a binary variable of whether someone has been diagnosed with depression, is correlated with marri\_2, DM1\_Dg, DM4\_dg, Dj4\_dg, DC2\_dg, DC4\_dg, DL1\_dg, DK4\_dg. The correlation coefficient between DF2\_dg and BP5 was 0.197(3sf). DF2\_pr is also a variable which shows whether someone is currently diagnosed with depression. We used BP5 instead of DF2\_dg and DF2\_pr in our final model because the processed dataset using DF2\_dg and DF2\_pr had too few data points. The validity of this choice will be referred back to in Section 4. To note, DF2\_dg and DF2\_pr was not used as a predictor variable.

Figure 1 also shows that similar diseases like the arthritis variables (DM1, DM2, DM3, DM4) are correlated with one another. In addition, DN1\_dg (Kidney Disease) and DM8\_dg (Gout), which are diseases that are closely related, show a correlation of 0.159.

**3.2 Feature Selection, Final Dataset, and Model Training**

After using AIC stepwise selection, DM1\_dg, DM2\_dg, DM3\_dg, DM4\_dg, DJ4\_dg , DJ6\_dg , DN1\_dg , DM8\_dg , marri\_2 , mh\_stress were shown to minimize the AIC value. Based on these variables, the final dataset was created as shown in Table 1. Afterwards, using LASSO regression, multicollinear variables were removed. Using a penalty of 0.05 lowered the CV the most from a CV error of 0.121 to 0.115. The remaining variables were DM3\_dg + DN1\_dg + DM8\_dg + marri\_2 + mh\_stress. Next, all combinations of bivariate interaction terms were checked to see which produced the lowest 10-fold-cross-validation error. However, all interactions terms were statistically insignificant in terms of p-value calculated by ANOVA and didn’t lower the CV error.

Using the features selected by stepwise AIC, we trained our logistic, decision tree, and random forest classifier. Figure 5 shows the importance of each feature. From the graph we can see that mh\_stress and DJ6\_dg were important variables, whereas, DM8\_dg and DN1\_dg were not influential. Similarly in Figure 4, the nodes were split by mh\_stress and DJ6\_dg. i.e. those who are stressed and have sinusitis were more likely to have depressive thoughts in the past two weeks. However, as we observe in Figure 4, the nodes were not homogenously seperated.

In the HN19 dataset, there were 5965 people who were not depressed and 714 people who felt depressed in the recent two weeks (0.893:0.107 ratio). In our processed final dataset as shown in Table 1, there were 383 people who were not depressed and 79 people who felt depressed in the recent two weeks (0.829: 0.171 ratio). In addition, although logistic classifiers are robust against imbalanced data, decision trees and random forest classifiers are more prone to bias towards the majority class. [11] Thus, to account for this class imbalance, we will use the F1.5-scores and Balanced Accuracy metrics to find the cutoff threshold. [11] [18]

**3.3 ROC Curve**

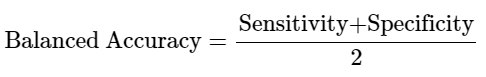
The ROCs of each model was plotted as shown in Figure 2. The AUC of ROCs for the Decision Tree, Logistic Classifier, and Random Forest models were 0.674, 0.748, 0.689, respectively. All three models were similar in ROC AUC performance and the ROC curves show significant overlap, making it difficult to differentiate the models. The similarity between the ROC curves are due to the class imbalance; for balanced data, the difference in the ROC curves would be more prominent. [18] According to the ROC curves, the logistic classifier performed the best, then the random forest, and then the decision tree. Since ROC does not take class imbalanced into account and may present an overly optimistic view on an algorithm’s performance for skewed data, the PR curve of the models were plotted too. [11] [18]

**3.4 PR Curve**

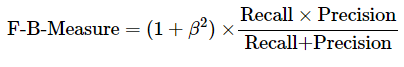
The PR curve visualizes the trade-off between precision and recall for different probability thresholds and is suited for detection of positive case rare events in imbalanced datasets. This is because both precisions and recall are both independent of the true negative rate, it is not dwarfed by the large true negative rate that is a byproduct of an imbalanced dataset. On the other hand, specificity, which is used in the ROC curve, has the true negative in the denominator, which gets dwarfed due to the high number of true negatives. Figure 3 shows the PR curves of the three models. Since 16.2% of people in the test set were depressed, there is a random baseline at 0.162, highlighting the class imbalance. The decision tree, random forest, and logistic classifier models have a PR-AUC of 0.426, 0.382, 0.484. As observed in Figure 3, there was less overlap between the PR curves and higher variation in the PR AUC. The logistic classifier was ultimately selected as the best model.

**3.5 Tuning Cutoff Threshold**

The overall classification accuracy is not fully indicative of performance ability for imbalanced data. To reiterate, a trivial or random classifier can still achieve a high accuracy if it predicts all cases to be part of the majority class. [11] Instead, we used the balanced/weighted accuracy to tune the cutoff threshold for the ROC Curve with the validation set. If the test set is used to tune the validation set, it no longer acts as an unbiased source to judge model performance. [11]



To tune the cutoff threshold for the PR curve, we used the F-score with a value of 1.5. [11] We set the Beta value to 1.5, meaning that we put more weight on recall than precision, thus preventing the model from producing false negatives. However, even when the Beta value was set to 1, the threshold probability values were the same for all models. [11]



**3.6 Metrics for Final Decision Tree Model**

As shown in table 4, compared to the default threshold, the decision tree classifier with a cutoff at 0.330 had a higher recall, but lower specificity and precision. A recall of 0.458 of the final model means that 45.8% of actually depressed people, were identified as depressed. Whereas, a precision value of 0.440 means that out of those who were predicted to be depressed, 49% of these predictions were correct. Specificity or the true negative rate is less of a concern as classifying patients as not depressed is not important relative to correctly classifying depressed patients.

**4. Discussion**

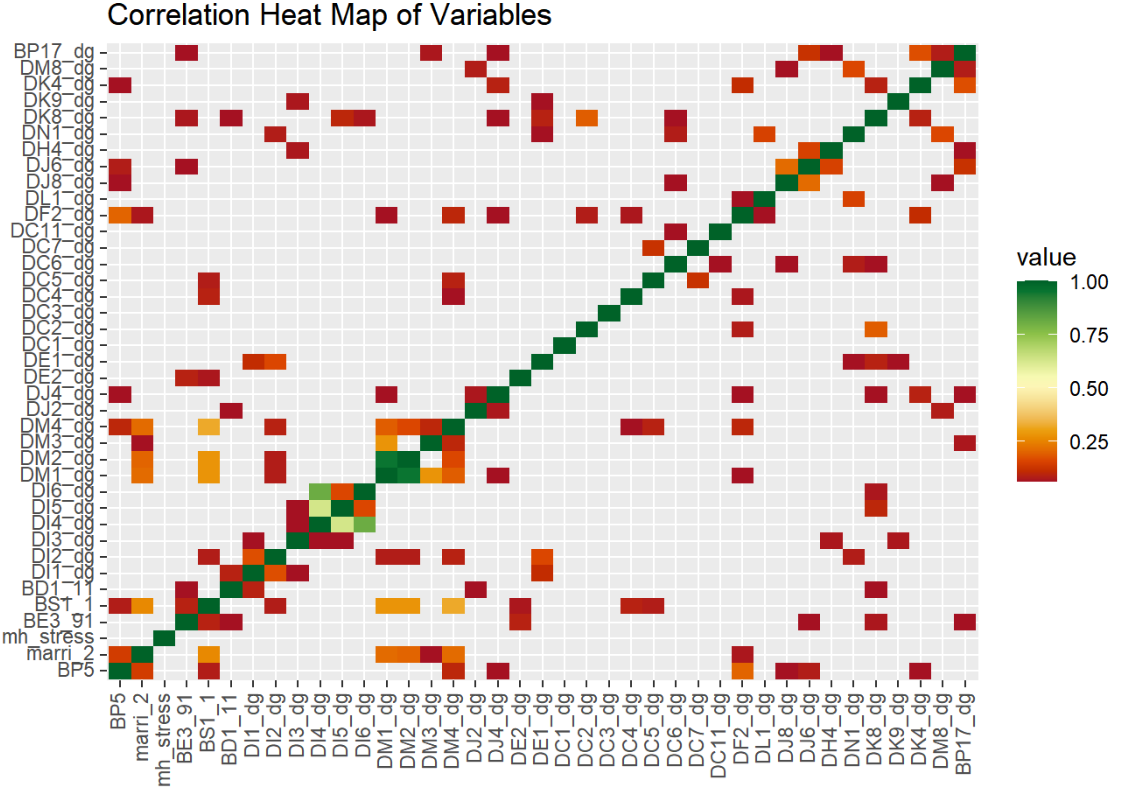
Table 5 shows a frequency table of DF2\_pr against BP5. In this table, 107 individuals had depressive thoughts for two weeks consequently and 81.3% of these individuals currently had depression. Possibly, the remaining 18.7% may be undiagnosed for depression or by chance had a difficult past two weeks on the day of the questionnaire. However, out of the 160 individuals currently diagnosed with depression, 54.4% of them had depressive thoughts for two consecutive weeks. Therefore, a limitation of our model is that it fails to identify a large proportion of currently depressed people as the people having depressive thoughts in the past two weeks and those who are clinically depressed are not necessarily equivalent. We can partially mitigate this problem by increasing the recall by trading off precision as we have done in Section 3.5 as this would increase the number of people classified to be depressed. However, there comes additional complications when misclassifying someone to be depressed when they are not may not be ideal. A non-depressed individual may believe that they are depressed, which leads to the development of depression, according to a phenomenon called the Self Fulfilling Prophecy Theory in the field of Psychology. [19] For example, prior research highlights that misdiagnosing patients with depression as the false positive group showed significantly higher functional impairment. [20] Ultimately, the proposed models should be used only as a preliminary screening tool to identify people who has a risk for depression for depressive symptoms in elders.

In addition, there are also other methods of dealing with class imbalance, such as SMOTE/over/under sampling or assigning a high weight to the minority class to heavily penalize misclassification of the minority class. [11] [18] By tuning the threshold, without changing the model parameters, we did not increase the overall effectiveness of the model. [18] Instead, alternating thresholds just adjust the trade-off between precision and recall to maximize our balanced frequency and f-1.5-measure. [18] To summarize, changing the threshold does not lead to any further separation between classes. [18] Unlike using alternative cutoffs, assigning unequal costs based on the different weight of errors have the potential to make true improvements to the classifier. [18] However, these methods are beyond the scope of this paper, and further research should be conducted implementing these methods to improve the predictive power of our algorithms. Another pitfall of our analysis is that to use PR curves, it is recommended that the test data has a similar class distribution to that of the true population. Our test dataset has a positive class prevalence of 16.32%, however, it is unclear whether this is the true depression prevalence rate in Korean elders. Although not the true population BP5 positive prevalence rate, 10.7% of people had depressive thoughts in the KNHANES 2019 dataset and 10.7% of people had depressive thoughts in the KNHANES 2018 dataset.

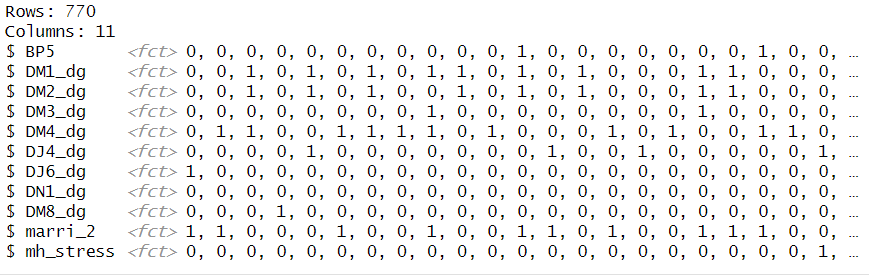
**Conclusion**

As the prevalence of depression increases in the Korean elder population, early intervention is crucial to prevent suicide. Due to underreporting of depression in the elder population and lack of robust objective surveys to identify depressed elders, we proposed three supervised learning classification models to identify depressed elders based on health and socioeconomic factors. Out of these three models, the Logistic Classifier model performed the best with a cutoff at 0.330. Further tuning of the model parameters to account for the imbalanced of depression in the population, and improve predictive power, should be performed.

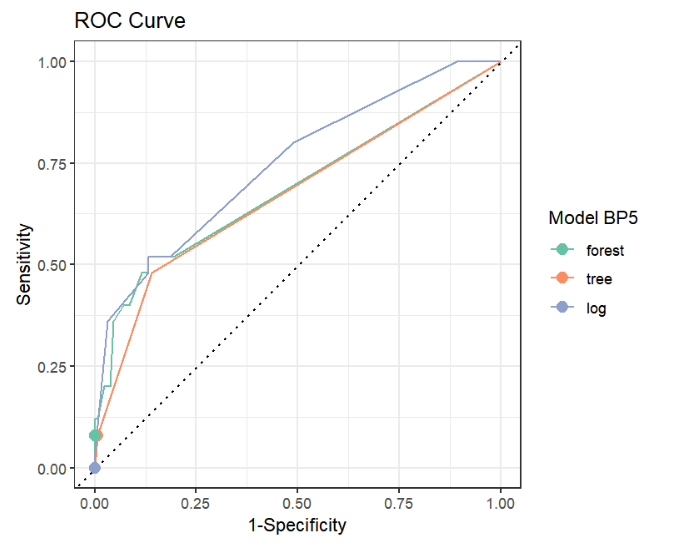
**Appendix**



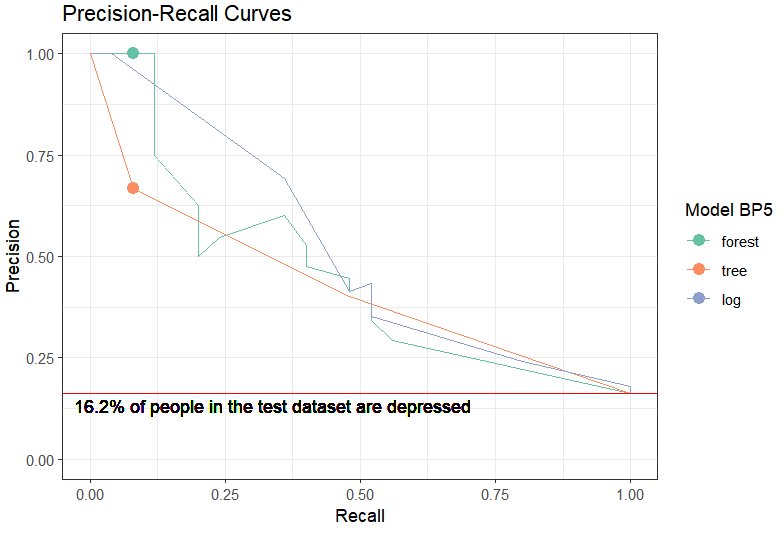
**Figure 1 Correlation Heat Map** Only variables with pearson correlation coefficients of at least 0.05 are colored.



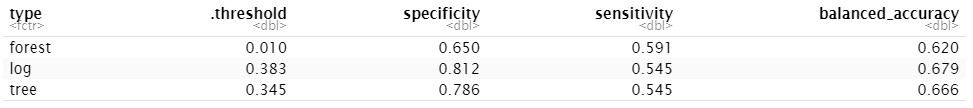
**Table 1 Processed Dataset** Our final dataset included 770 individuals. All variables were binary categorical variables.



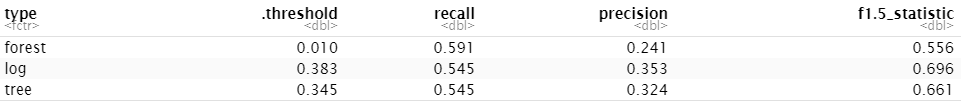
**Figure 2 ROC Curve**



**Figure 3 Precision-Recall Curve**

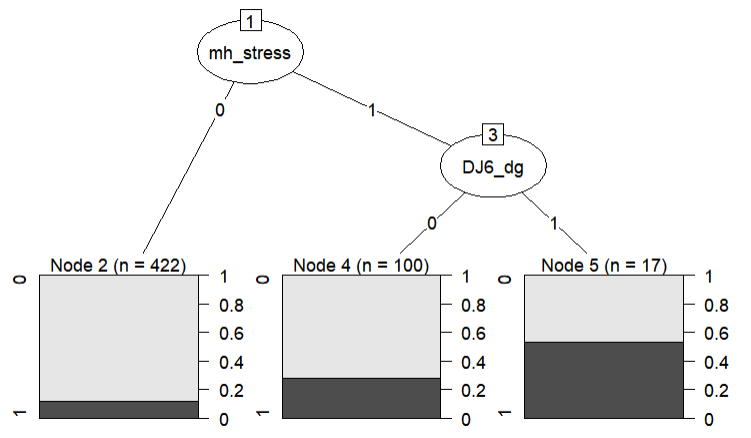


**Table 2 Cutoff Values that Maximize Balanced Accuracy**

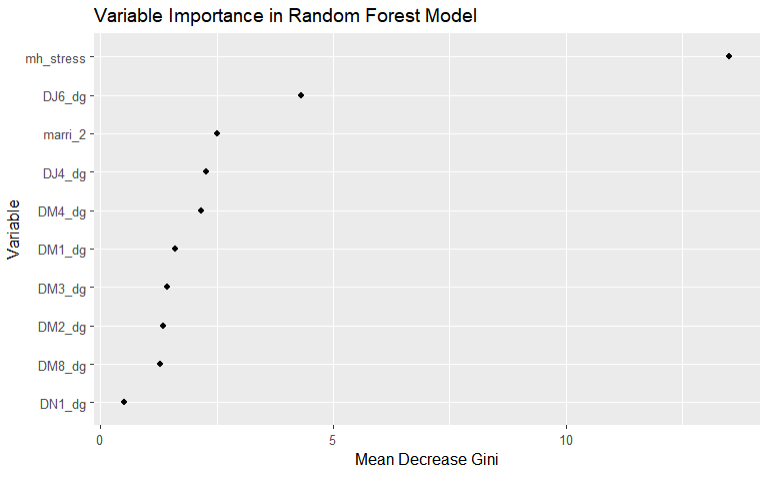


**Table 3 Cutoff Threshold that Maximize F1.5-Measure**

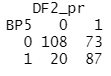
|  |
| --- |
|  |
| **Table 4 Metric Comparison of Metrics for Logistic Classifier at Probability Cutoff of 0.330 and 0.5.** |



**Figure 4 Decision Tree Visualizations**



**Figure 5 Feature Importance Based on Random Forest Model**



**Table 5 Frequency Table Comparing those that are currently depressed to those who have had depressed thoughts in the past two weeks.**

**Reference**

[1] 김봉정. (2020). 노인의 우울증에 영향을 미치는 요인: 국민건강영양조사 제7기 자료. 보건정보통계학회지, 45(2), 165-172.

[2] 최희아, 이동숙 and 김영주. (2011). 한국 성인여성들의 우울증과 자살생각에 대한 분석 -제4기 국민건강영양조사자료를 이용하여-. Journal of The Korean Data Analysis Society, 13(2), 709-720.

[4] 저널: Journal of Health Informatics and Statistics

[5] Seo, J., Choi, B., Kim, S. et al. The relationship between multiple chronic diseases and depressive symptoms among middle-aged and elderly populations: results of a 2009 korean community health survey of 156,747 participants. BMC Public Health 17, 844 (2017). https://doi.org/10.1186/s12889-017-4798-2

[6] Ma, Y., Xiang, Q., Yan, C., Liao, H., & Wang, J. (2021). Relationship between chronic diseases and depression: the mediating effect of pain. BMC psychiatry, 21(1), 436. https://doi.org/10.1186/s12888-021-03428-3

[7] Davis, Jesse & Goadrich, Mark. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning, ACM. 06. 10.1145/1143844.1143874.

<https://arxiv.org/pdf/2204.02526.pdf>

[8]

[9]

[10]

[11] Chen, Chao & Breiman, Leo. (2004). Using Random Forest to Learn Imbalanced Data. University of California, Berkeley.

[12] Nagar S, Patel HB, Nagar N, Mahyavanshi D, Nagar SS, Godara N. Study on Factors Associated with Depression among Elderly and Comparison of Two Scales Used for Screening. *Indian J Community Med*. 2021;46(3):446-449. doi:10.4103/ijcm.IJCM\_627\_20

[13] O'Connor DW, Rosewarne R, Bruce A: Depression in primary care. 2: general practitioners' recognition of major depression in elderly patients. Int Psychogeriatr. 2001, 13: 367-374. 10.1017/S104161020100775X.

Phelan, E., Williams, B., Meeker, K. *et al.* A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam Pract* **11**, 63 (2010). https://doi.org/10.1186/1471-2296-11-63

[14] Phelan, E., Williams, B., Meeker, K. *et al.* A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam Pract* **11**, 63 (2010). https://doi.org/10.1186/1471-2296-11-63

[15] Allan CE, Valkanova V, Ebmeier KP. Depression in older people is underdiagnosed. *Practitioner*. 2014;258(1771):19-3.

[16] Awata S. *Seishin Shinkeigaku Zasshi*. 2005;107(10):1099-1109.

[17] Gratten, J., Visscher, P.M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med* **8**, 78 (2016). <https://doi.org/10.1186/s13073-016-0332-x>

[18] Kuhn, M., & Johnson, K: Chapter 16 *Applied predictive modeling*. New York, Springer. 2013

[19] Weaver J, Filson Moses J, Snyder M. Self-Fulfilling Prophecies in Ability Settings. *J Soc Psychol*. 2016;156(2):179-189. doi:10.1080/00224545.2015.1076761

[20] Leon AC, Portera L, Olfson M, et al. False positive results: a challenge for psychiatric screening in primary care. *Am J Psychiatry*. 1997;154(10):1462-1464. doi:10.1176/ajp.154.10.1462

[21] Pikhartova J, Bowling A, Victor C. Is loneliness in later life a self-fulfilling prophecy?. *Aging Ment Health*. 2016;20(5):543-549. doi:10.1080/13607863.2015.1023767