References

[1] W. H. Organization, "Cardiovascular diseases (cvds)," https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2017, [Online: accessed 31-8-2021].

[2] N. Poulter, "Coronary heart disease is a multifactorial disease," *American Journal of Hypertension*, vol. 12, no. S6, pp. 92S–95S, 10 1999. doi: 10.1016/S0895-7061(99)00163-6. [Online]. Available: https://doi.org/10.1016/S0895-7061(99)00163-6

[3] J. Navarro-Pérez, D. Orozco-Beltran, V. Gil-Guillen, F. Pallares, V.and Valls, A. Fernandez, A. M. Perez-Navarro, C. Sanchis, A. Dominguez-Lucas, J. M. Martin-Moreno, J. Redon, and M. Tellez-Plaza, "Mortality and cardiovascular disease burden of uncontrolled diabetes in a registry-based cohort: the escarval-risk study," in *BMC cardiovascular disorders*, vol. 18(1), no. 180, 2018. doi: 10.1186/s12872-018-0914-1

[4] PubMed, https://pubmed.ncbi.nlm.nih.gov/, 2021, [Online: accessed 31-8-2021].

[5] C. Krittanawong, H. U. Virk, S. Bangalore, Z. Wang, W. J. Kipp, R. Pinotti, H. Zhang, S. Kaplin, B. Narasimhan, T. Kitai, U. Baber, J. L. Halperin, and W. H. W. Tang, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific Reports*, 2020. doi: 10.1038/s41598-020-72685-1. [Online]. Available: https://doi.org/10.1038/s41598-020-72685-1

[6] E. Soria-Olivas, J. D. Martín-Guerrero, J. Redón, M. Tellez-Plaza, and J. Vila-Francés, "Improving mortality prediction in cardiovascular risk patients by balancing classes," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015. doi: 10.1109/ICDMW.2015.76 pp. 480–484.

[7] F. Mateo, E. Soria-Olivas, M. Martínez-Sober, M. Tellez-Plaza, J. Gómez-Sanchís, and J. Redon, "Multi-step strategy for mortality assessment in cardiovascular risk patients with imbalanced data," in *24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016*, 2016. [Online]. Available: http://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2016-60.pdf

[8] R. C. Deo, "Machine learning in medicine," *Circulation, PMC*, vol. 20, pp. 1920–1930, 2018.

[9] A. Subasi, "Practical machine learning for data analysis using python," in *Practical Machine Learning for Data Analysis Using Python*. Academic Press, 2020. ISBN 978-0-12-821379-7 pp. 8–13.

[10] R. Soni and K. Mathai, "Improved twitter sentiment prediction through cluster-then-predict model," *ArXiv*, vol. abs/1509.02437, 2015.

[11] S. Kaufman, S. Rosset, and C. Perlich, "Leakage in data mining: Formulation, detection, and avoidance," vol. 6, 01 2011. doi: 10.1145/2020408.2020496 pp. 556–563.

[12] PyCaret, https://pycaret.org/, 2021, [Online: accessed 31-8-2021].

[13] t. PyCaret, "Pycaret, transformations," https://pycaret.org/transformation/, 2021, [Online: accessed 31-8-2021].

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," in *Artificial Intelligence (cs.AI)*, vol. 16, 2002. doi: 10.1613/jair.953 pp. 321–357.

[15] J. Brownlee, "Smote for imbalanced classification with python," https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/, 2021, [Online: accessed 31-8-2021].

[16] Numpy, https://numpy.org/, 2021, [Online: accessed 31-8-2021].

[17] Pandas, "Pandas," https://pandas.pydata.org/, 2017, [Online: accessed 31-8-2021].

[18] SciPy, https://www.scipy.org/, 2021, [Online: accessed 31-8-2021].

[19] V. w. P. Matplotlib, https://matplotlib.org/, 2021, [Online: accessed 31-8-2021].

[20] s. d. v. Seaborn, https://seaborn.pydata.org/, 2021, [Online: accessed 31-8-2021].

[21] M. L. in Python, "Machine learning in python," https://scikit-learn.org/stable/, 2021, [Online: accessed 31-8-2021].

[22] S. Weisberg, *Applied Linear Regression*. John Wiley & Sons, Inc., 2005. ISBN 9780471704096

[23] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques; Adaptive computation and machine learning*. MIT Press, 2009. ISBN 9780262013192

[24] M. Sánchez-Montañés, P. Rodríguez-Belenguer, A. J. Serrano-López, E. Soria-Olivas, and Y. Alakhdar-Mohmara, "Machine learning for mortality analysis in patients with covid-19," *Internatinal Journal of Environmental Research and Public Health*, vol. 17, 2020.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction To Statistical Learning*. Springer, 2013. ISBN 9781461471387

[26] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[27] T. Afonja, "Accuracy paradox," https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b, 2017, [Online: accessed 31-8-2021].

[28] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," vol. 41, 2014. doi: 10.1007/s10115-013-0679-x p. 647–665.

[29] M. L. McHugh, "The chi-square test of independence," *Biochemia Medica*, vol. 23, pp. 143–149, 2013.

[30] A. Schneider, G. Hommel, and M. Blettner, "Linear regression analysis," *Evaluation of Scientific Publications*, vol. 107, pp. 776—-782, 2010.

Appendix A

Results overview

Table A.1 – XGBoost Results through different experiments in test data.

Activity	**Model**	**F1**	**Recall**	**Prec.**	**Accuracy**	**AUC**
Baseline	XGBoost	0.2328	0.2639	0.2082	0.9695	0.8408
Undersampling	XGBoost	0.0846	0.8090	0.0446	0.6926	0.8308
Oversampling	XGBoost	0.1079	0.0833	0.1529	0.9758	0.7672
SMOTE	XGBoost	0.0889	0.0764	0.1063	0.9725	0.7821

Table A.2 – SVM Results through different experiments in test data.

Activity	**Model**	**F1**	**Recall**	**Prec.**	**Accuracy**	**AUC**
Baseline	SVM	0.0975	0.7083	0.0524	0.7698	0.7396
Undersampling	SVM	0.1213	0.6840	0.0665	0.8260	0.7563
Oversampling	SVM	0.1024	0.7604	0.0549	0.7659	0.7632
SMOTE	SVM	0.1024	0.7604	0.0549	0.7659	0.7632