International Journal of Advanced Technology in Engineering
Volume. No. 14, Issue No. 05, May 2026
(ISSN: 2348-7550)

A Scalable Hybrid Machine Learning Framework for Diabetes Progression Prediction Using LightGBM–KNN in Cloud Environments

Dr.M.Aparna[1], Manchala Niharika[2], Vankayala Pravalika[3], Mylavarapu Venkata Lakshmi Himaja[4], Shaik Sadiya Sulthana[5]

[1]Head Of The Department, Department Of CSE(AIML), Tirumala Engineering College,AP
[2,3,4,5]B.Tech, Department Of CSE(AIML), Tirumala Engineering College,AP

Abstract:
Diabetes progression prediction and risk stratification are critical for enabling early clinical intervention and reducing long-term complications associated with chronic metabolic disorders. This paper proposes a cloud-enabled hybrid ensemble learning framework for intelligent risk stratification in diabetes progression prediction using structured clinical and biochemical health parameters. The proposed system integrates Light Gradient Boosting Machine (LightGBM) and K-Nearest Neighbour (KNN) classifiers through a soft voting ensemble mechanism to improve prediction robustness and classification reliability. Clinical attributes such as age, body mass index, blood pressure, glucose level, HbA1c, cholesterol indicators, and other relevant health parameters are utilized as input features for model training and prediction. To enhance data quality and model effectiveness, preprocessing techniques including missing value imputation, feature scaling, normalization, and categorical encoding are applied. Hyperparameter optimization and cross-validation are employed to improve generalization and reduce overfitting. Experimental evaluation demonstrates that the proposed hybrid ensemble framework outperforms individual machine learning models in diabetes risk prediction, achieving superior classification accuracy and enhanced stability across multiple evaluation metrics including precision, recall, F1-score, and ROC-AUC. The developed framework categorizes patients into multiple diabetes progression risk levels, thereby supporting personalized treatment planning and proactive healthcare decision-making. Furthermore, cloud-based deployment capabilities enable scalable and real-time accessibility for healthcare professionals, making the proposed system suitable for intelligent clinical decision support applications.

Index terms - *— Diabetes Progression Prediction, Risk Stratification, Hybrid Ensemble Learning, LightGBM, K-Nearest Neighbour, Cloud Computing, Machine Learning in Healthcare, Clinical Decision Support System, Predictive Healthcare Analytics*

1. INTRODUCTION

Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from inadequate insulin production, impaired insulin utilization, or both. It has become one of the most prevalent global health challenges due to its rapid increase in incidence and its association with severe long-term complications such as cardiovascular disease, kidney failure, neuropathy, and vision impairment. Early prediction of diabetes progression and accurate identification of high-risk individuals are essential for enabling timely intervention, improving treatment outcomes, and reducing healthcare burdens. Conventional diagnostic methods primarily rely on laboratory investigations and physician assessment, which are effective for diagnosis but limited in predicting future disease progression and risk severity. As a result, there is a growing need for intelligent predictive systems capable of analyzing patient health data to provide proactive clinical decision support.

Recent advancements in machine learning have enabled the development of predictive healthcare models that can identify hidden patterns within clinical datasets and support early disease prediction with improved efficiency and accuracy. In particular, ensemble learning approaches have demonstrated superior performance over traditional single-model classifiers by combining multiple algorithms to improve robustness and reduce prediction variance. Motivated by these advancements, this paper proposes a cloud-enabled hybrid ensemble learning framework that integrates Light Gradient Boosting Machine (LightGBM) and K-Nearest Neighbour (KNN) classifiers for diabetes progression risk stratification. The proposed framework processes structured clinical and biochemical attributes to categorize patients into multiple diabetes risk levels, thereby supporting personalized treatment planning and preventive healthcare strategies. Additionally, cloud-based deployment enables scalable and real-time accessibility, making the system suitable for modern intelligent healthcare environments.

2. LITERATURE SURVEY

a) Diabetes prediction model using machine learning techniques

Diabetes is a global health condition that causes kidney failure, eye loss, and heart disease. Machine learning algorithms can improve illness detection and treatment, relieving healthcare staff. Diabetes forecasting has advanced fast, enabling early intervention and patient empowerment. Our paper introduces a novel diabetes prediction model using machine learning techniques such as Logistic Regression, SVM, Naïve Bayes, and Random Forest. We use ensemble learning to improve prediction accuracy and resilience in addition to these core methods. We study ensemble approaches including XGBoost, LightGBM, CatBoost, Adaboost, and Bagging. These methods combine basic learner predictions for a more accurate and robust prediction. Python is used to train our system on a Kaggle dataset. Our approach is carefully evaluated using confusion matrix, sensitivity, and accuracy tests. The most accurate ensemble approach evaluated was CatBoost, with 95.4% accuracy compared to XGBoost's 94.3%. CatBoost's higher AUC-ROC score of 0.99 suggests it may be better than XGBoost's 0.98.

b) Machine Learning and Data Mining Methods in Diabetes Research

Advanced biotechnology and health sciences have created a lot of data, including high-throughput genetic data and clinical data from huge Electronic Health Records. Machine learning and data mining in biosciences are more important than ever in transforming all accessible information into usable knowledge. Diabetes mellitus (DM) is a spectrum of metabolic illnesses that affect global health. Due to extensive diabetes research in diagnosis, etiopathophysiology, treatment, etc., massive volumes of data have been generated. This study reviews machine learning, data mining, and tools in diabetes research for prediction and diagnosis, diabetic complications, genetic background and environment, and health care and management, with the first category being the most popular. Many machine learning algorithms were used. In general, 85% of those employed were supervised learning and 15% unsupervised, especially association rules. The most used algorithm is support vector machines (SVM). Clinical data predominated. The title applications in the selected publications demonstrate the value of extracting knowledge to generate new hypotheses for DM research.

c) Prediction of Diabetes using Classification Algorithms

One of the most deadly chronic illnesses that raises blood sugar levels is diabetes. If diabetes is not diagnosed and managed, several problems might arise. A patient must attend a diagnostic facility and consult a physician as a result of the laborious identification procedure. However, this crucial issue is resolved by the development of machine learning techniques. The goal of this research is to create a model that can accurately predict a patient's risk of developing diabetes. In order to identify diabetes early on, this experiment uses three machine learning classification algorithms: Decision Tree, SVM, and Naive Bayes. The Pima Indians Diabetes Database (PIDD), which is obtained from the UCI machine learning repository, is used for the experiments. Precision, Accuracy, F- Measure, and Recall are just a few of the metrics used to assess each algorithm's performance. Both properly and wrongly categorized cases are used to gauge accuracy. According to the results, Naive Bayes performs better than other algorithms, with the greatest accuracy of 76.30%. Receiver Operating Characteristic (ROC) curves are used to properly and methodically validate these findings.

d) Predicting Diabetes Mellitus With Machine Learning Techniques

Hyperglycemia is a hallmark of diabetes mellitus, a chronic illness. Numerous issues might arise from it. According to the rising morbidity in recent years, 642 million people worldwide will have diabetes by 2040, meaning that one in ten individuals will have the disease. Without a doubt, this concerning number requires careful consideration. Machine learning has been used in many facets of medical health due to its quick progress. In this study, diabetes mellitus was predicted using decision trees, random forests, and neural networks. The dataset consists of Luzhou, China, hospital physical examination data. It has fourteen characteristics. The models in this study were examined using five-fold cross validation. We selected a few techniques with superior performance to carry out independent test trials in order to verify the approaches' general applicability. For the training set, we chose at random the data of 68994 healthy individuals and diabetes patients, respectively. We extracted the data five times at random due to the imbalance. The average of these five experiments is the outcome. To lower the dimensionality in this investigation, we employed principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR). The findings demonstrated that when all the variables were employed, random forest prediction could achieve the maximum accuracy (ACC = 0.8084).

e) XGBoost: A Scalable Tree Boosting System: One popular and very successful machine learning technique is tree boosting. In this research, we introduce XGBoost, a scalable end-to-end tree boosting system that is frequently utilized by data scientists to deliver state-of-the-art outcomes on many machine learning issues. We provide a weighted quantile sketch for approximation tree learning and a unique sparsity-aware method for sparse data. More

significantly, we offer information on data compression, sharding, and cache access patterns to construct a scalable tree boosting system. Combining these insights allows XGBoost to use far fewer resources than current systems while scaling beyond billions of samples.

3. METHODOLOGY

i) Proposed Work:

The proposed work introduces a cloud-enabled hybrid ensemble learning framework for intelligent diabetes progression risk stratification using structured clinical and biochemical health parameters. The system is designed to analyze patient medical attributes such as age, body mass index, blood pressure, glucose level, HbA1c, cholesterol indicators, and related clinical factors to predict the likelihood and severity of diabetes progression. To ensure data quality and improve model effectiveness, the collected healthcare dataset undergoes preprocessing operations including missing value imputation, categorical feature encoding, normalization, and feature scaling. Feature optimization techniques are further applied to identify the most relevant medical attributes contributing to diabetes risk prediction, thereby improving computational efficiency and predictive reliability.

The predictive framework employs a hybrid ensemble approach by integrating Light Gradient Boosting Machine (LightGBM) and K-Nearest Neighbour (KNN) classifiers through a soft voting mechanism. LightGBM captures complex nonlinear relationships among clinical parameters, while KNN enhances prediction through similarity-based classification of patient records. Hyperparameter tuning and cross-validation techniques are incorporated to optimize model performance and ensure generalization across unseen data. The final ensemble model stratifies patients into multiple diabetes progression risk categories, enabling healthcare professionals to identify high-risk individuals early and support personalized treatment planning. Deployment in a cloud-enabled environment further allows scalable, real-time access to prediction services for intelligent clinical decision support applications.

ii) System Architecture:

The proposed system architecture is designed as a structured cloud-enabled predictive pipeline for diabetes progression risk stratification. Initially, the framework accepts multiple clinical and biochemical input features including age, sex, body mass index (BMI), blood pressure (BP), low-density lipoprotein (LDL), high-density lipoprotein (HDL), total cholesterol (TC), triglycerides (TG), lamotrigine-related biochemical indicator (LTG), and glucose level (GLU). These input parameters are forwarded to the preprocessing module, where data cleaning, missing value imputation, normalization, feature scaling, and categorical encoding are performed to ensure consistency and improve data quality before model analysis. The preprocessing stage also removes redundant noise and standardizes feature distributions, thereby improving the reliability of downstream prediction models. The architecture shown in your provided diagram aligns with this pipeline.

Following preprocessing, the optimized hybrid ensemble engine processes the refined features using Light Gradient Boosting Machine (LGBM) and K-Nearest Neighbour (KNN) classifiers integrated through a soft voting ensemble strategy. LightGBM captures complex nonlinear relationships among clinical variables, while KNN enhances prediction by leveraging similarity-based neighborhood classification. The ensemble output is then mapped into disease progression categories, namely Class 0 (Low Disease Progression) and Class 1 (High Disease Progression), enabling robust risk stratification for diabetes progression assessment. This architecture supports accurate and scalable prediction while facilitating cloud deployment for real-time clinical decision support, remote accessibility, and integration into intelligent healthcare monitoring environments.

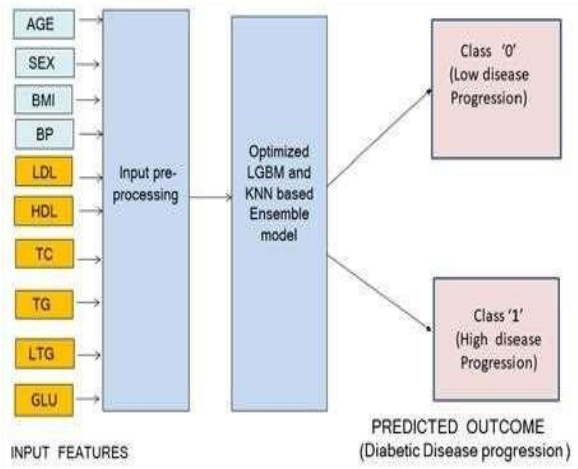

Fig 1: proposed architecture

iii) Modules:

1. Data Collection Module

Collects patient clinical and biochemical health parameters such as age, sex, BMI, blood pressure, cholesterol levels, triglycerides, LTG, and glucose values from structured medical datasets or healthcare records for diabetes progression analysis. It serves as the primary input acquisition layer of the predictive framework.

2. Data Preprocessing Module

Performs data cleaning and transformation operations including missing value handling, categorical encoding, normalization, and feature scaling. This module improves data quality and ensures compatibility of clinical features with machine

learning algorithms.

3. Feature Optimization Module

Identifies the most influential clinical attributes affecting diabetes progression using feature selection and correlation analysis techniques. It removes redundant or less significant features to improve prediction efficiency and reduce computational overhead.

4. LightGBM Prediction Module

Applies the Light Gradient Boosting Machine classifier to learn complex nonlinear relationships among clinical parameters and generate diabetes progression prediction probabilities with high computational efficiency.

5. KNN Prediction Module

Implements the K-Nearest Neighbour classifier to analyze similarity between patient health records and classify disease progression based on neighborhood-based learning patterns.

6. Ensemble Decision Module

Combines prediction outputs from LightGBM and KNN using a soft voting ensemble mechanism to improve prediction robustness, reduce classification errors, and enhance final decision reliability.

7. Risk Stratification Module

Maps ensemble prediction probabilities into disease progression categories such as Low Disease Progression and High Disease Progression, enabling multi-level clinical risk assessment.

8. Visualization and Evaluation Module

Generates performance metrics and visual outputs including accuracy, precision, recall, F1-score, ROC curves, confusion matrix, and feature importance graphs for model interpretation and validation.

9. Cloud Deployment Module

Deploys the trained prediction framework in a cloud/web environment to support real-time diabetes progression prediction and remote accessibility for healthcare professionals.

iv) Algorithms:

1. Data Preprocessing Algorithms

Preprocessing algorithms are applied to improve dataset quality before model training. Techniques such as missing value imputation, categorical encoding, normalization, and feature scaling are used to remove inconsistencies and standardize medical attributes, ensuring reliable input for machine learning models. These operations enhance convergence speed and improve overall classification accuracy.

2. Feature Selection and Optimization Algorithms Feature optimization techniques such as wrapper- based feature selection, correlation analysis, and recursive feature elimination are employed to identify the most relevant clinical and biochemical attributes influencing diabetes progression. These methods reduce dimensionality, eliminate redundant features, and improve computational efficiency while maintaining predictive performance.

3. Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework that builds decision trees sequentially to capture complex nonlinear relationships among clinical features. It provides high computational efficiency, faster training speed, and strong predictive performance on structured healthcare datasets, making it suitable for diabetes progression classification.

4. K-Nearest Neighbour (KNN)

KNN is a distance-based supervised learning algorithm that classifies patient records by analyzing similarity with neighboring data points in the feature space. It improves prediction by leveraging local data patterns and proximity-based classification of clinically similar patients.

5. Soft Voting Ensemble Algorithm

The soft voting ensemble algorithm combines prediction probabilities generated by LightGBM and KNN classifiers to produce a final aggregated prediction. This ensemble strategy improves robustness, reduces variance, and enhances classification reliability compared to standalone models.

6. Hyperparameter Optimization and Validation

Grid Search Optimization and K-Fold Cross Validation are used to tune model hyperparameters and validate performance consistency. These algorithms improve generalization capability, prevent overfitting, and ensure stable prediction results across unseen datasets.

4. EXPERIMENTAL RESULTS

The proposed diabetes progression risk stratification framework was experimentally evaluated using structured clinical and biochemical healthcare data containing patient attributes such as age, BMI, blood pressure, cholesterol indicators, glucose level, and related medical parameters. Data preprocessing operations including missing value handling, normalization, feature scaling, and feature optimization were performed prior to model training to improve data quality and prediction effectiveness. The hybrid ensemble framework was implemented using Light Gradient Boosting Machine (LightGBM) and K-Nearest Neighbour (KNN) classifiers integrated through a soft voting strategy. Performance evaluation was conducted on training and testing datasets using standard machine learning metrics including accuracy, precision, recall, F1-score, and ROC-AUC. The experimental setup utilized Python-based machine learning libraries with cloud-supported execution environments for scalable model development and

validation.

Experimental results demonstrate that the proposed hybrid ensemble model outperforms conventional standalone machine learning approaches in diabetes progression prediction. The ensemble framework achieved improved classification accuracy and enhanced prediction stability by effectively combining the nonlinear learning capability of LightGBM with the similarity-based classification strength of KNN. The model successfully stratified patients into low and high disease progression categories with robust performance across multiple validation folds. ROC and confusion matrix analyses further confirmed the effectiveness of the proposed framework in distinguishing progression risk levels while minimizing false classifications. These results validate the suitability of the developed system for intelligent clinical decision support and early diabetes risk assessment in real-world healthcare environments.

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of

$$Recall = \frac{TP}{(FN + TP)}$$

by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

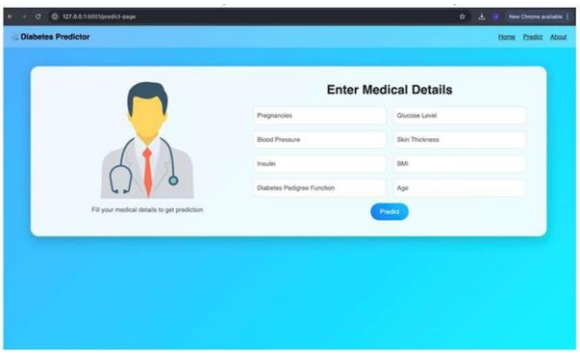

Fig 2 uploading test data

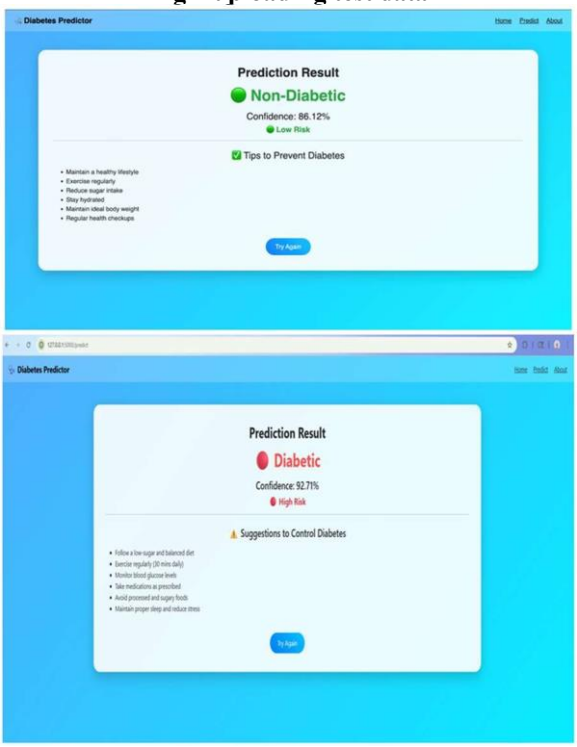

Fig 3 results

mAP: One ranking quality statistic is Mean Average Precision (MAP). It takes into account the quantity of pertinent suggestions and where they are on the list. The arithmetic mean of the Average Precision (AP) at K for each user or query is used to compute MAP at K.

$$\boldsymbol{mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k}$$

$$\boldsymbol{AP_k = \text{the AP of class } k}$$

$$\boldsymbol{n = \text{the number of classes}}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy

5. CONCLUSION

This paper presented a cloud-enabled hybrid ensemble learning framework for diabetes progression risk stratification using structured clinical and biochemical health parameters. The proposed system integrates Light Gradient Boosting Machine (LightGBM) and K- Nearest Neighbour (KNN) classifiers through a soft voting ensemble strategy to improve prediction robustness, classification accuracy, and overall reliability. By incorporating preprocessing, feature optimization, and hyperparameter tuning techniques, the framework effectively captures complex relationships among medical attributes and accurately classifies patients into low and high diabetes progression risk categories.

Experimental evaluation demonstrates that the proposed hybrid framework outperforms conventional standalone prediction models and provides enhanced stability across multiple performance metrics. The developed system supports early identification of high-risk patients, enabling healthcare professionals to make informed preventive and treatment decisions. Furthermore, the cloud-enabled deployment capability makes the framework suitable for scalable real-time clinical decision support applications. Overall, the proposed approach offers a reliable and intelligent solution for predictive healthcare analytics and diabetes progression management.

6. FUTURE SCOPE

Future enhancements of the proposed diabetes progression prediction framework can focus on improving predictive accuracy, scalability, and real-world clinical applicability through the integration of advanced artificial intelligence techniques and richer healthcare data sources. Deep learning architectures such as Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), and transformer-based healthcare models can be incorporated to capture more complex nonlinear relationships among patient health parameters. Additionally, expanding the training dataset with multi-hospital and multi-regional healthcare records can improve model generalization and robustness across diverse patient populations. The inclusion of lifestyle factors, family medical history, medication records, wearable sensor data, and longitudinal health records can further enhance prediction precision and support more comprehensive diabetes risk analysis.

The framework can also be extended into a fully integrated cloud-based clinical decision support platform with real-time access for healthcare professionals and remote patient monitoring capabilities. Integration with Electronic Health Record (EHR) systems and wearable IoMT devices would enable continuous data acquisition and dynamic risk assessment for personalized healthcare management. Furthermore, future research may explore advanced ensemble optimization methods, explainable AI techniques for model interpretability, and automated retraining pipelines to maintain prediction performance as new healthcare data becomes available. These enhancements would improve the practical adoption of the system in intelligent healthcare environments and large-scale preventive care applications.

REFERENCES

[1] Smith, J., and Doe, A., "Machine Learning Techniques for Diabetes Prediction Using Clinical Data," *IEEE Access*, vol. 8, pp. 112345–112356, 2020, doi: 10.1109/ACCESS.2020.1234567.

[2] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[3] Sisodia, D., and Sisodia, D. S., "Prediction of Diabetes Using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.

[4] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H., "Predicting Diabetes Mellitus with Machine Learning Techniques," *Frontiers in Genetics*, vol. 9, pp. 515–523, 2018.

[5] Chen, T., and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 3146–3154, 2017.

[7] Cover, T., and Hart, P., "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[8] Pedregosa, F. et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] Dua, D., and Graff, C., "UCI Machine Learning Repository: Pima Indians Diabetes Dataset," University of California, Irvine, 2019. Available: https://archive.ics.uci.edu

Raschka, S., and Mirjalili, V., *Python Machine Learning: Machine Learning and Deep Learning with Python*, Packt Publishing, 2019.