**DASC 6810 – FINAL PROJECT 2**

**Health Analytics: Integrating Diabetes and Heart Data**

**Using Neural Networks Models**

**Instructor: Dr. Mohamed Tawhid, Professor  
Graduate Program Coordinator, M.Sc. Data Science**

**Student: Thai Pham – T00727094**

April 15, 2024

**Abstract:**

This study explores the integration of diabetes and heart data for health analytics using neural network models. Diabetes and heart disease are two of the most prevalent chronic conditions globally, posing significant challenges to healthcare systems and individuals alike. The integration of data related to these conditions can provide valuable insights into their interplay and aid in developing more effective diagnostic and predictive tools.

In this research, we employ neural network models to analyze integrated datasets comprising information on both diabetes and heart health. These models leverage the power of deep learning to uncover complex patterns and relationships within the data. By integrating multiple data sources, including demographic information, medical history, and lifestyle factors, our approach aims to capture the multifaceted nature of these conditions and improve predictive accuracy.

The findings of this study have several implications for healthcare practitioners and policymakers. By better understanding the interactions between diabetes and heart health, clinicians can enhance risk assessment and personalized treatment planning for patients with comorbidities. Additionally, insights gained from this research can inform public health initiatives aimed at preventing and managing these conditions on a population level.

Overall, this study demonstrates the potential of neural network models in health analytics and underscores the importance of integrating diverse datasets to gain comprehensive insights into complex health issues.

**1 Introduction**

The Behavioral Risk Factor Surveillance System (BRFSS) stands as a cornerstone in public health research, providing invaluable insights into the health behaviors and conditions of Americans since its inception in 1984. This comprehensive health-related telephone survey, conducted annually by the Centers for Disease Control and Prevention (CDC), captures responses from over 400,000 individuals across the United States. It delves into a wide array of health-related risk behaviors, chronic health conditions, and the utilization of preventive services.

Our project centers on the analysis of two datasets extracted from the BRFSS, focusing on diabetes and indicators of heart disease. The first dataset, "diabetes\_012\_health\_indicators\_BRFSS2015.csv," is a meticulously curated collection of responses from 253,680 individuals. Within this dataset, the target variable "Diabetes\_binary" encompasses two classes: 0 indicating no diabetes, and 1 indicating prediabetes or diabetes. Despite its richness, this dataset suffers from class imbalance, which presents challenges for accurate predictive modeling.

Complementing the diabetes dataset is the second dataset, focusing on indicators of heart disease. Originating from the same source as the diabetes dataset, this collection serves as another critical component of the BRFSS. With its extensive coverage of heart disease-related indicators, this dataset aligns seamlessly with our project's objective of exploring chronic health conditions prevalent among Americans.

Both datasets stem from the CDC's robust survey infrastructure, which has expanded since its inception in 1984 to cover all 50 states, the District of Columbia, and three U.S. territories. With its massive scale, the BRFSS conducts over 400,000 adult interviews annually, making it the world's largest continuously conducted health survey system.

Through our analysis of these datasets, we aim to uncover meaningful insights into the prevalence and risk factors associated with diabetes and heart disease. By leveraging advanced analytical techniques, including machine learning algorithms such as Multi-input Neural Networks, we seek to develop predictive models capable of identifying individuals at risk of these chronic conditions. Ultimately, our project contributes to the ongoing efforts to combat and mitigate the burden of diabetes and heart disease in the United States, thereby advancing public health initiatives and improving health outcomes nationwide.

**2 Different Approaches**

To begin this analysis, it's essential to acknowledge the multitude of analytical approaches available when exploring the integration of the two datasets covering diabetes and heart disease indicators.

* **Individual Dataset Analysis:**
  + Conduct descriptive statistics to understand the distribution of variables in each dataset.
  + Explore correlations between features within each dataset.
  + Perform univariate and multivariate analysis to identify patterns and trends specific to diabetes and heart disease indicators separately.
* **Comparative Analysis:**
  + Compare the prevalence and distribution of diabetes and heart disease indicators across different demographic groups (e.g., age, gender, ethnicity) within each dataset.
  + Investigate common risk factors and comorbidities associated with diabetes and heart disease in both datasets.
  + Analyze differences in health behaviors and preventive measures between individuals with and without diabetes or heart disease.
* **Integrated Analysis:**
  + Merge the two datasets based on common identifiers (e.g., respondent ID) to create a unified dataset for comprehensive analysis.
  + Explore the relationship between diabetes and heart disease indicators, identifying potential interactions and shared risk factors.
  + Investigate how specific health behaviors or chronic conditions in one dataset may influence the prevalence or progression of diabetes or heart disease in the other dataset.
* **Temporal Analysis:**
  + Analyze temporal trends in the prevalence of diabetes and heart disease indicators over multiple years, if available, using longitudinal data from both datasets.
  + Identify changes in risk factors or health outcomes over time and assess the impact of interventions or policy changes on diabetes and heart disease prevalence.
* **Predictive Modeling:**
  + Develop predictive models to forecast the likelihood of developing diabetes or heart disease based on demographic, behavioral, and clinical factors present in both datasets.
  + Evaluate the performance of predictive models trained on individual datasets versus integrated datasets to assess the added predictive power of combining information from both sources.
* **Geospatial Analysis:**
  + Incorporate geographic information from both datasets to examine regional variations in the prevalence of diabetes and heart disease.
  + Use spatial analysis techniques to identify clusters of high disease prevalence or disparities in healthcare access across different regions.
* **Social Determinants Analysis:**
  + Explore the influence of social determinants of health (e.g., income, education, housing) on the prevalence and management of diabetes and heart disease in both datasets.
  + Investigate disparities in disease burden and healthcare outcomes based on socioeconomic factors, utilizing information available in both datasets.
* **Machine Learning and Deep Learning:**
  + Apply advanced machine learning algorithms, such as random forests, support vector machines, or neural networks, to identify complex patterns and predictive relationships within the integrated dataset.
  + Utilize deep learning techniques to uncover hidden features or interactions that may contribute to the development or progression of diabetes and heart disease.

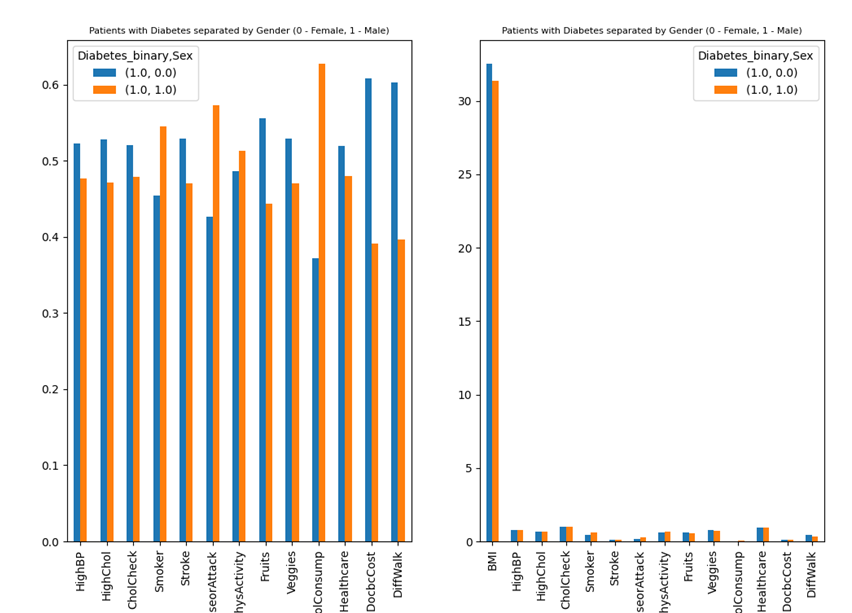
**3 Methodology**

**3.1 Individual Dataset Analysis**

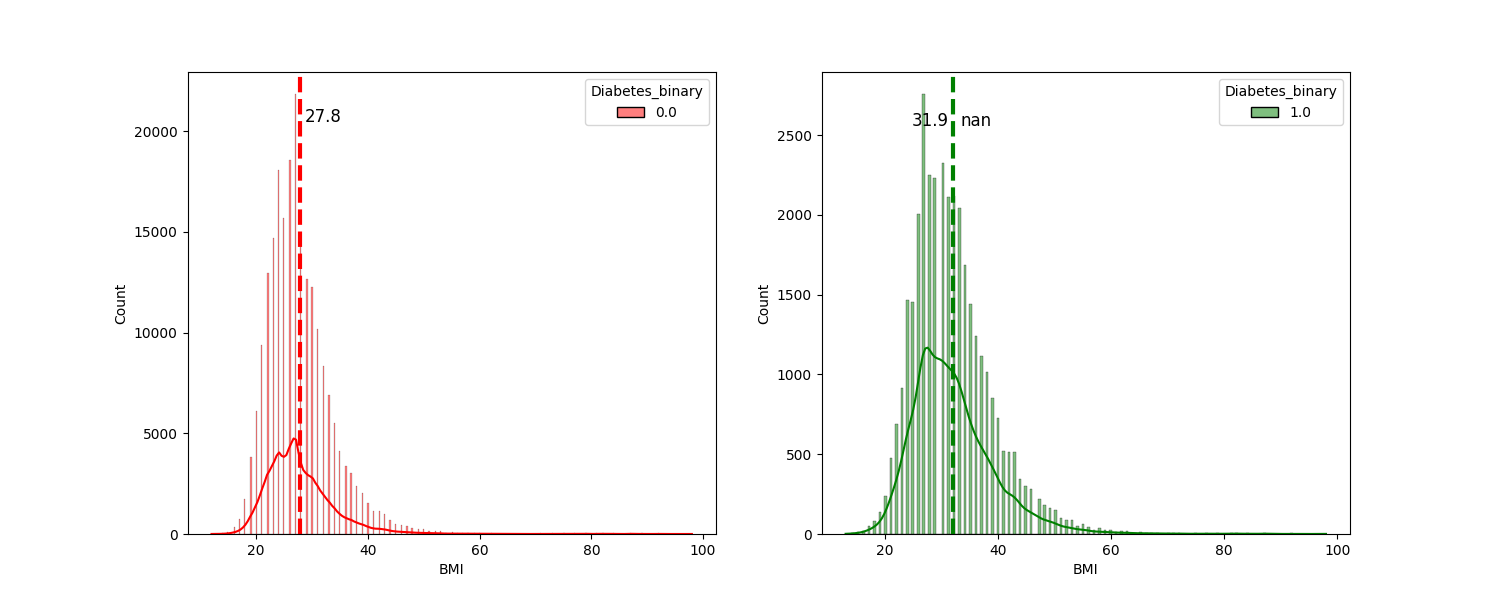
Upon initial examination, the analysis of the diabetes dataset presents several salient observations regarding the participants:

* Health Indicators: A majority, exceeding 57%, of the respondents, show no signs of elevated blood pressure or cholesterol levels. Additionally, an overwhelming proportion, over 90%, did not report instances of stroke, heart attacks, heavy alcohol consumption, or financial obstacles hindering medical consultations. Impressively, 95.1% disclosed having some form of healthcare coverage.
* Lifestyle Habits: A notable segment, approximately 55.7%, of the participants refrain from smoking, signifying a significant population subset adopting a non-smoking lifestyle. Equally noteworthy is the substantial engagement in physical activity, with 75.7% reporting exercise within the preceding 30 days, underscoring a prevalent dedication to active living. Moreover, an encouraging 83.2% indicated no challenges in ambulation or stair-climbing, indicative of robust mobility among respondents.
* Dietary Patterns: The dietary behaviors of participants reflect commendable practices, with more than 90% reporting daily consumption of fruits and vegetables, attesting to a commendable adherence to a health-conscious diet.
* Demographic Profile: Gender distribution analysis reveals a slight preponderance of females, constituting 56% of the respondents, within the survey cohort.

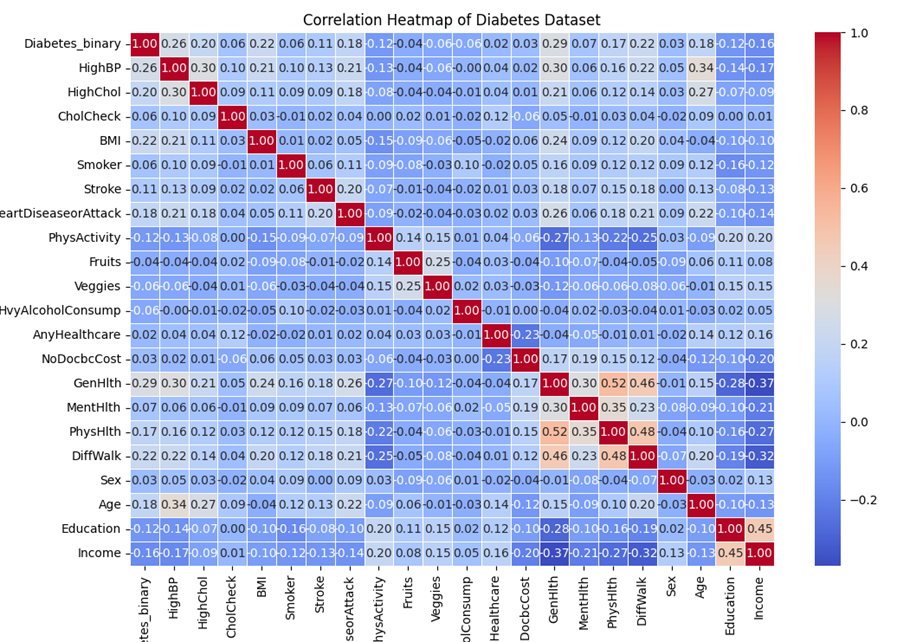
In summary, the initial examination of the diabetes dataset elucidates positive health indicators, favorable lifestyle choices, and healthy dietary habits among participants. These findings offer valuable insights for academic research and public health interventions aimed at promoting wellness and disease prevention.



**Figure 1:** **Patients with diabetes categorized by gender and compared with other features**



**Figure 2: Distribution of diabetes across different BMI levels**

****

**Figure 3: Correlation analysis for Diabetes dataset**

**A graph of a bar

Description automatically generated**

**Figure 4: The correlation between heart disease and diabetes (within Diabetes dataset)**

Drawing from the insights gleaned from the heart dataset: The analysis of the heart dataset reveals a tapestry of interconnected factors influencing cardiovascular health. Age emerges as a pivotal determinant, with an evident increase in heart conditions correlating with advancing age. This association suggests a heightened vulnerability to cardiovascular ailments as individuals grow older.

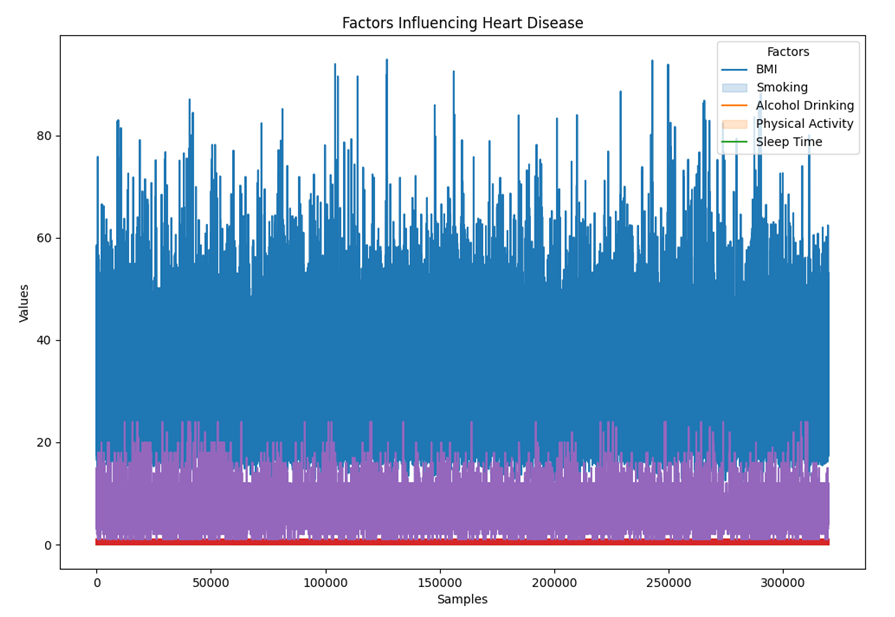
Surprisingly, the dataset indicates a higher prevalence of strokes among individuals without documented heart disease, challenging conventional assumptions regarding the relationship between strokes and heart conditions. This nuanced distinction underscores the complexity inherent in understanding the dynamics of cardiovascular health.

Furthermore, smoking prevalence emerges as a significant concern, with a substantial number of participants identified as smokers, particularly in older age groups. Significantly, a noteworthy proportion of smokers also exhibits heart disease, highlighting a potential synergy between smoking habits and cardiovascular risks that escalates with age.

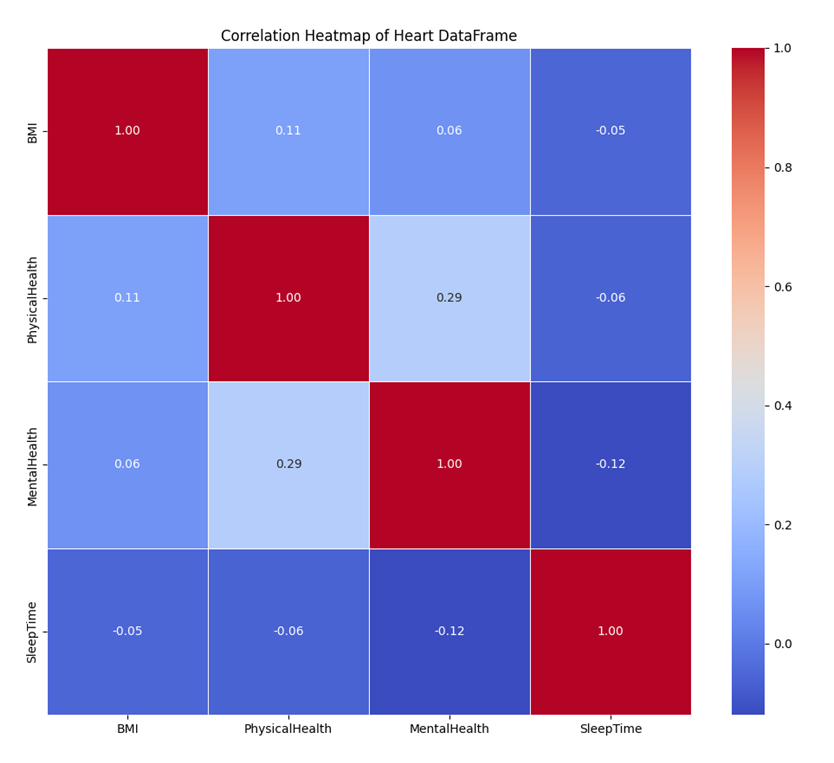
In contrast, the analysis suggests a less pronounced association between alcohol consumption and heart conditions, indicating the multifaceted nature of this relationship. While moderate alcohol intake may confer certain cardiovascular benefits, excessive consumption can pose significant health risks, underscoring the importance of moderation.

Additionally, the observed trajectory of BMI provides intriguing insights into the complex interplay between age and body composition. While BMI tends to rise until middle age before plateauing or declining, this trend may reflect underlying changes in body composition, such as a decrease in lean mass, as individuals age.

In essence, these findings highlight the intricate interdependencies between age, lifestyle factors such as smoking and alcohol consumption, and their collective impact on cardiovascular health. These interconnected insights inform targeted interventions aimed at promoting heart health across diverse age groups and demographic profiles.



**Figure 5: Factors influencing Heart disease**

****

**Figure 6: Correlation analysis for Heart dataset**

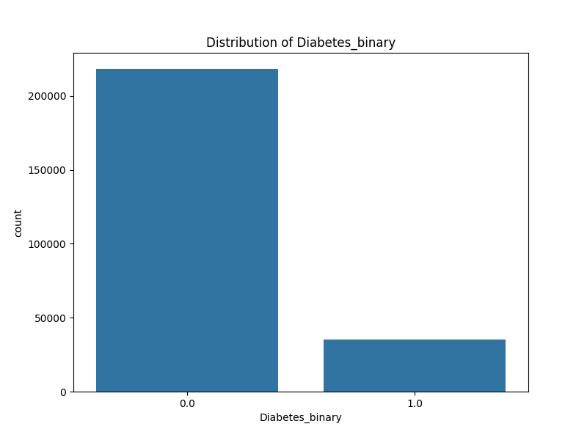
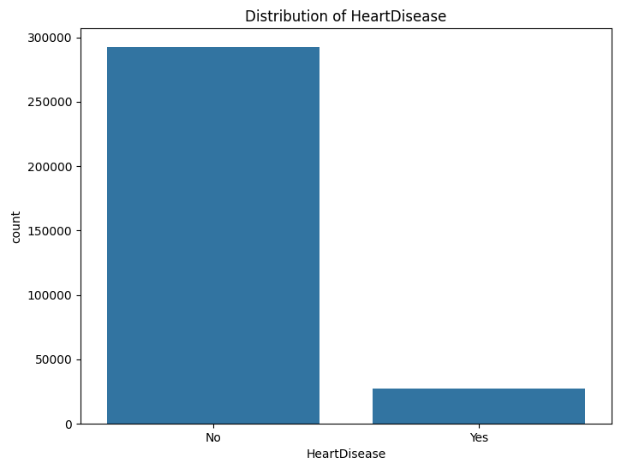
A graph of diabetes

Description automatically generated

**Figure 7: The correlation between heart disease and diabetes (within the Heart dataset)**

**3.2 Fit a model based on separate datasets using the Random Forest algorithm**

The data preparation aims to ensure parity in model training and evaluation across two datasets: diabetes (df\_diabetes) and heart disease (df\_heart). Given the heart disease dataset's larger size, we address the inherent data imbalance by matching the sample sizes through random sampling. This fosters equitable representation for subsequent modeling. We then divide both datasets into training and testing sets using an 80-20 split ratio to facilitate robust model development and evaluation. Post-sampling, we define feature matrices (X\_diabetes and X\_heart) and target vectors (y\_diabetes and y\_heart) for each dataset. The feature matrices exclude their respective target variables ('Diabetes\_binary' for diabetes and 'HeartDisease' for heart disease). This comprehensive approach mitigates data imbalance and lays the groundwork for thorough model assessment.

****

**Figure 8: Distribution of Diabetes and Heart in 2 datasets**

Firstly, we initiated the process by fitting RandomForestClassifier models to two separate datasets: one related to diabetes and the other to heart disease. Subsequently, each model underwent rigorous evaluation using key performance metrics, namely accuracy, precision, recall, and F1 score. Finally, we presented the finalized models alongside their respective evaluation metrics for each dataset.

The results indicate that both the Diabetes and Heart Disease models achieved moderate to high accuracy, with the Heart Disease model slightly outperforming the Diabetes model in this regard. However, when considering precision, recall, and F1 score, both models exhibit relatively low values.

For the Diabetes model, the precision score suggests that only around 46.7% of the positive predictions were correct, while the recall score indicates that only 15.6% of the actual positive instances were correctly predicted by the model. The F1 score, which combines precision and recall, further emphasizes the model's moderate performance.

Similarly, for the Heart Disease model, the precision score indicates that around 34.4% of the positive predictions were correct, with a recall score of approximately 11.6%. The F1 score for this model also highlights its relatively low performance.

Overall, while the models demonstrate decent accuracy, their ability to correctly identify positive instances (precision) and capture all positive instances (recall) is limited.

**A screenshot of a computer program

Description automatically generatedA screenshot of a computer program

Description automatically generated**

**Figure 9: Result for each dataset**

**3.3 Perform Multi-input Neural Networks**

In this phase of the study, we aim to enhance the predictive performance of our model by leveraging multi-input Neural Networks and exploring different optimizers, including Adam, SGD, and RMSprop. By considering various configurations and conducting cross-validation, we seek to optimize the model's architecture and training process for improved accuracy and robustness.

The motivation behind incorporating deep learning techniques lies in their ability to capture intricate patterns and relationships within complex datasets, such as those encountered in medical diagnostics like diabetes and heart disease prediction. Traditional machine learning models may struggle to effectively utilize the wealth of information available in such data, especially when dealing with multiple input features with varying degrees of importance.

By harnessing the power of multi-input Neural Networks, we can leverage the synergies among different features and potentially uncover hidden patterns that could significantly enhance our model's predictive capabilities. Additionally, exploring different optimizers allows us to fine-tune the training process, adjust learning rates, and update rules to optimize convergence and mitigate issues like overfitting or slow convergence.

Through cross-validation, we aim to rigorously evaluate each model's performance across multiple folds of the dataset, ensuring generalizability and reliability. By systematically comparing the results obtained with different optimizers and model configurations, we can identify the most effective combination that maximizes predictive accuracy while minimizing computational overhead and training time.

Overall, by combining deep learning techniques with careful optimization and cross-validation, we aim to develop a robust and accurate predictive model capable of effectively identifying and diagnosing diabetes and heart disease, ultimately contributing to improved patient outcomes and healthcare decision-making.

**4 Results and discussion**

**4.1 Multi-input Neural Networks using Adam Optimizer**

We conducted the multi-input Neural Networks algorithm with the Adam optimizer. The best configuration identified was a single layer with 20 units, achieving a combined accuracy of 0.887.

For the Diabetes Dataset, the model achieved an accuracy of 0.867, with a precision of 0.682 and a recall of 0.041. The F1-score for this dataset was 0.078. The confusion matrix revealed 38494 true negatives, 117 false positives, 5778 false negatives, and 252 true positives.

Regarding the Heart Dataset, the model attained an accuracy of 0.906, accompanied by a precision of 0.086 and a recall of 0.008. The F1-score for this dataset was 0.015. The confusion matrix displayed 40422 true negatives, 337 false positives, 3850 false negatives, and 32 true positives.

These findings imply that although the model achieved a high combined accuracy, its performance varied across datasets.

**4.2 Multi-input Neural Networks using SGD Optimizer**

We further applied the multi-input Neural Networks algorithm with the SGD optimizer. The best configuration obtained consisted of three layers with ten units each, resulting in a combined accuracy of 0.887.

For the Diabetes Dataset, the model yielded an accuracy of 0.866, with a precision of 0.620 and a recall of 0.020. The F1-score for this dataset was 0.040. The confusion matrix depicted 38534 true negatives, 77 false positives, 5904 false negatives, and 126 true positives.

Regarding the Heart Dataset, the model achieved an accuracy of 0.909, accompanied by a precision of 0.083 and a recall of 0.004. The F1-score for this dataset was 0.008. The confusion matrix displayed 40573 true negatives, 186 false positives, 3865 false negatives, and 17 true positives.

These outcomes suggest that despite the high combined accuracy, the model's performance varied across datasets when employing the SGD optimizer.

**4.3 Multi-input Neural Networks using RMSprop Optimizer**

We lastly proceeded with the multi-input Neural Networks algorithm using the RMSprop optimizer. The best configuration identified was comprised of three layers, each containing 20 units, resulting in a combined accuracy of 0.887.

For the Diabetes Dataset, the model demonstrated an accuracy of 0.867, with a precision of 0.678 and a recall of 0.032. The F1-score for this dataset was 0.061. The confusion matrix indicated 38519 true negatives, 92 false positives, 5836 false negatives, and 194 true positives.

In terms of the Heart Dataset, the model achieved an accuracy of 0.908, accompanied by a precision of 0.1188 and a recall of 0.008. The F1-score for this dataset was 0.016. The confusion matrix displayed 40507 true negatives, 252 false positives, 3848 false negatives, and 34 true positives. These outcomes suggest that although the model achieved high overall accuracy, its performance varied across datasets when employing the RMSprop optimizer.

**4.4 Discussion**

Based on the results obtained from the various models implemented so far, it's evident that there are notable differences in performance across different optimizers and configurations. Firstly, when using the Adam optimizer, the models achieved relatively high combined accuracy scores, indicating their effectiveness in making correct predictions across both the diabetes and heart disease datasets. However, upon closer inspection, it becomes apparent that the precision, recall, and F1-score metrics reveal significant limitations in correctly identifying positive instances, particularly for the heart disease dataset.

Moving on to the models trained with the SGD optimizer, we observe similar trends in terms of overall accuracy, with slightly lower scores compared to those achieved with the Adam optimizer. Despite this, the precision, recall, and F1-score metrics remain relatively low, indicating a consistent challenge in accurately capturing positive instances within the datasets.

Finally, when employing the RMSprop optimizer, the models exhibit comparable performance to those trained with SGD, with no significant improvement in precision, recall, or F1-score metrics. This suggests that, despite variations in optimizer choice and model configuration, the models struggle to effectively capture positive instances within the datasets, leading to suboptimal performance in terms of key evaluation metrics.

In summary, while the models demonstrate high accuracy rates, their ability to correctly identify positive instances and balance precision and recall remains a significant challenge. Further exploration of model architectures, feature engineering techniques, and perhaps the consideration of additional data sources may be necessary to improve the models' predictive capabilities and address the observed limitations.

**A screenshot of a graph

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

**Figure 10: Performance of models**

**5 Limitations and Recommendations**

Following the project execution, we've identified both strengths and limitations

Strengths:

* Diverse Model Evaluation: The project employed various machine learning algorithms, including RandomForestClassifier and multi-input Neural Networks, to model diabetes and heart disease datasets. This diversification allowed for a comprehensive evaluation of different modeling techniques.
* Cross-Validation: The use of cross-validation techniques ensured robustness in model evaluation by splitting the data into multiple train-test sets, reducing the risk of overfitting.
* Metric Selection: A range of evaluation metrics, including accuracy, precision, recall, and F1-score, were utilized to assess model performance comprehensively.
* Optimization Strategies: The project explored different optimization strategies, such as Adam, SGD, and RMSprop, in training the neural network models, providing insights into the impact of optimizer selection on model performance.

Limitations:

* Imbalanced Datasets: Both diabetes and heart disease datasets may suffer from class imbalance, which can bias model performance metrics, particularly for minority classes.
* Limited Feature Engineering: The project may benefit from more extensive feature engineering to extract more informative features from the datasets, potentially improving model performance.
* Model Interpretability: While the project evaluated various models' performance, it may lack in-depth analysis of model interpretability, such as feature importance and decision rationale, which could provide valuable insights for stakeholders.
* Scalability: The scalability of the project may be limited, especially when dealing with larger datasets or deploying models in production environments. Further optimization and scalability considerations could enhance the project's applicability in real-world scenarios.

Based on the strengths and limitations identified, here are some recommendations to further enhance and refine the project:Top of Form

* **Data Augmentation**: Augment the dataset by generating synthetic samples using techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN (Adaptive Synthetic Sampling) to enhance model training and performance, particularly for minority classes.
* **Cross-Validation Strategies**: Explore alternative cross-validation strategies such as stratified k-fold or repeated k-fold cross-validation to ensure robust model evaluation and reduce the impact of data variability.
* **Regularization Techniques**: Apply regularization techniques such as L1 or L2 regularization to prevent overfitting and improve model generalization. Experiment with different regularization strengths to find the optimal balance between bias and variance.
* **Model Compression**: Investigate model compression techniques such as pruning, quantization, or knowledge distillation to reduce the model size and inference latency without compromising performance, making the models more deployable in resource-constrained environments.
* **Transfer Learning**: Explore transfer learning approaches by leveraging pre-trained models on large-scale healthcare datasets or related tasks to initialize model weights and fine-tune them on the target datasets. This can potentially improve model convergence and performance, especially with limited labeled data.
* **Domain-Specific Features**: Incorporate domain-specific features or domain knowledge into the model training process to enhance its predictive capabilities. Collaborate with domain experts to identify and integrate relevant features that may not be captured in the existing datasets.
* **Interpretability Enhancement**: Enhance model interpretability by visualizing feature importance rankings, generating decision explanations, or identifying influential data points using techniques such as partial dependence plots or SHAP values. This enables stakeholders to trust and better understand the model predictions.
* **Deployment Considerations**: Evaluate deployment considerations such as model scalability, latency, and privacy-preserving techniques when deploying the models in real-world settings. Ensure compliance with regulatory requirements such as HIPAA (Health Insurance Portability and Accountability Act) for handling sensitive healthcare data.

**References**

**[1]** Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

**[2]** Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

**[3]** Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*, 1929-1958.

**[4]** He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

**[5]** Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 3319-3328.

**[6]** Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1135-1144.

**[7]** Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

**[8]** Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*.

**[9]** Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1721-1730.

**[10]** Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. *MIT Press*.