Project report on Data Driven Analysis for Remote Work and Mental Health

December 05, 2024

1. INTRODUCTION

The advent of remote work, accelerated by global events such as the COVID-19 pandemic, has reshaped traditional workplace dynamics. While remote work offers flexibility and convenience, it also introduces unique challenges that may impact employees' mental well-being.

Key concerns include increased stress levels, difficulty in maintaining work-life balance, and the prevalence of mental health conditions. These factors can vary significantly across industries, regions, and demographics. Organizations increasingly recognize the importance of fostering a healthy remote work environment to maintain productivity and employee satisfaction.

This study aims to explore the impact of remote work on mental health using a data-driven approach. By leveraging analytical methods and machine learning techniques, we investigate the relationships between remote work factors and mental health outcomes. Our goal is to identify patterns, predict stress levels, and assess the satisfaction of employees with remote work arrangements.

The findings will provide valuable insights for organizations seeking to implement effective policies and practices to support their workforce in a remote or hybrid working model.

2. DATA PREPROCESSING

The dataset contains several columns with missing values that need to be addressed. For example, the Mental_Health_Condition column has 1196 missing values, and the Physical_Activity column has 1629 missing values. These missing values can lead to inconsistencies in data analysis and model training, so handling them appropriately is crucial.

For the Mental_Health_Condition column, since it is likely a categorical variable, a suitable approach would be to replace the missing values with 'Unknown' or the most frequently occurring category in this column. This will ensure that no rows are excluded due to missing values while maintaining the integrity of the data.

The Physical_Activity column, which appears to be a numerical variable, could be imputed with the mean or median of the column. Imputing with the mean is effective when the data is symmetrically distributed, while the median is better for skewed distributions to reduce the influence of outliers.

Columns without missing values, such as Age, Gender, Job_Role, and Work_Location, still require preprocessing to prepare them for analysis. For example, categorical columns like Gender and Work_Location should be encoded into numerical representations. One-hot encoding could be used for these variables to create binary columns for each category, which is useful for algorithms that require numerical inputs.

The numerical columns, such as Age, Years_of_Experience, and Hours_Worked_Per_Week, need to be scaled to bring them into a uniform range. Scaling techniques like Min-Max scaling or standardization ensure that features contribute equally to the model's performance, especially for distance-based algorithms like KNN.

Additionally, it is important to check for outliers in columns such as Hours_Worked_Per_Week and Physical_Activity. Outliers can be detected using statistical methods like the Interquartile Range (IQR) or visualization techniques like boxplots. Any detected outliers should be handled by capping, transformation, or removal, depending on their significance in the dataset.

The dataset also includes columns like Access_to_Mental_Health_Resources, Satisfaction_with_Remote_Work, and Social_Isolation_Rating, which could provide insights into correlations and trends. These variables should be prepared for analysis through encoding or scaling, as necessary.

Finally, before proceeding to analysis or modeling, all columns must be checked to ensure consistent data types. For instance, columns intended to represent categorical data should not have numerical data types. This consistency will help avoid issues during data analysis or while training machine learning models.

```
Employee_ID                          0
Age                                  0
Gender                               0
Job_Role                             0
Industry                             0
Years_of_Experience                  0
Work_Location                        0
Hours_Worked_Per_Week                0
Number_of_Virtual_Meetings           0
Work_Life_Balance_Rating             0
Stress_Level                         0
Mental_Health_Condition           1196
Access_to_Mental_Health_Resources    0
Productivity_Change                  0
Social_Isolation_Rating              0
Satisfaction_with_Remote_Work        0
Company_Support_for_Remote_Work      0
Physical_Activity                 1629
Sleep_Quality                        0
Region                               0
dtype: int64
```

3. EXPLORATORY DATA ANALYSIS

A. Employees in the 40s age range might be underrepresented compared to other age groups, indicating a potential gap or anomaly in the dataset.

B. Employees' distribution is uniform across different experience levels, with periodic dips around specific intervals (e.g., near 10, 20, and 30 years). These dips could indicate a potential lack of representation in those ranges or specific trends in the workforce composition.

C. Employees are evenly distributed across the three categories: Remote, Hybrid, and Onsite.

D. Diverse range of opinions about remote work satisfaction among employees, with no single category overwhelmingly dominant.

E. Stress levels are relatively evenly distributed across all work locations. No significant differences in stress levels are observed among the different work setups, indicating that stress might be influenced by factors other than work location.

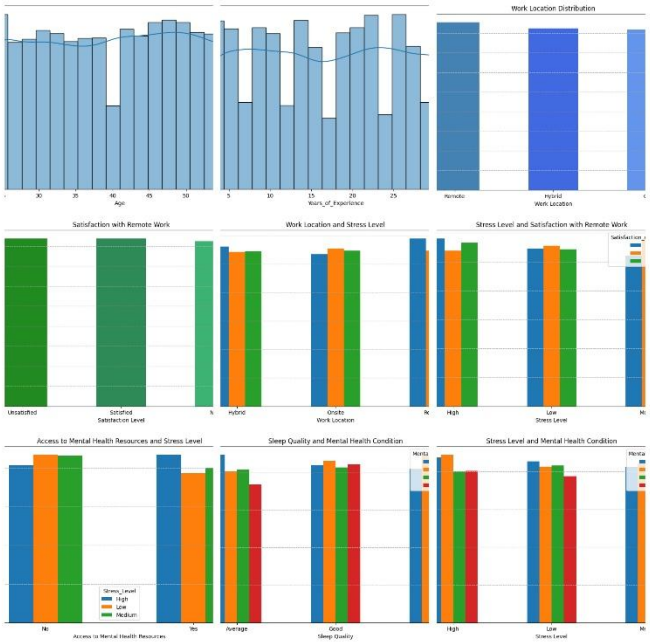
F. Individuals with low stress levels generally report higher satisfaction with remote work compared to those with high stress levels. Satisfaction levels are relatively

balanced across the stress categories, though high stress shows a slight increase in dissatisfaction. Neutral satisfaction is consistent across all stress levels, suggesting that factors other than stress may also influence remote work satisfaction.

G. Employees with no access to mental health resources tend to have slightly higher stress levels (High and Medium) compared to those with access. Providing access to mental health resources appears to help reduce high stress levels.

H. Employees with good sleep quality are distributed more evenly across mental health conditions. Poor sleep quality is associated with a higher occurrence of burnout and depression, highlighting the importance of sleep for mental health.

I. Anxiety and burnout are more prevalent in employees with high stress levels. Low stress levels are linked with fewer instances of burnout and anxiety, indicating the strong connection between stress and mental health conditions.



J. Across all work locations, daily physical activity slightly reduces social isolation compared to occasional or weekly physical activity. The difference in social isolation ratings is minimal, indicating that work location has limited influence on this aspect.

K. Employees in high-stress environments report lower satisfaction across all job roles. Job roles such as HR and Marketing show a broader range of satisfaction levels compared to more technical roles like Software Engineers.

L. The variability in work-life balance ratings is similar across stress levels.

M. The work-life balance is consistent across all age groups, with no noticeable differences influenced by age or satisfaction with remote work.

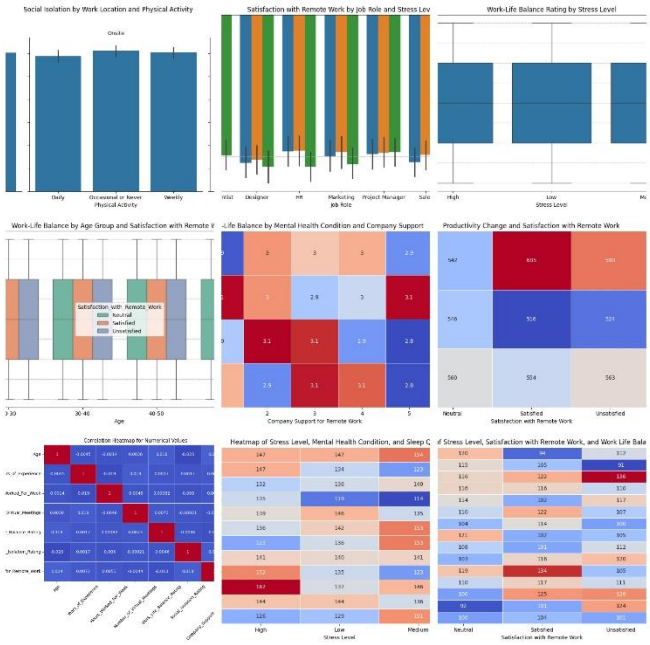
N. Burnout and Depression have slightly higher work-life balance ratings (around 3.1) when company support is moderate (2 or 3). Lower company support (1) generally results in slightly lower ratings (2.9) for most conditions. Unknown mental health conditions and low support (5) show the lowest ratings (2.8), indicating potential gaps in addressing employee needs.

O. Decreased Productivity is most common among those who are satisfied (605) and unsatisfied (590). Increased Productivity aligns more with neutral satisfaction (546) compared to satisfied or unsatisfied groups. No Change in Productivity is evenly distributed across all satisfaction levels, with slightly higher counts for neutral satisfaction (560).

P. Weak correlations across numerical variables as most values are close to 0. No strong relationships between key factors like Work-Life Balance Rating, Social Isolation Rating, or Hours Worked Per Week. Years of Experience and Age are slightly negatively correlated, likely reflecting diverse workforce demographics.

Q. High Stress Level is strongly associated with Poor Sleep Quality, especially for conditions like Burnout and Anxiety. Medium Stress Level occurs more often with Good Sleep Quality when mental health conditions like Burnout or Depression are present. Low Stress Level frequently aligns with Unknown Mental Health Conditions or Poor Sleep Quality.

R. High or Medium stress levels often correlate with both satisfied and neutral satisfaction levels. Low stress appears less impactful unless paired with a strong work-life balance. Work-life balance ratings above 3 are consistently associated with satisfied or neutral satisfaction levels.



4. DATA MODELLING

A. STRESS LEVEL PREDICTION:

Bayesian Optimization was utilized to fine-tune the hyperparameters of an XGBoost Classifier for predicting Stress_Level. This approach efficiently explored the hyperparameter space and aimed to maximize the model's performance on a multi-class classification task.

The dataset consisted of features such as Work_Hours_Per_Week, Sleep_Quality, and Physical_Activity, with Stress_Level as the target variable. It was observed that the dataset had an imbalanced distribution of classes. XGBoost was chosen as the modeling algorithm due to its robustness in handling both structured data and class imbalance.

The hyperparameter ranges included max_depth between 3 and 10, learning_rate between 0.01 and 0.1, n_estimators between 50 and 300, subsample between 0.4 and 1.0, colsample_bytree between 0.6 and 1.0, and regularization parameters reg_alpha and reg_lambda, both ranging from 0.0 to 1.0. The optimization process used a 5-fold stratified cross-validation to evaluate the model's performance, ensuring consistent evaluation and

preservation of class distribution across folds. The f1_macro metric was selected as the objective for optimization, balancing performance across all classes.

The Bayesian Optimization process began with 20 initial random points to build the surrogate model. It then performed 50 iterations to explore the parameter space further. The optimized hyperparameters included an intermediate tree depth, a moderately low learning rate, and a reasonable number of boosting rounds. Subsampling and feature sampling fractions were set close to 0.9, and regularization parameters were adjusted to balance underfitting and overfitting.

After optimization, the model's performance was evaluated. The accuracy was 34%, indicating poor predictive power. Precision, recall, and F1-scores for all classes hovered around 0.33 to 0.36, demonstrating weak separability among the classes. The confusion matrix highlighted significant misclassification across all classes, suggesting challenges in learning meaningful patterns from the data.

The poor performance, despite hyperparameter optimization, suggests underlying limitations in the dataset or feature representation. The dataset may require additional feature engineering to capture more informative patterns. Addressing class imbalance with techniques such as SMOTE or class-weighted loss functions could also improve model performance.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.35      0.37      0.36       335
           1       0.33      0.31      0.32       339
           2       0.34      0.34      0.34       326

    accuracy                           0.34      1000
   macro avg       0.34      0.34      0.34      1000
weighted avg       0.34      0.34      0.34      1000

Confusion Matrix:
 [[125 112  98]
 [119 105 115]
 [116  98 112]]
```

B. HEALTH RISK PREDICTION

Logistic Regression was employed to predict whether an individual's health is at risk (Is_Health_In_Risk) based on key factors such as stress level, sleep quality, and mental health conditions. This model was chosen due to its simplicity and effectiveness in binary classification tasks.

Feature engineering was an essential part of the modeling process. A custom function, classify_health_risk, was designed to account for the interplay between stress, sleep quality, and mental health conditions. If an individual had a Stress_Level classified as "High," they were automatically labeled as being at health risk. For those with "Medium" stress levels, the model incorporated additional criteria such as poor sleep quality or known mental health conditions to classify them as at risk. Similarly, individuals with "Low" stress levels but poor sleep quality or a mental health condition were also marked as at risk. This function enabled the model to capture complex relationships between multiple features and their contribution to health risks.

Hyperparameter tuning was conducted using GridSearchCV to optimize the Logistic Regression model. The tuning process explored various values of C, representing the inverse of regularization strength, along with different solver algorithms. The best combination was identified as C = 10 and the liblinear solver. Five-fold cross-validation with the f1 score as the evaluation metric ensured robust and reliable parameter selection.

The model's initial evaluation demonstrated strong overall accuracy of 94%. For Class 0 (not at risk), recall was perfect at 1.00, ensuring that all instances were correctly identified. However, the precision was relatively low at 0.63, indicating a higher number of false positives. For Class 1 (at risk), the model achieved an excellent precision of 1.00 and a recall of 0.94, showing strong performance for this class. Despite the high accuracy, the imbalance between precision and recall for Class 0 suggested room for improvement.

To address this, the decision threshold of the model was adjusted. After the adjustment, precision for Class 0 improved to 0.64, while recall remained perfect at 1.00. This adjustment ensured that all instances of Class 0 were correctly identified, albeit with some false positives still present. Class 1 retained its excellent precision and strong recall, ensuring minimal impact on its performance. The overall accuracy improved slightly to 95%, with a more balanced trade-off between precision and recall across both classes.

The Logistic Regression model demonstrated the value of feature engineering in capturing complex relationships among predictors. Hyperparameter tuning played a critical role in achieving optimal performance, while threshold adjustment helped balance precision and

recall for the minority class. Further improvements could involve incorporating additional features such as physical activity or diet, as well as validating the model on an independent dataset to confirm generalizability. These steps would help refine the model further and enhance its application in identifying individuals at health risk.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.64      1.00      0.78        98
           1       1.00      0.94      0.97       902

    accuracy                           0.95      1000
   macro avg       0.82      0.97      0.88      1000
weighted avg       0.97      0.95      0.95      1000

Confusion Matrix:
 [[ 98   0]
 [ 54 848]]
```

C. SATISFACTION WITH REMOTE WORK

The Random Forest Classifier was utilized to predict Satisfaction_with_Remote_Work. This model was selected for its capability to handle complex, non-linear relationships and its ability to reduce overfitting through ensemble techniques. Bayesian Optimization was employed to fine-tune the model's hyperparameters, allowing efficient exploration of the parameter space. The hyperparameters considered during optimization included the number of estimators (n_estimators), maximum depth of the trees (max_depth), the minimum number of samples required to split a node (min_samples_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf).

The ranges for these hyperparameters were defined as follows: n_estimators ranged between 50 and 300, max_depth ranged between 3 and 20, min_samples_split ranged between 2 and 10, and min_samples_leaf ranged between 1 and 5. A 5-fold stratified cross-validation was performed to ensure that the model evaluation was consistent and accounted for class imbalances. The f1_macro score was used as the optimization metric, emphasizing the importance of balanced performance across all classes. The Bayesian Optimization process began with 20 random initial points to explore the parameter space and continued for 50 iterations to refine the parameter search.

The best parameters obtained from the optimization process included n_estimators set to 95, max_depth set to 15, min_samples_split set to 6, and min_samples_leaf set to 4. These parameters provided an optimal balance between model complexity and generalization. After training

the model with these parameters, the performance was evaluated on standard classification metrics. The model achieved an overall accuracy of 36 percent, which was indicative of poor predictive performance. The precision, recall, and F1-scores for all classes ranged between 0.28 and 0.40, reflecting weak differentiation among the classes. The confusion matrix further revealed significant misclassifications across all classes, suggesting that the model failed to capture meaningful patterns in the dataset.

The low performance of the model despite hyperparameter tuning pointed to limitations in the dataset's features or potential issues with class imbalance. The features used in the model may not have been sufficient to predict Satisfaction_with_Remote_Work effectively. Additionally, the imbalance in class distribution likely impacted the model's ability to classify minority classes accurately. To address these issues, it is recommended to perform advanced feature engineering to derive more meaningful predictors. Techniques such as resampling through SMOTE or incorporating class-weighted training can help mitigate the effects of class imbalance. Furthermore, experimenting with alternative algorithms such as Gradient Boosting or Neural Networks may provide better predictive capabilities for this complex task.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.28      0.32      0.30       174
           1       0.39      0.39      0.39       231
           2       0.40      0.37      0.38       221

    accuracy                           0.36       626
   macro avg       0.36      0.36      0.36       626
weighted avg       0.36      0.36      0.36       626

Confusion Matrix:
 [[55 62 57]
 [77 89 65]
 [63 77 81]]
```

D. SATISFIED WITH REMOTE WORK

The Random Forest Classifier was employed to predict satisfaction with remote work. This section outlines the steps taken, from feature engineering to model evaluation, and highlights the results of the analysis.

A function was developed to classify satisfaction with remote work based on key predictors such as satisfaction level, stress level, and work-life balance rating. The function assigned a value of 1 to rows where satisfaction was categorized as 'Satisfied' or 'Neutral' and either the stress level encoded was greater than 1 or the work-life balance rating exceeded 3. Rows not meeting these conditions were assigned a value of 0. This step ensured that

critical features relevant to satisfaction and well-being were adequately represented in the dataset.To prepare the dataset for analysis, records corresponding to 'Onsite' work locations were excluded, as the focus was on remote or hybrid work scenarios. Furthermore, rows where satisfaction with remote work was marked as 'Unsatisfied' and the work-life balance rating exceeded 3 were removed. These actions helped eliminate noise and outliers, ensuring that the dataset was clean and reliable for modeling.

Hyperparameter tuning was conducted using Bayesian Optimization to enhance the performance of the Random Forest Classifier. The parameter bounds were defined as follows:

Number of estimators (n_estimators): 50–300

Maximum depth (max_depth): 3–20

Minimum samples split (min_samples_split): 2–10

Minimum samples leaf (min_samples_leaf): 1–5

The optimization process yielded the following best parameters:

max_depth: ~3.3

min_samples_split: ~4.6

min_samples_leaf: ~4.6

n_estimators: ~60.8

These parameters ensured the model was finely tuned for performance without overfitting.

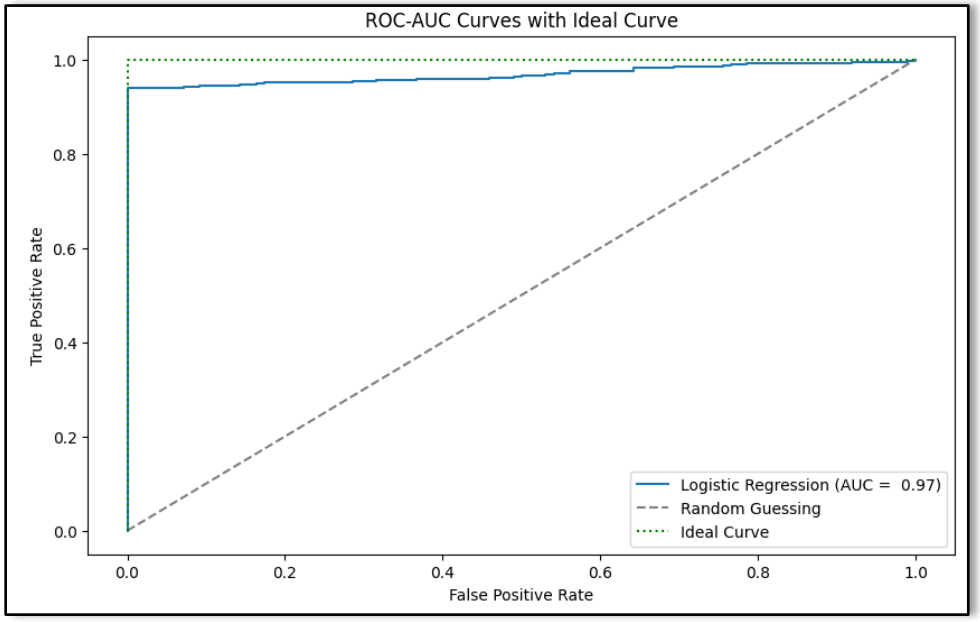
The trained Random Forest Classifier was evaluated using standard metrics, including precision, recall, and F1-score. For Class 0, the model achieved perfect precision (1.00), strong recall (0.90), and an F1-score of 0.95. For Class 1, the precision was 0.88, recall was perfect at 1.00, and the F1-score was 0.93. The overall accuracy of the model was 94%, indicating robust predictive capabilities. Despite these strong results, it was noted that recall for Class 0 was slightly lower, highlighting a potential area for improvement.

The feature importance analysis revealed that the work-life balance rating was the most significant predictor, with an important score of 0.538. Other important features included the stress level encoded (0.328) and years of experience (0.028). Additional features such as age, the number of virtual meetings, and company support for remote work also contributed to the model but with comparatively lower importance. This analysis provided valuable insights into the factors influencing satisfaction with remote work.

5. RESULTS AND VISUALIZATION

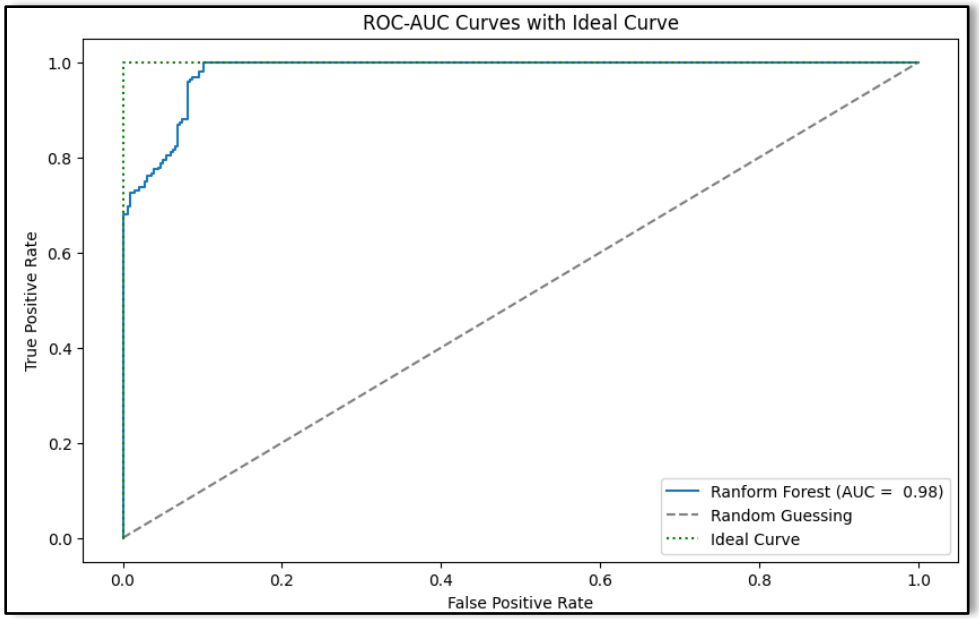
The ROC-AUC curve analysis highlights the model's performance in classification tasks, emphasizing its ability to differentiate between the classes effectively.

The ROC-AUC curve achieves a high AUC value of 0.97, indicating excellent performance in distinguishing between the two classes. An AUC value close to 1 reflects a highly effective classifier. The curve closely aligns with the "Ideal Curve," demonstrating low false positive rates and high true positive rates. This alignment suggests the model's robust predictive power and minimal errors in classification.



The ROC-AUC curve achieves an even higher AUC value of 0.98, showcasing near-perfect discrimination between the classes. The curve closely matches the "Ideal Curve," reinforcing the model's effectiveness in minimizing false positives and maximizing true positives. This result signifies that the Random Forest model is exceptionally proficient at capturing patterns and making accurate predictions for the classification task.

Both analyses underscore the importance of the ROC-AUC curve as a comprehensive metric for evaluating classification models, offering insights into their ability to balance sensitivity and specificity effectively.


ROC-AUC Curves with Ideal Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Ranform Forest (AUC = 0.98)
Random Guessing
Ideal Curve