Uso de framework de aprendizje máquina para la implementación de Predicción de diabetes.

Alan Alberto Mota Yescas

September 7, 2024

Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos
Profesor: Jesús Adrián Rodríguez Rocha
Fecha de entrega: September 7, 2024

Contents

1 Introducción

En este reporte se aborda el análisis de un dataset con el objetivo de prevenir la aparición de diabetes en pacientes que la desarrollarán en el futuro mediante el uso de las técnicas de machine learning. Para tal fin, se desarrolla un modelo de Regresión Logística utilizándolo en la exploración previa de los datos junto con el entrenamiento del propio modelo y la búsqueda de hiperparámetros óptimos basados en datos de entrenamiento. La finalidad principal radica en la mejora de las capacidades predictivas, en particular, al intentar predecir diabetes usando datos clínicos relevantes y significativos. Para asegurarse de que el modelo sea robusto.

2 Metodología

Esta sección describe los pasos seguidos para preparar, procesar y analizar los datos antes de aplicar el modelo de regresión logística. Incluye desde la carga de los datos, hasta la exploración visual, normalización de datos, correlación y la separación en conjuntos de entrenamiento y prueba.

2.1 Carga y preprocesamiento de datos

En este análisis, los datos fueron cargados desde un archivo CSV que contiene información clínica de pacientes. El dataset incluye diversas variables relevantes para la predicción de la diabetes, las cuales se deben procesar adecuadamente antes de entrenar el modelo.

El dataset contiene las siguientes columnas, que describen distintas características de los pacientes:

- **Pregnancies**: Número de embarazos.
- **Glucose**: Concentración de glucosa en sangre.
- **BloodPressure**: Presión sanguínea diastólica (mm Hg).
- **SkinThickness**: Grosor del pliegue cutáneo tricipital (mm).
- **Insulin**: Niveles de insulina en suero (µU/ml).
- **BMI**: Índice de masa corporal (peso en kg / altura en m^2).
- **DiabetesPedigreeFunction**: Función de pedigrí de la diabetes, que mide la probabilidad de desarrollar diabetes basada en factores hereditarios.
- **Age**: Edad de los individuos.
- **Outcome**: Variable objetivo que indica si el paciente tiene (1) o no tiene (0) diabetes.

2.2 Estadísticas descriptivas

Estas estadísticas permiten observar la distribución de las características, identificando valores promedio, mínimos, máximos y rangos de las variables que serán utilizadas para predecir la diabetes. Los resultados de las estadísticas descriptivas se muestran en la tabla a continuación.

El dataset cuenta con un total de 768 observaciones y las siguientes variables:

- **Pregnancies**: El número de embarazos oscila entre 0 y 17, con una media de 3.85 embarazos. Esto sugiere que la mayoría de las mujeres en el dataset han tenido entre 3 y 4 embarazos, aunque algunas no han tenido ninguno.
- **Glucose**: La concentración de glucosa en sangre tiene una media de 120.89, con un rango de valores entre 0 y 199. Un nivel de glucosa cercano a 0 es inusual y probablemente representa datos faltantes o erróneos, lo que podría requerir un mayor análisis.
- **BloodPressure**: La presión sanguínea diastólica tiene un valor promedio de 69.1 mmHg, con valores que oscilan entre 0 y 122 mmHg. Al igual que con la glucosa, un valor de 0 para la presión arterial es atípico y podría indicar datos faltantes.
- **SkinThickness**: El grosor del pliegue cutáneo tricipital varía entre 0 y 99 mm, con una media de 20.5 mm. Los valores de 0 nuevamente podrían representar datos no disponibles en lugar de valores reales.
- **Insulin**: Los niveles de insulina muestran una media de 79.8, con un rango extremadamente amplio de 0 a 846, lo que sugiere la presencia de valores atípicos.
- **BMI (Índice de Masa Corporal)**: El IMC tiene una media de 31.99, con valores que oscilan entre 0 y 67.1. Los valores cercanos a 0 también podrían representar datos faltantes.
- **DiabetesPedigreeFunction**: La función de pedigrí de la diabetes, que mide la susceptibilidad hereditaria, tiene un valor promedio de 0.47, con un rango de 0.078 a 2.42.
- **Age**: La edad de los pacientes varía entre 21 y 81 años, con una media de 33.24 años.

La tabla siguiente resume estas estadísticas descriptivas:

Variable	Media	Desviación Estándar	Mínimo	Máximo
Pregnancies	3.85	3.37	0	17
Glucose	120.89	31.97	0	199
BloodPressure	69.1	19.36	0	122
SkinThickness	20.54	15.95	0	99
Insulin	79.8	115.24	0	846
BMI	31.99	7.88	0	67.1
DiabetesPedigreeFunction	0.47	0.33	0.078	2.42
Age	33.24	11.76	21	81
Outcome	0.35	0.48	0	1

Estas estadísticas proporcionan una primera impresión del rango y la variabilidad de cada una de las características del dataset. Se observa que varias variables contienen valores atípicos o posiblemente datos faltantes (como aquellos con valor 0), lo que podría afectar el rendimiento del modelo y requerir un preprocesamiento adicional.

Estas estadísticas proporcionan una primera impresión del rango y la variabilidad de cada una de las características del dataset. Se observa que varias variables contienen valores atípicos o posiblemente datos faltantes (como aquellos con valor 0), lo que podría afectar el rendimiento del modelo y requerir un preprocesamiento adicional, donde en la siguiente sección serán tratados.

2.3 Visualización de variables principales

A continuación se muestran los histogramas de las principales variables para observar sus distribuciones.

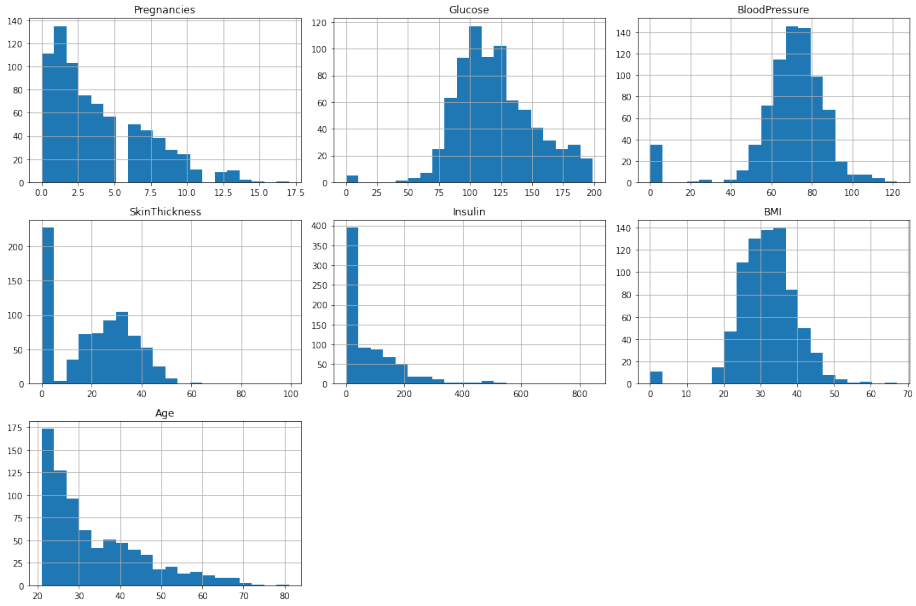

Figure 1: Histogrmas de variables principales

Observamos lo siguiente:

- **Pregnancies**: La mayoría de las personas tienen entre 0 y 10 embarazos, pero hay algunos valores más altos.
- **Glucose**: La glucosa tiene una distribución cercana a la normal, pero hay valores cercanos a 0, lo que podría ser un error o dato faltante.
- **BloodPressure**: Hay un número significativo de valores en 0, lo que no es realista y sugiere datos faltantes.
- **SkinThickness**: La mayoría de los valores están entre 0 y 50, pero también hay muchos valores de 0.
- **Insulin**: Hay muchos valores de 0, lo que indica datos faltantes.
- **BMI**: La distribución es más normal, pero nuevamente hay algunos valores de 0.
- **Age**: La edad tiene una distribución sesgada hacia los más jóvenes.

2.4 Normalización de datos

Como se observó en los histogramas, algunas columnas como los niveles de glucosa, presión arterial, grosor del pliegue cutáneo, insulina y el índice de masa corporal (BMI) contenían valores de 0. Dado que estos valores no son válidos en un contexto clínico (ya que, por ejemplo, un nivel de glucosa o presión arterial de 0 no es posible en humanos), se consideraron como datos faltantes.

Para tratar estos valores, se implementó un proceso de imputación. Primero, los valores de 0 en las columnas afectadas fueron reemplazados por valores faltantes (NaN), de manera que pudieran ser tratados adecuadamente. Posteriormente, para cada una de estas columnas, se utilizó la imputación de la mediana como estrategia de reemplazo. Esta técnica asegura que las características numéricas se completen de manera robusta, sin afectar significativamente la distribución de los datos ni introducir sesgos, lo que es crucial para garantizar que el modelo predictivo se entrene con datos consistentes.

Finalmente, se verificó que no quedaran valores faltantes en las columnas tratadas, donde nuestro output fue el siguiente:

```
Glucose          0
BloodPressure    0
SkinThickness    0
Insulin          0
BMI              0
dtype: int64
```

Así, asegurando que todas las variables relevantes estuvieran completamente preparadas para su uso en el modelo de regresión logística.

2.5 Correlación de variables

A continuación, se genera un heatmap de correlación para visualizar las relaciones entre las variables.

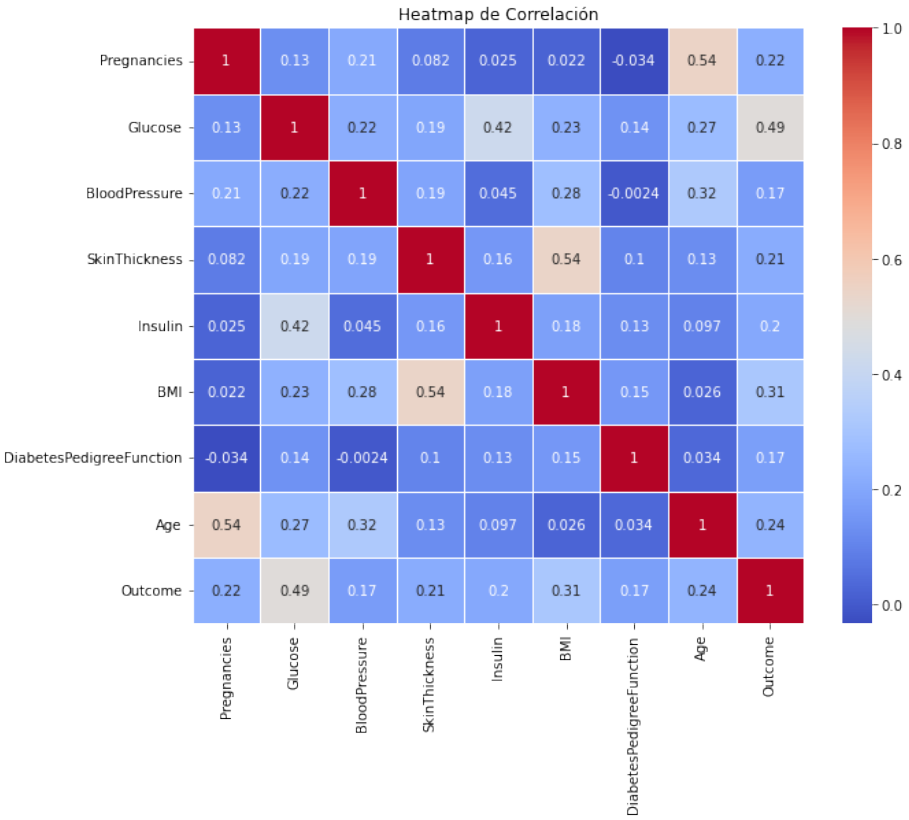

Figure 2: Heatmap de correlación

El heatmap de correlación permite visualizar las relaciones entre las distintas variables del conjunto de datos, proporcionando una primera aproximación sobre qué características tienen una mayor influencia en el resultado de la predicción de diabetes. Las correlaciones se miden en un rango de -1 a 1, donde los valores cercanos a 1 indican una fuerte correlación positiva, y los valores cercanos a -1, una fuerte correlación negativa.

Entre las observaciones más destacadas:

- **Glucose** muestra una alta correlación positiva con el Outcome (0.47), lo que sugiere que a medida que aumenta el nivel de glucosa en sangre, también lo hace la probabilidad de que un paciente desarrolle diabetes. Esto es consistente con la literatura médica, donde los niveles elevados de glucosa son uno de los principales indicadores de diabetes.

- **Age** y **Pregnancies** presentan correlaciones positivas moderadas con el Outcome. En particular, la correlación de Age indica que los pacientes de mayor edad tienden a tener un mayor riesgo de desarrollar diabetes. Del mismo modo, Pregnancies (número de embarazos) está moderadamente correlacionada, lo que podría estar relacionado con los cambios metabólicos y hormonales que ocurren durante y después del embarazo, afectando la salud metabólica de las mujeres.

- **BMI (Índice de Masa Corporal)** y **DiabetesPedigreeFunction** también tienen correlaciones moderadas con el Outcome. Un mayor BMI, que indica sobrepeso u obesidad, es un factor de riesgo conocido para la diabetes tipo 2. Asimismo, la DiabetesPedigreeFunction refleja la predisposición genética de un paciente a desarrollar diabetes, y su correlación con el Outcome confirma la importancia de la herencia genética en la probabilidad de desarrollar esta enfermedad.

Además de estas observaciones clave, es importante señalar que otras variables como **BloodPressure** (presión arterial) e **Insulin** no muestran una correlación fuerte con el Outcome. Esto no significa que no tengan relevancia clínica, pero en este conjunto de datos en particular, su relación con la diabetes no es tan marcada como la de otras variables. Estas observaciones sugieren que el modelo de predicción debe prestar especial atención a variables como **Glucose**, **BMI**, y **Age**, ya que parecen tener un mayor peso en la probabilidad de desarrollar diabetes.

2.6 Separación de datos

Se realizó la separación de las variables independientes ($\mathbf{X}$) de la variable objetivo ($\mathbf{y}$), donde $\mathbf{X}$ contiene todas las características del dataset y $\mathbf{y}$ representa el resultado de diabetes (Outcome). Posteriormente, los datos fueron divididos en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%). Esta división permite entrenar el modelo con una porción de los datos mientras se reserva el conjunto de prueba para evaluar el rendimiento del modelo, garantizando una adecuada generalización. Además, se mantuvo la proporción original de la variable objetivo en ambas particiones para asegurar un balance representativo en los conjuntos. Con esta división, los datos están listos para el entrenamiento del modelo de regresión logística.

3 Modelado

En esta sección se presentará la implementación del modelo de regresión logística, explicando su aplicación para predecir la diabetes en base a las características del dataset. Además, se detallará el proceso de entrenamiento y evaluación del modelo, incluyendo las métricas de rendimiento como la precisión, *recall* y F1-score. Finalmente, se incluirá una visualización del comportamiento del modelo, destacando su capacidad para clasificar correctamente los casos de diabetes y los desafíos en la identificación de la clase positiva (diabetes).

3.1 Regresión Logística

La Regresión Logística es un modelo de clasificación que sirve para estimar la probabilidad de que una observación pertenezca a una de las dos clases posibles. La probabilidad de padecer o no diabetes: Este modelo se basa en una función logística que convierte el resultado de una combinación lineal de las variables de entrada en probabilidades que van de 0 a 1.

En el contexto presentado, analiza las características de los pacientes y la posibilidad de que presenten diabetes, y según dicha probabilidad, asigna una clase u otra: 0 para los que no tienen diabetes y 1 para los diabéticos. El principal punto a favor de la regresión logística es la interpretabilidad y la aptitud para trabajar con datos categóricos y continuos, lo que lo convierte en el modelo más conveniente para realizar estimaciones de este tipo. Aunque no es un modelo muy complejo, es ampliamente utilizado, especialmente en casos cuando se requiere una combinación de rendimiento y capacidad de interpretar las predicciones.

3.2 Evaluación del modelo

Se procedió con el entrenamiento del modelo de regresión logística utilizando los datos previamente divididos en conjuntos de entrenamiento y prueba. El modelo se entrenó con los datos escalados del conjunto de entrenamiento. Posteriormente, se realizaron predicciones en el conjunto de prueba utilizando el modelo entrenado, obteniendo así las predicciones para cada paciente.

Para evaluar el rendimiento del modelo, se generó un reporte de clasificación que incluye las métricas clave como precisión, *recall* (sensibilidad) y F1-score para ambas clases: pacientes con diabetes (clase 1) y pacientes sin diabetes (clase 0). Los resultados del reporte indican lo siguiente:

- **Precisión**: El modelo tiene una precisión del 75% para la clase 0 (sin diabetes) y del 60% para la clase 1 (con diabetes), lo que significa que es más preciso al identificar a las personas que no tienen diabetes.
- **Recall (Sensibilidad)**: La sensibilidad es del 82% para la clase 0, pero baja a un 50% para la clase 1, lo que sugiere que el modelo tiene dificultades para detectar casos de diabetes.

- **F1-score**: El F1-score, que combina precisión y *recall*, es de 0.78 para la clase 0 y 0.55 para la clase 1, confirmando que el modelo funciona mejor al identificar individuos sin diabetes.

La precisión del modelo es del 71%, pero muestra un mejor rendimiento en la identificación de individuos sin diabetes que en aquellos con diabetes.

3.3 Curva ROC y AUC

La Curva ROC (*Receiver Operating Characteristic*) es una representación gráfica que evalúa el rendimiento del modelo de regresión logística en términos de su capacidad para distinguir entre las clases positivas (diabetes) y negativas (sin diabetes). Esta curva traza la tasa de verdaderos positivos (*True Positive Rate*) frente a la tasa de falsos positivos (*False Positive Rate*) para diferentes umbrales de clasificación.

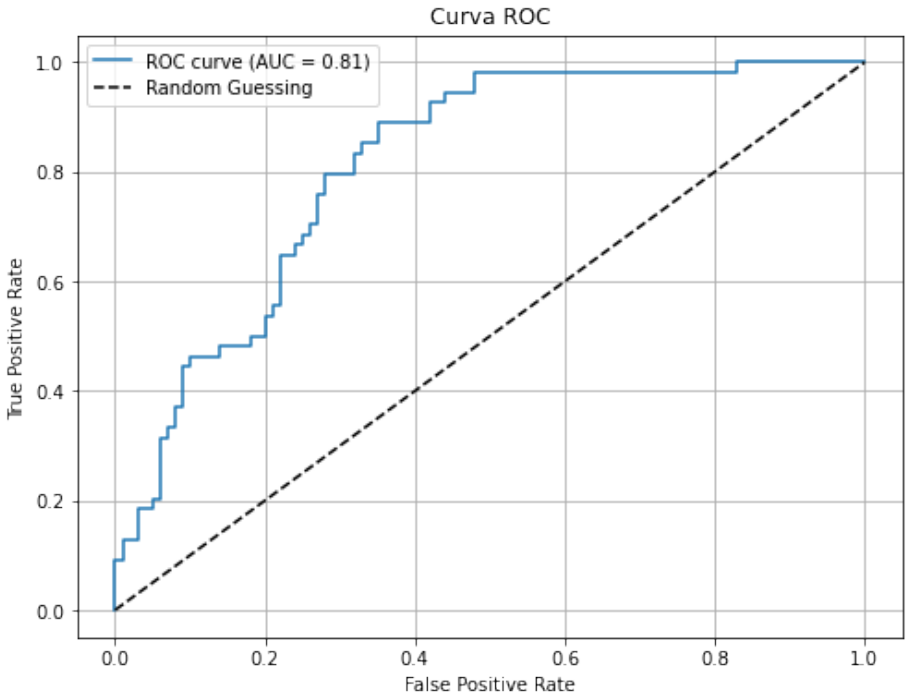

Figure 3: Histogrmas de variables principales

La Curva ROC refleja la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) del modelo de regresión logística, proporcionando una visión clara de su capacidad de clasificación. En esta gráfica, la curva se eleva rápidamente, lo que indica que el modelo es capaz de identificar correctamente un alto porcentaje de casos positivos (pacientes con diabetes)

con un bajo porcentaje de falsos positivos. Sin embargo, a medida que aumenta el FPR, el crecimiento de la TPR se vuelve menos pronunciado, lo que sugiere que el modelo pierde efectividad a medida que se reducen los umbrales de clasificación.

El valor del Área bajo la curva (AUC) es de 0.81, lo que sugiere que el modelo tiene un buen rendimiento general, ya que es capaz de distinguir entre pacientes con y sin diabetes en el 81% de los casos. Este valor de AUC está significativamente por encima de 0.5, lo que indica que el modelo es mucho mejor que una clasificación aleatoria. Sin embargo, también hay margen para mejorar, ya que un AUC más cercano a 1 representaría un modelo casi perfecto.

4 Optimización del Modelo

La optimización de hiperparámetros es un paso crucial en el desarrollo de modelos de *machine learning*, ya que permite mejorar el rendimiento del modelo ajustando los parámetros que no se aprenden directamente del entrenamiento.

Tras realizar la optimización de hiperparámetros mediante el uso de *GridSearchCV*, se exploraron diferentes combinaciones de parámetros clave para el modelo de regresión logística. El hiperparámetro **C**, que controla el nivel de regularización, fue evaluado en distintos valores (0.01, 0.1, 1, 10, 100), junto con el *solver* (método de optimización utilizado para ajustar el modelo) y la *penalty* (tipo de penalización aplicada). *GridSearchCV* permitió probar estas combinaciones usando validación cruzada, asegurando que el modelo fuera evaluado de manera robusta en diferentes particiones de los datos de entrenamiento.

El resultado de esta optimización identificó los mejores hiperparámetros para el modelo:

- **C**: 1, lo que indica que el modelo aplica una regularización moderada, evitando tanto el sobreajuste como el subajuste.
- **Solver**: *lbfgs*, un método de optimización eficiente que es comúnmente utilizado en problemas de clasificación.
- **Penalty**: *l2*, que es la penalización estándar para la regularización en regresión logística, favoreciendo soluciones que reduzcan la magnitud de los coeficientes del modelo.

Una vez encontrados los mejores hiperparámetros, se entrenó nuevamente el modelo con esta configuración óptima y se evaluó su rendimiento en el conjunto de prueba. El reporte de clasificación resultante mostró una mejora en las métricas generales en comparación con el modelo previo, confirmando que la optimización de los hiperparámetros ha permitido un ajuste más adecuado del modelo a los datos.

Este proceso asegura que el modelo tiene el mejor equilibrio posible entre simplicidad y rendimiento, evitando el sobreajuste mientras maximiza la precisión en la predicción de la diabetes.

5 Conclusión

El desarrollo de este modelo ha seguido un enfoque estructurado, desde el preprocesamiento de los datos hasta la optimización de los parámetros clave. Se tomaron decisiones importantes que influyeron en el rendimiento final del modelo, y se utilizaron diversas herramientas de análisis para evaluar su capacidad de predicción. El trabajo permitió explorar el impacto de múltiples variables en la predicción de resultados relevantes, lo que facilitó la identificación de patrones y tendencias importantes en los datos. Este trabajo representa una base sólida para la predicción y análisis de datos en contextos similares. La optimización del modelo permitió mejorar sus capacidades predictivas, aunque queda espacio para continuar refinando el enfoque y explorando otros modelos que podrían ofrecer un rendimiento aún mejor en escenarios más complejos.

Referencias

- Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265. Disponible en: `https://www.kaggle.com/uciml/pima-indians-diabetes-database`

- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Vol. 398). John Wiley & Sons.

- Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression: A Self-Learning Text* (Vol. 3). Springer Science & Business Media.

- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.

- Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (pp. 437-478). Springer Berlin Heidelberg.