

## ¿Cuál es el método estadísticamente válido para convertir rangos categóricos de tiempo de búsqueda de empleo en variables continuas para modelos de regresión no lineal?

¿Cómo se justifica científicamente el uso del punto medio (midpoint) en datos de encuestas con intervalos de tiempo para predecir la inserción laboral?

¿Qué impacto tiene la discretización de variables de tiempo en la precisión de los modelos de Gradient Boosting comparado con el uso de datos continuos?

## Modelar tiempos de búsqueda de empleo: métodos para categorías, midpoints y Gradient Boosting

### 1. De categorías de tiempo a variables continuas

No hay un método “único” estándar específico para rangos de búsqueda de empleo, pero la literatura de duración y de encuestas usa tres enfoques principales:

- **Modelos de duración/supervivencia:** tratan el tiempo como variable continua u observada por intervalos, evitando colapsar a un solo punto (p.ej. Kaplan-Meier, Cox, modelos de riesgo con ANN) (Benhamed, 2025; Cygu et al., 2023; Andonovikj et al., 2023; Boškoski et al., 2021).
- **Regresión no lineal/ML con tiempo continuo:** cuando se dispone del conteo exacto de días hasta empleo, se usa directamente como continuo en árboles oblicuos, Gradient Boosting, redes, etc. (Andonovikj et al., 2023; Boškoski et al., 2021).
- **Interval-censored survival:** cuando solo se conoce un intervalo (“3–6 meses”), la recomendación estadística es modelar explícitamente el intervalo, no reducirlo a un valor puntual, mediante modelos de supervivencia con censura por intervalos.

En estudios de transición escuela-trabajo y desempleo se prefiere supervivencia (Kaplan-Meier, hazard no lineal) precisamente para evitar supuestos fuertes sobre la distribución dentro del intervalo (Benhamed, 2025; Baydur & Xu, 2024; Boškoski et al., 2021).

### Opciones prácticas

Situación de los datos	Tratamiento recomendado	Citaciones
Tiempo exacto en días/meses	Usar continuo en regresión/GBM o supervivencia	(Benhamed, 2025; Andonovikj et al., 2023; Boškoski et al., 2021)
Solo intervalos de encuesta	Modelos de supervivencia con censura por intervalos (no midpoint)	(Benhamed, 2025; Cygu et al., 2023; Boškoski et al., 2021)
Pocos intervalos amplios y sin herramientas de supervivencia	Midpoint como aproximación, pero con análisis de sensibilidad	(Benhamed, 2025; Boškoski et al., 2021)

FIGURE 1 Tratamientos recomendados según tipo de medición de tiempo

## 2. Justificación y límites del uso del midpoint

El uso del **punto medio del intervalo** implica asumir que el tiempo real es uniformemente distribuido dentro del rango. En datos de duración de desempleo, la evidencia muestra **hazard no lineal y decreciente** en el tiempo (Benhamed, 2025; Baydur & Xu, 2024; Boškoski et al., 2021), lo que hace poco realista la uniformidad; por tanto, el midpoint es como máximo una **aproximación ad-hoc** y debería acompañarse de análisis de sensibilidad (p.ej. usar extremos del intervalo o simulaciones dentro del rango).

## 3. Efecto de discretizar tiempo en modelos Gradient Boosting

Los estudios que comparan tratar el tiempo como continuo frente a discretizado muestran:

- Incorporar covariables y tiempos en forma **rica y continua** mejora la discriminación en supervivencia; Gradient Boosting y GBM-survival superan a modelos más simples cuando se conserva la estructura temporal (Cygu et al., 2023; Andonovikj et al., 2023; Boškoski et al., 2021).
- Enfatizan que decisiones de discretización (p.ej. agrupar tiempos en categorías gruesas) pueden **degradar la precisión** y la capacidad de capturar no linealidades en el riesgo (Cygu et al., 2023; Andonovikj et al., 2023; Boškoski et al., 2021).

En resumen, para predicción de inserción laboral con Gradient Boosting, **es preferible usar el tiempo continuo o modelos de supervivencia con censura por intervalos**, y usar midpoints solo como solución de compromiso explícitamente reconocida como aproximación.

*These search results were found and analyzed using Consensus, an AI-powered search engine for research. Try it at <https://consensus.app>. © 2026 Consensus NLP, Inc. Personal, non-commercial use only; redistribution requires copyright holders' consent.*

## References

- Benhamed, A. (2025). Non-parametric Duration Models in the First Job for Young Graduates in Tunisia. *Journal of Posthumanism*. <https://doi.org/10.63332/joph.v5i6.2075>
- Cygu, S., Seow, H., Dushoff, J., & Bolker, B. (2023). Comparing machine learning approaches to incorporate time-varying covariates in predicting cancer survival time. *Scientific Reports*, 13. <https://doi.org/10.1038/s41598-023-28393-7>
- Baydur, I., & Xu, J. (2024). STATISTICAL DISCRIMINATION AND DURATION DEPENDENCE IN A SEMISTRUCTURAL MODEL. *International Economic Review*. <https://doi.org/10.1111/iere.12696>
- Andonovikj, V., Boškoski, P., Džeroski, S., & Boshkoska, B. (2023). Survival analysis as semi-supervised multi-target regression for time-to-employment prediction using oblique predictive clustering trees. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2023.121246>
- Boškoski, P., Perne, M., Ramesa, M., & Mileva-Boshkoska, B. (2021). Variational Bayes survival analysis for unemployment modelling. *Knowl. Based Syst.*, 229, 107335. <https://doi.org/10.1016/j.knosys.2021.107335>